Advances in Knowledge Organization
Vol. 5 (AKO-5)
published by the
International Society for Knowledge Organization

Editorial Board
Daniel Benediktsson, Debrecen, Hungary
Robert Fugmann, Idstein, Germany
Roland Hjerpe, Linköping, Sweden
Barbara Keim, Frankfurt, Germany
Norbert Meder, Köln, Germany
Roy Rada, Pullman, WA; USA
Yulius Shrejder, Moscow, Russia
Elaine Svenonius, Los Angeles, CA, USA
Rudolf Ungvary, Budapest, Hungary

ISSN 0938-5495
Fourth Conference of the
International Society for Knowledge Organization (ISKO)
held at the Library of Congress
Washington, DC, USA, July 15-18, 1996

Organizers:

Conference Chair:
Dr. Sarah Thomas
Office of the Director for Public Service Collections,
Library of Congress, Washington, DC

Chair, Local Arrangements Committee:
Dr. Jolande Goldberg
Cataloging Policy and Support Office
Library of Congress, Washington, DC

in cooperation with
The International Society for Knowledge Organization
Woogstr. 36a, D-60431 Frankfurt, Germany

Program Chair:
Prof. Dr. Rebecca Green
College of Library and Information Services
University of Maryland, College Park, MD

Program Committee:
Prof. Hanne Albrechtsen
Prof. James D. Anderson
Prof. Kenneth Bakewell
Prof. Dr. Clare Beghtol
Prof. Pauline A. Cochrane
Dr. Ingetraut Dahlberg
Prof. Dr. Brigitte Endres-Niggemeyer
Prof. Dr. Raya Fidel
Dr. Robert Fugmann
Mr. Alan Gilchrist
Prof. Dr. M.A. Gopinath
Prof. Dr. Roland Hjerppe
Prof. Dr. Hemalata Iyer
Prof. Krishan Kumar
Prof. Tamiko Matsumura
Prof. A. Neelameghan
Dr. Giliola Negrini
Dr. Bluma C. Perlitz
Prof. Dr. Dagobert Soergel
Prof. Dr. Nancy Williamson
Knowledge Organization and Change

Proceedings of the
Fourth International ISKO Conference
15-18 July 1996
Washington, DC, USA

organized by the
Office of the
Director for Public Service Collections
Library of Congress
the ISKO General Secretariat and
OCLC Forest Press

Edited by
Rebecca Green

Frankfurt/Main
INDEKS VERLAG
1996
(Proceedings of the ... international ISKO conference ... ; 4)
(Advances in knowledge organization ; Vol. 5)
ISBN 3-88672-024-1
NE: Green, Rebecca [Hrsg.]; 2. GT

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in other ways and storage in databanks. Duplication of this publication or parts thereof is only permitted under the provisions of the German Copyright Law, a copyright fee must always be paid.

(c) 1996 by INDEKS Verlag, Frankfurt/Main
Cover Design: AVIVA W.Dahlberg
Printers: Druckerei Guntrum II, KG, Schiltz/Hessen
Printed in Germany
ISBN 3-88672-023-1; ISSN 0938-5495
Contents

Preface

Keynote address:
Roland HJERPPE: Go with the Flow, or Abide by the Side, or Watch the Waves?
Challenges of Change for Knowledge Organization
10

1. Library of Congress Classification
Rebecca S. GUENTHER: Bringing the Library of Congress Classification into the
Computer Age: Converting LCC to Machine readable Form
26
Jolande GOLDBERG: Library of Congress Classification: Shelving Device for
Collections or Organization of Knowledge Fields?
33

2. Knowledge Organization in Cross-Cultural and Cross-Linguistic Settings
Snunith SHOHAM and Moshe YITZHAKI: The Impact of Cultural and Technological
Changes on Titles Content and Their Use in the Process of Information Retrieval
43
Marcia Lei ZENG: Towards A Unified Medical Language in a Diverse Cultural
Environment
51
Shaoyi HE: Concept-based vs. Word-based Measures of Medical Information Transfer
via English-Chinese and Chinese-English Translation of Medical Titles
59

3. The Role of Relationships in Knowledge Organization
Pat MOLHOLT: Standardization of Inter Concept Links and Their Usage
65
Rebecca GREEN: Development of a Relational Thesaurus
72
Carol BEAN: Analysis of Non-Hierarchical Associative Relationships Among
Medical Subject Headings (MESH): Anatomical and Related Terminology
80

4. Knowledge Organization in the Online Environment, I
Javier GARCIA MARCO: Hypertext and Indexing Languages: Common Challenges
and Perspectives
87
Pauline ATHERTON COCHRANE and Eric H. JOHNSON: Visual Dewey: DDC in a
Hypertextual Browser for the Library User
95
Karen M. DRABENSTOIT: Classification to the Rescue: Handling the Problems of
Too Many and Too Few Retrievals
107
5. Management of Change in Knowledge Organization Schemes

Eduard R. SUKIASYAN: Change as a Problem of Classification System Development 119

Ia C. MCILWAINE: New Wine in Old Bottles: Problems of Maintaining Classification Schemes 122

6. Thesauri and Metathesauri, I

Ron DAVIES: Thesaurus-aided Searching in Search and Retrieval Protocols 137

Victoria FRANCU: Building a Multilingual Thesaurus Based on UDC 144

Ingetraut DAHLBERG: Library Catalogs in the Internet: Switching for Future Subject Access 155

Dagobert SOERGEL: SemWeb: Proposal for an Open, Multifunctional, Multilingual System for Integrated Access to Knowledge about Concepts and Terminology 165

7. Knowledge Organization and Images

Elin K. JACOB and Debra SHAH: Is a Picture Worth a Thousand Words? Classification and Graphic Symbol Systems 174

Olivia C. FROST: The University of Michigan School of Information Art Image Browser: Designing and Testing a Model for Image Retrieval 182

Corinne JÖRGENSEN: The Applicability of Selected Classification Systems to Image Attributes 189

8. Knowledge Organization in the Business and Economic Environment

Harold C. STEYER, Jr., Ana Flavia FONSECA, Diane D. HOPKINS, Marc NODELL, Irene L. TRAVIS, and William S. WAHL: The World Bank's Information Management Architecture: A Blueprint for Building Institutional Information Services 198

Philip C. MURRAY: Business Productivity and Organization of Knowledge: A Look at the Emerging Requirements 206

J.F. SCHREINEMAKERS and J.P.J.M. ESSERS: Critical Notes on the Use of Knowledge in Knowledge Management 213

9. User Focus in Knowledge Organization

Michelle M. FOSS: Facilitating the Interaction of User and Public Knowledge Organization with User Profiles, User Views and User Education Plans 216

Mirja IVONEN: Selection of Search Terms as a Meeting Place of Different Discourses 224
A. Steven POLLITT, Mark TREGLOWN, Martin P. SMITH, Patrick A. J. BRAINTVELT, Geoffrey P. ELLIS, Janet E. FINLAY, and Steven J. WADE: Empowering Users for Improved Database Access and Analysis through the Application of Knowledge Structure Views, Progressive Refinement Techniques and a Design Approach Driven by Usability

10. Knowledge Organization in the Online Environment, II

Rebecca GREEN: Description in the Electronic Environment

Diane VIZINE-GOETZ: Online Classification: Implications for Classifying and Document-like Object Retrieval

Gerry MCKIERNAN: Parallel Universe: The Organization of Information Elements and Access in a World Wide Web (WWW) Virtual Library

Michiaki IWAZUME, Hideaki TAKEDA, and Toyoaki NISHIDA: Ontology-based Information Capturing from the Internet

11. Interdisciplinary Approaches to Knowledge Organization

Lynne C. HOWARTH: An Exploratory Study into Requirements for an Interdisciplinary Metathesaurus


Massimo RAGUCCI: An Inductive Approach towards the Integration of General Information Systems for Agriculture: The Case of CERETHEIS, with Particular Examples

Jeffrey T. HUBER and Mary L. GILLASPY: Controlled Vocabulary for HIV/AIDS: An Evolving Nosological: Record of a Diseased Body of Knowledge

12. Interplay of Epistemology and Knowledge Organization

Hope A. OLSON: Dewey Thinks Therefore He Is: The Epistemic Stance of Dewey and DDC

Roberto POLI: Ontology for Knowledge Organization


13. Natural Language Processing

A.R.D. PRASAD: PROMETHEUS: An Automatic Indexing System

Toshiyuki MATSUO and Toyoaki NISHIDA: Intelligent Support for Construction and Exploration of Advanced Technological Information Space
Widad MUSTAFA ELHADI and Christophe JOUIS: Evaluating Natural Language Processing Systems as a Tool for Building Terminological Databases 346

14. Thesauri and Metathesauri, II

Faina CITKINA: Terminology Organization and Change 355

Michèle HUDON: Preparing Terminological Definitions for Indexing and Retrieval Thesauri: A Model 363

Nancy WILLIAMSON: Deriving a Thesaurus from a Restructured UDC 370

15. Dewey Decimal Classification

Joan S. MITCHELL: The Dewey Decimal Classification at 120: Edition 21 and Beyond 378

Gregory R. NEW: Revision and Stability in Dewey 21: The Life Sciences Catch Up 386

Julianne BEALL: Dewey for Windows 396

16. Dewey 120th Anniversary Address:

Francis MIKSA: The DDC, the Universe of Knowledge, and the Post-Modern Library 406

List of Contributors 413

Name and Subject Index 418
Preface

This volume documents the formal presentations delivered at the 4th International Conference of the International Society for Knowledge Organization (ISKO) in Washington, DC, July 16-18, 1996. The conference was sponsored by the Library of Congress, where the meetings were held, in cooperation with OCLC Forest Press. The conference was also supported by the Classification Research Special Interest Group of the American Society for Information Science (ASIS SIG/CR) and the College of Library and Information Services (CLIS) of the University of Maryland. Previous international conferences were held in Copenhagen (1994), Madras (1992), and Darmstadt (1990).

ISKO '96 addressed "Knowledge Organization and Change," a theme outlined in the keynote address by Roland Hjerppe. This theme was developed through a series of sessions on: Knowledge Organization in Cross-Cultural and Cross-Linguistic Settings, The Role of Relationships in Knowledge Organization, Knowledge Organization in the Online Environment (2 sessions), Management of Change in Knowledge Organization Schemes, Thesauri and Metathesauri (2 sessions), Knowledge Organization and Images, Knowledge Organization in the Business and Economic Environment, User Focus in Knowledge Organization, Interdisciplinary Approaches to Knowledge Organization, Interplay of Epistemology and Knowledge Organization, and Natural Language Processing. Although various classification schemes received attention throughout the conference, there were also sessions specific to the Library of Congress Classification and the Dewey Decimal Classification. There was also one session for software demonstrations.

A variety of experiences rounded out the three days of paper presentations. Pre-conference activities included a demonstration session at the Digital Library Visitors' Center, a guided tour of the Library of Congress, and a reception sponsored by OCLC Forest Press in commemoration of the 120th anniversary of Melvil Dewey's Decimal Classification. The reception was preceded by a talk by Fran Miksa on "The DDC, the Universe of Knowledge, and the Post-Modern Library." The next evening a banquet was held at the Supreme Court of the United States, where the group was addressed by Douglas Bennett of the American Council of Learned Societies. The final evening of the conference, a reception was hosted for the group by the German Embassy. Two post-conference excursions—one to the National Agricultural Library, the other to the National Library of Medicine—rounded out the occasion.

Producing the proceedings of a conference is a mammoth undertaking. Of course, none of this volume would exist but for the thoughtful effort of those who proposed, wrote, and presented, as well as of the program committee, who reviewed the initial abstracts and gave valuable feedback. I acknowledge the help of Susan Westenbarger with the initial editing of the full papers. And as anyone who knows anything about ISKO will surely expect, I gratefully acknowledge the unflagging effort of Ingetraut Dahlberg, who is responsible for the index, and, who, through INDEKS Verlag, is fully responsible for the proceedings being anything more than a pile of pages on my desk.

Rebecca Green
Program Chair
17 May 1996
Roland Hjerppe  
Liblab - Department of Computer and Information Science, Linköping University

Go with the Flow, or Abide by the Side, or Watch the Waves? Challenges of Change for Knowledge Organization

Abstract: Internet itself and the dynamic and diffuse nature of documents and collections on Internet are discussed. Some previously obscured assumptions of current and past modes of knowledge organization, based on a rejection of change, become apparent against this background. Three types of consequences found in the examination of these assumptions are outlined: that knowledge organization is a matter for everyone, there is a need for a re-definition, and for several partial re-orientations of aims, for knowledge organization. A number of proposals on areas for reflection, research and action are presented and the three types of responses to the challenges of change are briefly discussed.

The future of the Internet is a real-time future. We'll see a vast shift from static sites today to a much more dynamic world with information always changing with real-time streams of media—audio, video and data. (Nova Spivack, 1996)

Preamble

The aim of this paper is to indicate trends, expose hitherto obscured assumptions, discuss implications of these findings, and propose directions for future action. The title could also have been something like: From static to dynamic, from use to user, from principled systems to ad-hoc principles. It is more focused on the bases of knowledge organization and external factors than on current practice. Three types of responses (there are others) to the challenges of change are discussed at the end. Some of the challenges of change for libraries and archives have been outlined in a companion paper (Hjerppe, 1996).

In order to cover more ground than the limits imposed by the context, a keynote paper for a conference, usually would permit, some of the traditional rhetorics have been sacrificed. It is thus not a traditional scholarly paper, focusing on a specific, well defined issue, with all the proper references to prior literature, etc. Most of the pointers provided are to electronic resources, i.e., URLs, and many of them are to collections of resources rather than primary material. A search for traditional literature on knowledge organization and change yielded surprisingly few results, most of them recent, e.g., Lockenhoff (1994) and Schipper (1994).

We are at the juncture where the new modes of communication and expression enabled by distributed hypermedia—Internet, are being tested and experienced but where scholarly communication still (with some experimental exceptions) operates in the traditional mode—writings intended for printing on paper. The writing of this paper has hence been persisted experiences of frustrations and impatience. Foremost among these have been the need to linearize the presentation of subjects that form networks, internally and between each other, and the inelasticity and passivity of pointers on papers compared to the immediacy and reach of links in the hypermedia environment of Internet. (An interesting question on the side is what form conference presentations will have in the future when the conference papers are integrated into the web at submission, which perhaps consists of a URL?)

1. Introduction

Constancy and change are the foreground and background, the ying and yang, for each other. The concept of change is extremely closely tied to space-time. The passing of time is noticed through changes in the surrounding space, and change implies a transition, from one state, before/there, to another, after/there. In epistemology and ontology there are two extreme positions that can be taken, the Heraclitean "panta rhei", or the Parmenidian view that change is an illusion.

Our perceptual apparatus has evolved to discern changes of different types within and without ourselves. The dynamic ranges of our senses are limited and instruments have thus been developed to extend these ranges. The dynamic ranges of our senses seem to be adapted to rates of change that we need to and can react to. There are lower bounds on the rate of change we can sense; we cannot experience directly, e.g., the growth of grass, or "the slow fires" of decomposing paper. There are upper bounds as well; we cannot see a bullet fired from a rifle nor hear sounds of very brief duration.

The changes we perceive because they are changes close in time are different from the changes we are aware of because of memories, e.g., recurring change: the circadian, lunar, seasonal rhythms, which provide a different kind of change, renewed returns.

Memories are, however, not reliable, or rather, they are reinterpreted or put away. Different means have therefore been found and invented to support memory and to enable communication across space and time.

The intent of this paper is, however, not to delve into philosophical nor into perceptual or cognitive issues but rather to discuss the consequences of brisk changing change—flux, for knowledge organization. Knowledge organization has had two foci: documents—as carriers of representations of knowledge, experience, expressions, opinions, . . ., and the abstractions of these represented entities—knowledge, in abstract. Both of these have been static or semi-static, immutable (the documents) or changing only slowly (abstractions of current state of knowledge).

2. Internet—the Changing, Amorphous Catalyst

The impetus for the reflections in this paper has been provided by the rapid emergence and acceptance of Internet. Internet is today to many the epitome of information technology (IT), and its popular image is the World Wide Web—W3, as seen through one of the W3-browsers. Although the roots of Internet go back to the late 1950’s and early 1960’s, and although the term Internet has been used since 1982 as designation for interlinked networks, and although Internet was a commonplace in the science and engineering domains of Western Academia since mid-1980, it wasn’t until 1994 that the public at large began to identify IT and Internet with each other, partly through the introduction of the graphical interface to W3, partly through the "discovery" of journalists of W3, and partly through the proliferation of computers.

There is already more written about Internet than anyone could ever hope to read and digest. The first place to look for information on Internet is of course Internet itself. The string "Internet" had in early April 1996 more than 5.5 million mentions among the 22 million W3-documents indexed by the Alta Vista search engine (internet is on the stop list, a non-discriminating word). Open text found more than 270 000 mentions. A Lycos search for Internet among its 37,643,037 unique URLs found 535,652 documents with the word internet. (The discrepancies can partly be explained by the fact that the first document listed by Lycos had more than 330 mentions of Internet!)

 Dialog had in early April 1996 a total of close to 800,000 mentions in 340 files, with a very skewed distribution, 215,000 mentions in just one file (KR/T Business News 1989-1996/April 9)
and 588,000 mentions in the 15 first files.

Much of it is repetitive, much of it is superfluous, and much of it uses the term Internet as a designator of place/mode, and many of the files that have a lot of mentions are full text files, newspapers, etc., but most of these mentions have been made in a few years.

More than half of the mentions in Dialog, almost 53%, or 418,797 are from 1995, a further 27% are from the first quarter of 1996. Approximately 70% of the mentions are thus very recent, from 1995 or 1996. A cumulated frequency/time distribution diagram, see fig. 1 below, shows this very clearly. That which seems to be a drop in the exponential growth is due to the fact that the figures for 1996 are for the first quarter of 1996 only.

2.1 What Is Internet?

There are no easy general answers to the question of what Internet is, but a few classes of answers are provided below to provide a context.

- General answers:
  - A very large global open information metanetwork, a network of networks
  - That which enables communication between a very large number of computers all over the world

- Technical answers:
  - Narrow definition: IP-Internet—That part (ca. 7 million, July 1995,) of all registered internets that can exchange Internet protocol packets, i.e., the set of packet exchange-networks that have a common name and address space: DNS, and use the same protocol: TCP/IP
  - Broad definition: IP-Intemet (above) plus all connected networks that can transfer traffic to a destination unit or process. This includes gateways at application levels (e.g., e-mail), store-and-forward networks, and networks that use non-IP protocols

- Resource emphasizing answers:
  - All the computers (10? million) that can be reached through the network, and
  - some of the programs and data that is in these computers, and
  - some of the services, information and experiences that these programs and data enable access to and use of

- Socially oriented answers:
  - The people (>30 million persons) who use, develop and can be reached through these networks and resources
  - The (sub)cultures and communities that have grown up in these networks

- Communicative role answers:
  - A new medium, or
  - access to everything that can be digitized,
  - for interaction, with data, information, equipment, people, organizations, . . . ,
  - for (in)human activities: knowledge development, learning, entertainment, action, business . . .

- Consequences/problem stressing answers:
  - Something that causes legal (e.g., copyright) problems, etc.
  - Something used for disseminating unwanted (e.g., pornography) material, etc.
  - Something leading to greater inequalities, etc.
Internet is at the same time many things:
- a view by the user,
- a groupware, a collective workplace,
- a publishing medium for individuals, groups, organizations,
- a means for making money,

Of all the characteristics of Internet there are two that are implicated more than the others in this paper and they are its diffuse and ever-changing nature. This diffuseness and change is apparent at all levels, from the highest, Internet as a whole, to one of the lower, the documents. What keeps it all together are protocols at all levels, which also evolve.

2.2 Internet Is in Flux

Internet is in flux, ever-changing, in many different ways. The ones that are of interest here are those have to do with documents and document collections, their contents, and the organization issues relating to these.

Whereas print-on-paper documents integrate presentation, structure and content in one physical container, the use of W3 documents is based on a distribution of labor. The documents contain their own structural descriptions (using HTML, an application of SGML) and these are used by viewers/browsers (separate from the documents) to generate a view/presentation of the document. (The division of labor is actually much more distributed than is possible to relate here, among devices, programs, protocols, etc., and people.)

The views/presentations of the documents is partly under the viewer’s/browser’s control which implies that different viewers/browsers will provide different views/presentations. The viewer/browser for a handheld device cannot provide the same presentation as that for a sophisticated workstation.

The views/presentations of the documents is partly under the user’s control (by setting parameters of the viewers/browsers) which implies that different users will have different views at the same time and the same user might have different views at different times. The viewer/browser at office can be different from the one at home, or the one at hand.

The views/presentations of the documents is partly under the creator’s control. The structure and content within the structure is determined by the creator, within the bounds imposed by HTML, which is continuously being revised to accommodate the changing demands of users.

Internet documents appear, and disappear. Some of them are mirrored, i.e., replicated, at several abodes. If they reappear, who can tell whether its a new document, a changed document, or an old one?

Internet documents change: their appearance, extent and contents can and does often change as time passes. The abode of Internet documents, which is manifest as their location/address, URL, changes.

Internet documents embrace all media, all modes of expression. Internet documents, being digital, can furthermore perform transformations between these media and modes of expression. Text can, e.g., be transformed to speech or Braille, sounds can be encoded and presented in symbolic form, as notes or text, depending on the sounds and the needs.

With Java, and compliant viewers/browsers, the document-viewer distinction is also eroded, processes/programs are also becoming integral parts of documents, or documents parts of processes/program.

Collections, however they are defined, are likewise in flux. When documents are in flux and
documents appear and disappear, how can collections be static?

2.3. Internet Is Diffuse

Internet is furthermore diffuse. The documents are diffuse, a document can be monolithic; one medium, one mode, one file, etc., or composed of many parts, many media, many modes, many files, etc. A document that thus is built as a network of nodes with links between them will often have links to other nodes, external to it, in some senses. Where does in that case a document start and end? What is a document? Is a collection of pointers a document? Is one pointer a document?

Collections are likewise diffuse. When the items in a collection are diffuse how could collections be distinct? Will the membership in a collection of an item remain stable when the item changes? Will the categorizations applied to it at one time remain applicable?

Categorizations are also diffuse, built as they are in part on diffuse bases. As functions and capabilities are extended the categorizations that once were easy become difficult, as is witnessed by the questions above.

Previously stable roles become diffuse. Authors take on some of the roles of publishers, the printers and distributors are replaced by Internet service providers and networks, software and hardware producers were previously not needed and are now indispensable. Publishers look for new opportunities to capitalize on their old competencies. Broadcasters venture into publishing and publishers provide broadcasting material.

3. Hidden Assumptions of Knowledge Organization and Their Consequences

Accepting this flux and diffuseness: what can we learn about knowledge organization from Internet?

One of the important lessons for knowledge organization is the realization that our present systems for and habits of categorization and description are based on a number of assumptions, hidden until now, that no longer are necessarily valid. The consequences of the acceptance of these assumptions are many and varied but only some of them are apparent at this level of analysis. The implications of the unearthing of these hitherto hidden assumptions will need to be discussed in the future.

The first set of these obscured assumptions is that documents:

- are static, immutable, with respect to structure, content and presentation
- are delimited, with respect to extent
- are monolithic, or decomposable
- have distinct characteristics

The second set of tacit assumptions (related to the first and third) is that identity:

- is (i.e., there are identities)
- is static
- is specific

The third set of these implicit assumptions is that collections:

- are circumscribed
- grow, in size mainly, through additions of items
- are seldom weeded
The fourth set of these silent assumptions is that systems for knowledge organization:

- are based on different types of manifestations of knowledge—documents, which
  - belong to distinct domains
- are structuring devices, for documents, based on abstractions of their content
- are static, change occasionally, reluctantly, in stages
- are isolated, separate from other activities
- have distinct and different kinds of users

Some of the consequences of these sets of assumptions are obvious, e.g.,

- documents are described and categorized
  - once
  - as a totality
  - as abstractions, i.e., each document is seen as an instance of some class
- concepts, documents, collections
  - are identifiable, nameable
  - are regarded as (essentially) the same, as time passes
- collections
  - remain separate (might be dispersed, but are seldom integrated)
  - are seldom reorganized
- systems for knowledge organization
  - differ for libraries, archives, museums, etc.
  - are designed and constructed by appliers for appliers, for use in/by institutions

There are of course other unstated assumptions, with consequences, that will become apparent as other developments enable us to re-examine our future bases.

4. Implications for Knowledge Organization

Change in general, and the examination of the hidden assumptions, has implications for all aspects of knowledge organization: epistemological bases, design principles, construction processes, implementation details, application practices, and end user use.

Seen from a temporal perspective it is possible to distinguish:

- the current situation and existing systems
  - their present and future applications,
  - their adaptations and revisions,
- present and future situations and future systems
  - their application to traditional documents and collections,
  - their application to documents and collections that change.

The situation at present cannot be sustained, except for the continuation of current practice for existing collections. Current systems are inadequate, they were built for different situations, different types of documents, different types of collections, different producers and users. Current modes of using them are also inadequate. Application by institutions with regularized procedures, etc., is for use by users in other situations, with other needs than those envisaged at the design and construction, etc., of the systems. Current manners of designing, building, etc., are also antiquated, have the old foci.

What guidelines are then available for discussing future systems for knowledge organization, if the arguments above are accepted? A number of differing guidelines can be envisaged contingent
4.1. Knowledge Organization Is a Matter for Everyone

Knowledge organization is a matter/necessity for everyone. Everyone using W3 and collecting bookmarks/hotlists faces the need to organize them, to build a knowledge organization, but the only immediately available tool is the simple hierarchical structure, as represented by lists with folder nestings. More and more people are using W3, at work and at home, and although it at present is inconceivable that "everyone" will use Internet the number of users is and will be much larger than the number of users of any existing system for knowledge organization. As the number of bookmarks/hotlists grows, as a consequence of continued use, the need for organization becomes manifest.

Knowledge organization in everyday life becomes more and more important as computers invade the lives of people outside work, including the homes. Computerization brings the need for organization, in abstract, as opposed to the organization of concrete objects, although the metaphors for and interfaces to these might be designed to resemble the traditional world, see, e.g., Bjerg (1994). Information seeking in the context of everyday life is one of the projects at the Department of Information Studies, University of Tampere, Finland (Annual Report, 1995) that, however, focuses seeking rather than organization.

Of the 3*2 possible worlds formed by the dimensions public-private, paper-digital, stable-dynamic (see table below, in which the dynamic, paper worlds have been grayed out since they are regarded as a contradicticio in adjecto) it is only those constituted as stable, public, paper or digital that at present have functioning systems for knowledge organization. The stable, private, paper or digital domains could conceivably be served by traditional systems for knowledge organization whereas the dynamic, digital, public or private domains are in need of new forms of knowledge organization. (It is at present not quite clear what the dynamic, digital, private domain represents but there is no doubt that there will be one.)

<table>
<thead>
<tr>
<th>Stable</th>
<th>Public</th>
<th>Private</th>
<th>Dynamic</th>
<th>Public</th>
<th>Private</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper</td>
<td>Libraries</td>
<td>Everyday life papers</td>
<td>Paper</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Digital</td>
<td>Electronic books, etc.</td>
<td>Everyday life files</td>
<td>Digital</td>
<td>&quot;Internet&quot;</td>
<td>?</td>
</tr>
</tbody>
</table>

Tools, for building and using systems, are needed. Some first attempts at providing assistance for the management of Internet related information at the individual and group level are now appearing. Being the first generation these tools organize links in tree- or Mac-Finder-like formats. Among these programs are:

- DragNet (http://www.onbasetech.com/DragNet.html)
- Emissary (http://www.twg.com/emissary/quintro.html)
- GrabNet (http://www.ffg.com/grabnet.html)
- Mariner (http://www.mariner.ncd.com/ProdInfo/index.htm)
- SmartMarks (http://home.mcom.com/comprod/smartmarks.html; an add-on to Netscape)

Some of them monitor and manage changes on the Internet and private Intranets, automatically notifying the user when specified information changes. Tools for management of addresses and activities, usually called Personal Information Managers—PIMs or Personal Digital Assistants—PDAs, abound, see Carlsen, Dieterich and Schneider-Hufschmidt (1994) for a view on universal personal information management. From a knowledge organizational perspective these tools are primitive.
There is thus a danger that all the mistakes of knowledge organization of the past will be repeated again and again in the development of these kinds of tools unless the knowledge and experiences from research on knowledge organization is communicated to and used in the building of tools. Likewise the aspirations, ideas and knowledge that these tool-builders have should be used in thinking about tools for knowledge organization in a changing environment. The capabilities and possibilities of formal and systematic approaches as realized in various knowledge based approaches need to be utilized.

Among the more interesting approaches, in general and in this context is CYC, a very large, multi-contextual knowledge base and inference engine, having "a huge amount of fundamental human knowledge: facts, rules of thumb, and heuristics for reasoning about the objects and events of modern everyday life. CYC is an attempt to do symbolic AI on a massive scale." CYC has a markedly low profile on Internet but an FAQ is available, see Whitten (1994).

The goals and criteria for knowledge organizations are for the changing environment not necessarily the same as for a static. More formalization could, perverse as it may seem, be one of the means needed to enable adaptation to change. Formalization enables knowledge based approaches to the handling of change. Two of the questions that need to be put and answered in this respect are: is there a need for re-defining systems for knowledge organization, and a re-orientation of their aims, in general and with respect to change?

4.2 Re-defining Systems for Knowledge Organization?

A system for knowledge organization is a strange thing, and, actually, strictly speaking, a misnomer. Knowledge cannot be organized. Abstractions of knowledge, e.g., concepts, given verbal representations, can, and collections of objects of various kinds, carrying representations of knowledge, often need to. A system for knowledge organization is therefore usually a system for organization of (expressions of) concepts or (descriptions of) objects.

Any such system is of course based on a stance, a conception of the nature of a kind of knowledge. Every system for knowledge organization has an epistemological foundation. Unfortunately these foundations are oftentimes neither deliberate nor explicit.

One of the findings from reflecting on the possible/potential epistemological foundations for systems for knowledge organization that accept or even assume change as an intrinsic quality is that only a small subset of all kinds of knowledge have been the focus and a limited number of epistemological bases have been used for past and present systems. A related finding is that the ones used are all similar, i.e., the unfocused kinds of knowledge and unused epistemological bases have a greater variety than the ones focused and used. A similar situation has been found in Information Systems, see, e.g., the analysis by Ivari (1991) who concludes that there is an identifiable orthodoxy in information systems research.

The ontological and epistemological bases of library and information science have been the subjects of occasional papers, examples of three recent ones being the papers by Budd (1995), Marco and Navarro (1993), and Dahlberg (1992).

In an attempt at describing systems for knowledge organization truly ab initio most concepts that are used: knowledge, organization, system, structure, relation, etc., would have to be defined, and somehow grounded in something. Recognizing this the workers in various fields of Artificial Intelligence have in the last few years started to build and use something they call formal ontologies—specifications of conceptualizations, for the sharing and reuse of knowledge among software entities, see Gruber (1993) for a discussion of design principles for such ontologies and Knowledge Sharing Effort public library (n.d.) for an overview of the Knowledge Sharing Effort, which is one of the generators of the formal ontologies, with further pointers to Ontolingua, a set
of tools for analyzing and translating ontologies. One compilation of links to further information on ontological modeling can be found at KBSI (n.d.).

What are then the differences between formal ontologies and systems for knowledge organization? One of the main differences is that formal ontologies are designed and built using strict representation formalisms, for use mainly by software entities, grounded, i.a., in logical theory, whereas systems for knowledge organizations have been built using semi-natural language, for use by people. Since software agents do not have any understandings of terms or concepts of the kinds that humans have, not having grown up in the world and in that process, i.a., acquired a language, there has been a need to replace that understanding with formalism and logic.

An application of these formalisms and representation methods to systems for knowledge organization will be necessary for many reasons and with that will come a need for re-examination and re-definition. In the processes of making digital versions of systems for knowledge organization, viz., Electronic Dewey™, the first steps in this direction have been taken. The needs for agreements on formats, i.e., representations, in order to exchange data, will also provide incentives for movements towards the same direction. The USMARC Format for Classification Data is in this respect still just a container with content intended for human use and simple computer-based handling; it cannot be used for knowledge sharing and reasoning the way ontologies are.

All of these, systems for knowledge organization and formal ontologies, are, however, at present mainly concerned with representing static knowledge, a description of the state at a specific point or span in time. As time passes new versions of the systems for knowledge organization are generated, and the old ones are discarded. Until recently enough time had passed between each version so that the questions of the relations between material—old and new, and versions—old and new, were solved by applying new versions to new material, with sedimented collections as a result. The sediments are the result of the applications of new versions, and hence the boundaries of the layers do not necessarily reflect the changes of each “subject”. One of the consequences is that current systems for knowledge organization are focused on mapping future knowledge states, documents and collections in terms of the situations at past times. In a situation where a very large number of people have access to vast and quickly growing numbers of changing and diffuse multimedia documentsstreams and collections this is not enough.

A re-definition of knowledge organization is, if the preceding arguments are accepted, necessary to accommodate the challenges of dynamic situations, documents, collections. Some proposals on the bases for such a redefinition will be made later, after an examination of whether a re-orientation of aims is also necessitated by the arguments so far.

4.3 Re-orienting the Aims of Systems for Knowledge Organization?

What are the aims of knowledge organization and of systems for knowledge organization? These might, in the context of this paper—a conference on knowledge organization, seem to be odd or unnecessary questions, but a consideration of the need for a re-orientation of the aims has to start with an examination of current aims and relate those to the needs and situations at hand.

There are of course differences between the aims of knowledge organization in general, in abstract, and the aims of systems of knowledge organizations, in general, and in specific, concrete cases. This section of a brief essay can mostly discuss aims (and means, since they are the operationalizations of aims) in general and abstract terms. The traditional aims and means of a system for knowledge organization could perhaps be summarized to be to provide an abstract tool, that:
• maps the state of knowledge, for a specified domain, at a certain time, by
• establishing a system of terms and functions, that
• denote accepted bodies of knowledge or concepts in that domain, and
• indicate the relations between these,
• provides rules and methods for
• generating combinations, and
• linearizing all instances,
• can be used to
• categorize specific items,
• sort collections of categorized items,
• find categorized items,

The aims and means of knowledge organization, as an activity and a discipline, can also be summarized to be, i.a., to investigate the various bases and principles of systems for knowledge organization, the methods and practices of designing, building, implementing, applying, using, etc., systems for knowledge organization.

At this general level of presenting aims there does not seem to be any need for re-orientation of aims. One of the problems with the descriptions above is that they are static, reflecting static views of static entities and systems. To paraphrase what Gertrude Stein once said of Oakland: there is neither any there there nor any when when. What is needed above all—and this is a proposed partial re-orientation, not a complete break with the previous—is an explicit awareness and recognition of time and change. We are all cognizant of the fact that knowledge evolves, but that has not had any impact on present modes of thinking about knowledge organization. We are behaving like an explorer or driver, who steers and builds maps by looking backwards, through a rear mirror, and only occasionally looks to the sides, to update the map.

One such re-oriented view could perhaps be modeled on organisms, recognizing birth, growth, and death for individuals, and evolution for species. Awareness and recognition of time and change bring new issues, new questions to knowledge organization, viz., recognition and identification of changes. In a time-based perspective questions about the development of bodies of knowledge, the states at specified times, and the changes would be natural and answerable. In the paper by Chen and Gaines (n.d.) there is, i.a., an interesting discussion of the need for what they call chronological awareness in group systems and some first explorations of systems for facilitating that.

Another of the problems with the descriptions above is that there is no recognition of the powerful search tools that are available today, enabling their users to create ad-hoc sets/categorizations using any available characteristics of the items in a collection. Related to these are on one hand the data mining, knowledge discovery and visualization tools being developed that provide assistance in detecting patterns, or do it by themselves, and on the other the filtering tools for message streams and agents/bots for, e.g., roaming on the net. Pointers to resources on such tools are maintained by, e.g., Piatetsky-Shapiro (n.d.), Becket (n.d.) and Finin (n.d.).

The synergies possible in combining search/filtering and pattern discovery tools and agents/bots with methods for knowledge organization in building tools for the generation of systems for knowledge organization imply that the design and construction of systems for knowledge organization for specific domains can today be experimental and tentative. Another of the partial re-orientations of aims for knowledge organization is thus the recognition of the needs and the means for experimental approaches, for handling streams, for discovery, and for temporary and short-lived solutions—ad-hocness.

Yet another of the problems with the descriptions above is that knowledge organization and
the design and construction of systems is seen as a concern for a specific, small group of people, with the users using the resulting systems, but not participating in the design and construction processes, except perhaps by proxies, hostages. Knowledge organization, is, as has been argued earlier above, a matter for everyone. One of the consequences is that there is a need for yet another partial reorientation of the aims of knowledge organization, towards facilitating the participation of users in all stages of knowledge organization work. Until now users have been seen as users of finished results, i.e., applications of systems of knowledge organization to items and collections, that enable them (the users) to find items they need.

One of the observations from watching user behavior on Internet from a knowledge organizing point of view is that very many of them are engaged in activities aimed at generating structures and order, both at the individual, group, and collective levels. It is also notable that the tools and approaches used are, from a knowledge organizational point of view primitive, and without much coordination, see, e.g., Kempf (1995) for testimony from the "inside" of a specific domain, and the paper by Burnett (1993) which discusses networked communication issues in terms that are reminiscent of self-organization. What can be witnessed on Internet now could be called the second steps toward self-organization, or autopoiesis, see, e.g., Whitaker (1995) for one introduction and Whitaker (1996) for a guide to Internet resources on Enactive Cognitive Science & Autopoiesis. Such activities have for a long time been part of the scientific enterprise, manifest as the citation networks of scholarly communication. The evolving networks being built the citations in individual papers, like this one, have until the advent of the W3, been time-consuming to map and use. The only tools available (apart from the scientific literature itself) have been the citation indexes produced by Institute for Scientific Information, Inc. Now all users of Internet are building similar but more complex structures, and since the links on W3 are untyped, all kinds of content are part of the same structures, which from some points of view is good, but from other points of view a nuisance. One more partial re-orientation of aims that is therefore proposed is towards all the issues around self-organization as applied to knowledge organization at different levels.

Knowledge organization has also for too long been concerned with what might be called medium level structures, organization of knowledge at the levels of subjects, collections of items. The micro-structures of knowledge organization, at the levels of discourse, the internal structures of individual items, has, for many reasons, not been a mainstream concern in knowledge organization. The macro-structures of knowledge organization, at the levels of different forms of knowledge representation, the structures of collections of collections, have received somewhat more attention in the form of broad systems of ordering. The inter-level structures have received some attention in, e.g., citation studies for micro-medium levels, and, e.g., switching languages for macro-medium levels.

The formalization of document structure popularized through HyperText Markup Language—HTML, an application of Standard Generalized Markup Language—SGML, and the diffuseness of W3-documents have the consequences that questions of level become more apparent. The relations between document internal and external structures, the possibilities of using explicit internal structures in conjunction with other techniques for elicitation of knowledge structures at different levels also raise the issues of interpretation vs. extraction both in terms of applications of systems for knowledge organization to documents or document parts and in terms of the generation of systems for knowledge organization. Further along come possibilities and questions that HyTime as a formalism for describing objects and events in space-time brings. The last (in this paper) proposal for partial re-orientation of aims for knowledge organization is thus towards the uses formalisms for document architectures in a wide sense, another kind of formalism than the ontologies. Many of the documents that we encounter on the net in the future will carry
descriptions of themselves that are based on some of the formalisms available in Cover (1996) which is an excellent collection of resources on SGML and related projects and applications.

5. Proposals

A number of brief proposals are made below, based on the analyses presented above. These proposals are to be regarded both as elaborations of the earlier thoughts and as items for further discussion.

Phenomenology could be one of the philosophical bases for change oriented knowledge organization, because it sees our lived experience of the world as the foundation of meaning, and because of the emphasis on knowing as a temporal process, see, e.g., Sharoff (1995) and van Gelder (1996) for discussions of time-consciousness from a phenomenological point of view, and Rieu (1995) for a discussion of the interactions between information technologies (in a very wide sense) and the humanities.

Qualitative research methodologies should be attended to even more than they are at present. They often involve making explicit—generation and exploration of—conceptual structures that are or were latent, ill-defined or unclear. Software tools that assist in the processes of uncovering and building of such structures can and are beneficially interacting with knowledge organization in its earliest stages. At the recent 20th Annual Conference of Gesellschaft für Klassifikation e.V., March 6-8, 1996—the program is available at Classification, Data Analysis and Knowledge Organization (1996)—there were six papers on uses of or software support for qualitative approaches. Accept, in line with earlier arguments a few guiding principles:

- knowledge organization is for most people a means, not an end in itself, therefore
- tools (for individuals) rather than systems for knowledge organization are needed, see, e.g., Murray (1996) specific tools can, however, provide specific systems,
- tools that facilitate collective structure building and organization, to speed up the self-organizing activities being carried out anyway, are needed, hence
- meta-tools, tools for building tools, abstract and concrete, are perhaps the most important of the constructive activities to be undertaken within knowledge organization in the near future,
- plurality, many systems, at many levels, should be provided, see, e.g., Lethbridge (1994), who summarizes: “Users need a tool that involves the synthesis of several techniques for organizing the knowledge.“
- ad-hocness, ability to generate structures on the spot are needed, but also
- continuity and maintenance, assistance in adaptation to changes for those areas that have a persisting interest, are increasingly needed as a complement to the ad-hoc structures that persist,
- visualization, in many forms, for displaying structures and relations, will be increasingly important as more types of users start using tools and systems, see, e.g., Fowler (1996), Korfhage (1995) and Judge (n.d.).

Behind us we have vast bodies of literature and documents of all kinds, not properly described or categorized even in the traditional mode, which will not or cannot as such be transferred to the digital domain in the foreseeable future, but which will be catalogued and categorized for the new access tools, see, e.g., Bearman (1994) for a museum perspective. Ahead of us we have bewildering and competing technological capabilities generating increasing numbers and forms of documents, with the concomitant needs for organization. Right now we have
(digital) documents appearing and disappearing at a probably hitherto unprecedented rate. Some efforts are being made to at least make partial snap-shots of the Web, e.g., the project KULTURARW3 in Sweden in which the goals are to regularly download all W3 documents in Sweden, whatever is meant by that.

Our concepts, however, do not change as rapidly as technology, although we do have to find new terms and concepts for what is new. The changes of underlying concepts for terms that remain are much more problematic. Many of the terms we use today did not have the same interpretations, connotations and associations a hundred years ago. The explanation and mapping of concepts and terms in a temporal perspective, as systems of knowledge organization are modified and built anew is a neglected area that will be ever more important as the rate of change increases.

Lastly, the ethics of knowledge organization and the activities surrounding it are largely disregarded. The quality of systems, services, documents, information, knowledge rests in the end on the ethics of the participants, and today there is precious little assistance or discussion to guide them in their deliberations.

6. Conclusion

A discussion of the types of responses to the challenges of change alluded to in the title of this paper has by design been evaded so far. The challenges of change are actually twofold: change as a(n external) phenomenon to recognize and deal with, and change as proposition, a summons to action, to come to terms with, to accept or reject, or dismiss. This paper has until now mainly been an attempt at showing why change as a phenomenon now has to be accepted as a concern for knowledge organization. There are, nevertheless, also explicit exhortations to change, e.g., in the arguments about the need for re-definition and re-orientation of aims. The reactions to these challenges do not, however, have to be simple and pure, absolute.

The designations of the three types of reactions were all chosen to allegorize together both the river which is never the same, into which one never can step twice, and attitudes to change. (There are of course many more types of reactions that could be symbolized by catch phrases, viz., build bridges, catch and carry, dig dikes, etc.)

Go with the flow is a deliberate yielding to the path of the force, full acceptance of change as a flow, always new. Go with the flow can in the context of knowledge organization be taken as a response that acquiesces in change. The foci in concrete terms could be attention to here and now issues, questions such as: What’s up? What’s on now? News, data/information streams, edu/info-tainment, shopping, trends/front mapping, etc., are some of the domains of interest.

Abide by the side has been and is the customary response. Behind this response is a disregard for the ephemeral and evanescent, and a care for the long term and the durable. The goal is to discern stable patterns, something to trust and use in the years to come. The foci in this response are hence the traditional ones. The river will of course not cease to flow.

Watch the waves, which is partly inspired by the short story/chapter “Reading a wave“ in Calvino (1983), is neither an acceptance nor a rejection of change. It is an interest in change as a phenomenon.

All three responses to the challenges of change are valid, but the first, going with the flow and the last, watching the waves are the ones considered more important. Although going with the flow is the response that perhaps is more and of more immediate value to users of systems for knowledge organization it is probably the understanding of change, the watching of the waves, that will provide the long term benefits to knowledge organization.
Two dangers constantly threaten the world: order and disorder. (Paul Valéry)

Notes
1. An interesting overview of "how people have achieved their humanity in part by attaining a fuller comprehension of their own place in time and space" can be found in Clark (1992).
2. There are many documents on Internet about Internet and its various aspects. A starting point with many links on computer-mediated communication can be found at December (1996); IFLA also has a good collection of links to information resources on Internet at IFLA.net (1996).
3. Ironically, the paper does not have a date or time stamp!
4. In the context of this paper their Research Front Database and SCI-MAP Software System are of particular interest, see Research Services Group: List of products & services (1996).

References
Cambridge University Press.


Knowledge Sharing Effort public library. (n.d.) (http://WWW-KSL.Stanford.EDU/ knowledge-sharing/)


Research Services Group: List of products & services. (1996). Institute for Scientific Information,


http://shr.stanford.edu/shreview/4-2/text/sharoif.html

http://techweb.cmp.com/ia/0122issue/0122earthweb.html


http://www.acm.org/sigois/auto/Main.html


http://www.mcs.net/~jorn/html/ai/cycfaq.html
Bringing the Library of Congress Classification into the Computer Age: Converting LCC to Machine-Readable Form

Abstract: The USMARC Format for Classification Data was developed for many reasons including the potential to provide online interactive linking with MARC bibliographic and authority records; manipulation of data to provide machine calculations; support classification number changes of selected records on a daily or weekly basis; allow for validation, global changes and reclassification of classification numbers in bibliographic records based on classification authority records. The paper reviews the use, development, structure, and content of the standard. The implementation of the format has resulted in the conversion of all the LCC schedules into a database with MARC classification records for searching and maintenance. The database represents the first time that all subject classes have been brought together for searching across disciplines in LCC. Features of the system and the conversion effort are reviewed. Upcoming enhancements of the software will increase the system's ability to navigate for subject oriented retrieval and make it more efficient as a catalogers' tool.

1. Introduction

The standard format for machine-readable cataloging in the United States is USMARC, which allows for the communication of bibliographic and related information between computer systems. Institutions can thus exchange records in a standardized format, so that systems can predict record content and know how to process the records. The family of MARC formats include: bibliographic, authority, holdings, classification, and community information.

The USMARC Format for Classification Data (Library of Congress, 1991) was completed and approved by the American Library Association's Machine-Readable Bibliographic Information Committee (MARBI) in June 1990. The Network Development and MARC Standards Office of the Library of Congress (LC) developed the format, and in 1991 initiated an experiment to use the format in the conversion of classification data. The experiment was extended to cover additional schedules, and the Cataloging Directorate at LC made a commitment in 1993 to complete the entire conversion, which was accomplished at the end of 1995. The implementation of the USMARC Classification Format for LCC will result in many benefits both for the Library of Congress and for other users of LCC.

2. Potential Uses for Online Classification

Online classification data has many potential uses for information access. It provides the authority for classification numbers, terms, and shelflist information; be used for printing and maintaining a classification scheme; enhance subject retrieval; assist the classifier; facilitates maintenance tasks for classification numbers in bibliographic records; and provide the basis for an online shelflist.

Control and validation. Online classification data may provide authority control for the classification number and caption (a heading that corresponds to a classification number(s) and describes the subject covered). An authoritative file of classification records may be used by the classifier to assign classification numbers to bibliographic records. It may also provide a system with the mechanism to validate the correct assignment of classification numbers. In addition,
Online classification data can support authority control for synthesized classification numbers, i.e., numbers that have been made more specific by adding other numbers from a table or other parts of the schedule to a base number. A synthesized classification number need not appear in the classification scheme itself, since it is built by following add instructions, which instruct the classifier to add or append other numbers from the schedule or a table to a base number.

**Printing and maintenance.** Online classification data is an efficient method for printing a classification schedule. A print program that uses the USMARC records for publishing the schedules may have different system requirements than the program for online user display.

The *Library of Congress Classification* is an enumerative scheme, with new classification numbers inserted where appropriate based on bibliographic material when the need arises. The Library of Congress publishes *LC Classification - Additions and Changes*, which communicates new, changed, and deleted numbers and captions. Classification numbers are updated weekly.

Prior to the conversion effort, *LCC* consisted of some forty-six separate schedules, developed over a period of time by different people. It was designed as a shelf location and browsing device and has been maintained as such (Chan, 1990, 9). Over the years, the Library of Congress revised some schedules using word processing software, although many were not in machine-readable form. Most of the schedules, including many that are machine-readable (but not in MARC), have required editing to make the references and notes more consistent with one another and to allow for the classification number records to be input into the USMARC format.

Producing the *LCC* from an online file will facilitate the cumbersome process that the Library of Congress has used to publish revised editions. Generally the publication of revised editions of specific subject areas have been sporadic, and some are badly out-of-date. Online classification data will allow for additions and changes to the classification schedules on a timely basis and for the ability to include back-of-the-book indexing on a routine basis.

**Providing subject access.** In previous research, scholars and researchers have recognized the possibilities of classification data providing subject access to library material. The DDC Online Project demonstrated the usefulness of classification data for subject access, browsing, and display, opening up potentially powerful new search strategies (Markey and Demeyer, 1986). Explorations on the use of the *LCC* for online subject access have also been conducted (Chan, 1986). Online classification data enhances retrieval of bibliographic records by providing a different type of subject access, through a classed catalog approach, rather than through controlled subject headings. It enables the library user to see the interrelationships between topics and classification numbers, and facilitates browsing from more general to more specific topics and numbers.

Shelflist access, usually unavailable to the library user prior to the online era, collocates material regardless of its circulation status, oversize dimensions, or institutional sublocation. It thus has value even in open-stacks libraries. In addition, more precise searching of bibliographic records is possible through classification numbers for certain types of searches. For example, a well-known individual author or work may have its own classification number in the *LCC*, and a search by class number may retrieve a listing of bibliographic records on that author or work more efficiently than a traditional author or title search. Having the classification data accessible online will facilitate this process.

**Assistance for the classifier.** Online access to classification numbers may save time for the classifier. The classifier may perform a keyword search to quickly locate the possible classification numbers through captions, notes, or index terms. The classifier may be able to trace the formation of synthesized numbers to assist him in classifying. In addition, online classification may assist the classifier in obtaining consistency in the assignment of classification numbers to library materials. The Library of Congress is exploring the integration of schedule and table numbers in its implementation so that the system provides the calculation function formerly done by the classifier.
Applying add instructions or tables to classification numbers and validating the accuracy of synthesized classification numbers can both save time for the classifier and increase accuracy. In addition, the system can provide online links between classification records and subject authority records both to facilitate the subject cataloging and classification function and to provide greater subject access to library material.

**Maintenance of classification numbers in bibliographic records.** An online system could use an automated classification database to facilitate a library's conversion to a different classification scheme. By maintaining the classification schedule online, an institution may be able to perform global updates of the bibliographic file when a classification number is added or changed. The classification format allows for tracing the history of a number and to provide specific links to the new or old number.

**Providing a basis for an online shelflist.** The online classification system provides a basis for an online shelflist, and is a building block for the assignment of item numbers. An automated shelflist would be a valuable resource for bibliographic material, especially if classification records can be linked to bibliographic records.

3. USMARC Classification Format

**Background.** Although classification schemes are textual in nature, there are several advantages to using a MARC format rather than a format intended to handle large amounts of text, such as Standard Generalized Markup Language (SGML).

Library of Congress Subject Headings (LCSH) serves as a useful model for classification data. LCSH are input into separate USMARC authority records with tracings provided from unused to used forms and links to broader, narrower, and other related terms. Tracings are an efficient method for assuring database integrity, as references for changing headings do not need to be found and changed in many places. LCSH are updated on a daily or weekly basis at the Library of Congress and historically have been controlled by LC. Essentially a database of USMARC classification records can provide the same type of control as name and subject authority records do; references can be supplied as tracings from unused to used numbers and terms. Changes are made to classification numbers and captions on a daily or weekly basis as well. Thus, the ability to communicate changes on a record-by-record basis, rather than an entire database, is essential.

Online classification data can be linked to other USMARC data. The consistency in structure and data elements between the USMARC formats for bibliographic, authority, especially subject authority, and classification data facilitates the linking between files. Most large library systems know how to process and interpret MARC data. Integrating classification records into this structure can provide advantages for future use of the records. In addition, for any interactive links to occur, such as validation of classification numbers in bibliographic records, global changes, or subject term and number comparisons, using the data in the same format will facilitate the process. MARC has proven to be flexible, efficient, and easy to maintain for library automation systems (Crawford, 1989, 7). It is used worldwide for storing, sharing, and manipulating bibliographic information by computer.

The Library of Congress is planning for a new MARC record distribution of classification records, based on the model of MARC bibliographic and authority record distribution. **LCC** schedules will be distributed in USMARC records, and, eventually, once conversion is complete, the entire scheme can be so distributed. Local institutions that choose to acquire the records can use them in local systems (assuming local implementation of the classification format). Those institutions will then be able to create local classification records, since many may wish to expand
certain areas of the classification, or provide local notes. Using the MARC format, institutions could create individual records or notes in existing records as needed locally, rather than be limited to using only the records as distributed by LC. They would then be able to communicate in a standard format these local records to other institutions if desired.

Structure of the Format. Content designation in the USMARC Format for Classification Data, as in the other USMARC formats, are codes and conventions used to explicitly identify data elements in a record. The goal is to characterize the data elements with sufficient precision to support manipulation of the data for various functions (Library of Congress, 1989, 4). Functions supported include display, both formatting in an online display and in producing a printed or other type of product, and online information retrieval. How an institution displays the data is not specifically covered in the USMARC formats, although they do provide for display constants, which are terms, punctuation, or spacing that are system generated for display. For instance, the hyphen separating the beginning and ending numbers of a classification number span is not carried in the record but may be generated as a display constant by the display system based on the structure of the classification number field of the record. The system itself, that is, the implementation of the format, will largely determine how well those intended functions are met. It is difficult to predict the limitations a system might encounter in using the data; however, the format itself has allowed for great specificity in coding to maintain optimal flexibility and to satisfy likely uses.

Similarities to other USMARC formats. The USMARC Classification Format retains much of the structure familiar from the USMARC bibliographic and authority formats. The Leader is a required element which defines the parameters for the processing of the record, and the Directory gives information on the fields, field length, and starting and ending character positions in the record so that a system can process a file of multiple records. The format has a 1XX block of fields, functioning like heading fields in authority records. Most records include a field 153 (Classification Number) that contains the authorized classification number and caption representing a topic, as well as the full caption hierarchy. In rare cases a field 154 (General Explanatory Index Term) can be used instead for terms that cannot be contained in the record for a classification number or span. The structure of reference notes (2XX and 3XX) and tracings (4XX and 5XX) mirrors the USMARC authority format and is used in a similar fashion. The 6XX block of fields contains notes, and tag names from the USMARC authority format that apply to classification have been retained. The 7XX block retains the structure of the USMARC bibliographic format added entries (6XX in bibliographic records), allowing for including in the classification record controlled authority headings when there is a relationship with the classification number. These fields are intended to supplement terms contained in other parts of the USMARC record, such as the caption data in 153, for additional subject access to the classification number.

4. Online Implementation

Online system. In 1992 the Network Development and MARC Standards Office initiated an experiment to test the new classification format, which resulted in the conversion of the H schedule (Social Sciences) to the USMARC format. The project used PC-based software called Minaret, which can create MARC records. Since Minaret allowed for the user to define data elements, indexes, and display forms, its flexibility proved essential for experimenting with a new format. Because the individual record approach does not adequately represent classification information, which must be viewed in its hierarchical context of subject terms, the developer of the software prepared an enhancement for LC to enable the user to access a classification browse
display. This display brings the classification records together according to defined specifications to display the data on the screen in a format similar to the page of a classification schedule, showing relationships between numbers and captions.

The Minaret program for classification at LC is now a UNIX-based multiuser system running on a minicomputer that is accessible to LC users using remote login (telnet). It fits into the LC environment of client/server applications and can be accessed through LC's Bibliographic Workstations within the context of the catalogers' other work. Thus, catalogers can access the database in one session while working in other LC internal files. Potentially they could search across databases or copy data from one to another. The system is currently not available for users outside of the Library of Congress.

The program provides links to other records to facilitate use of the classification schedule. The user can jump to another number referred to through the tracing in the record by pressing the enter key. In addition, the user can access classification tables when the record refers the user to a table or subarrangement, allowing him/her to create a more specific classification number.

Several indexes are available to search the schedule: classification number, record control number, classification caption, index term (primarily those that would appear in a printed index), and a combined caption/term index. In addition, a keyword index has been recently incorporated, searching terminology in captions, index terms, and notes. The indexes provide quick access to numbers and topics as well as the potential for effective maintenance of the schedules on a frequent basis.

USMARC records created in Minaret can be exported as MARC or ASCII for use by other systems or programs.

Conversion. The conversion effort continued with outside contractors delivering converted USMARC records for several schedules between 1993 and 1995. Staff of the Cataloging Policy and Support Office converted the remaining schedule, also providing subject expertise in those areas. As of the end of 1995, all schedules (except for a few subclasses) had been converted and incorporated into a master classification database.

The Library of Congress has undergone the conversion of the LCC schedules very quickly. The effort of automating over forty volumes of classification data with more than 16,000 total pages in a relatively short period of time was immense. The conversion began as an experiment to test the format, and gradually grew into the reality of USMARC records that would be used for printing, online display, and distribution.

LC made the assumption early in the project that it would be easier to modify the data in the schedules once it had been keyed. Thus, inputting staff has generally revised schedules only where needed to make the information explicit or to correct obvious errors. One exception has been the J schedule (Political science), which was so outdated that LC performed a thorough revision before input into the online classification database. Staff is reviewing the data as the USMARC records are used by the print program developed by the Cataloging Distribution Service for new editions of separate schedules. Although all schedules have been converted, there will be a delay before the data is distributed because of the review process.

A study of LCC prepared by Nancy Williamson has assisted the Library of Congress in identifying problem areas. The conversion experience confirmed the need for many of the recommendations given in the study (Williamson, 1994). In many cases LC will need to investigate further to consider how much revision is needed to make the machine-readable data more usable. Daily use of the growing classification database is providing valuable experience for determining even better ways to access and utilize the data.

There are specific problem areas that will require further study. One is the application of table numbers and other types of subarrangements. In the original schedules, some tables
appeared at the end of the schedule, others within the text, and yet others by footnotes referencing other locations. In addition, how those tables are applied to a number has been inconsistent from one schedule to the next. In some cases the number from the table is appended to the base number; in other cases it is necessary to perform addition to calculate a number resulting from adding from a table to a base number.

Because it is highly desirable for a classification system to perform the function of synthesizing numbers (and all the information to do so is already present in the records), it will be necessary to make the data consistent so that rules are predictable. In addition, most of the subarrangements that were coded according to the USMARC Classification Format as internal tables (embedded in records in repeatable 763 fields) during the early stages of the conversion will be changed to separate table records to facilitate a calculator function.

Another problem is the consistency in the use of index terms. The indexes to the schedules provide rich terminology, but they vary greatly because schedules were developed individually and revised at different times. The depth of indexing from one schedule to the next is dependent upon the subject matter and the perspective of the indexer (Williamson, 1994, 10). As the different schedules are being brought together into one master classification database these inconsistencies have become obvious. A thorough review is needed both within and between separate schedules. LC is exploring these indexing issues.

Cataloger use. As part of the initial online classification experiment one cataloger on the Business and Economics Team of the Social Sciences Cataloging Division has used the classification database daily for classification of new material since 1992. The cataloger experimented with enhancing the database with additional index terms related to classification numbers. As the online data has been reviewed and approved, catalogers in other subject areas have begun to participate in the use of the system. Responses have been generally positive to the use of online classification data, and use of the system will be extended to all catalogers responsible for applying classification numbers.

5. Software Enhancements

The Library of Congress plans to upgrade the Minaret software being used for the internal classification system. Some specific enhancements include: a classification number calculator, which will allow the system to create synthesized classification numbers by calculating the result of adding a table number (or other subdivision) to a base classification number; allow for the user to jump to numbers in notes, which will provide additional functionality in navigation among classification numbers; separation of schedules so that catalogers can isolate portions of the entire scheme; graphical user interface to enable more flexibility in the presentation of data; cataloger notes and bookmarks, enabling a cataloger to create personal notes in his/her "copy" of the schedule; mapping classification numbers to subject terms, to correlate subject terms in bibliographic records with classification numbers; and, expansion of hierarchies, enabling the user to display hierarchies at each level and expand them as needed.

6. Distribution of Classification Data.

The Cataloging Distribution Service (CDS) has developed a print product using the USMARC records. In late 1994, it published the first schedule using USMARC records, Schedule H (Social Sciences). By the end of the first quarter of 1995, new editions of seven new schedules had been published, with nine more in the process of review. Once a schedule is published it will be maintained both online and manually for a period of time, until review of all schedules is
completed. Since additions and changes are made to the LCC classification schedules on a weekly basis, the USMARC records in the classification database are changed accordingly. The new publishing process will enable CDS to publish revised editions of schedules on regular basis compared to the previous manual effort.

CDS has recently issued a CD-ROM product called Classification Plus, which includes all schedules that have undergone review with links to Library of Congress Subject Headings. In addition a MARC record distribution is planned to begin in 1995. It is expected that other institutions, system vendors and/or utilities will develop classification systems, using the MARC record distribution for most of the data.

7. Conclusions

When the USMARC classification format was completed, doubt was expressed that the Library of Congress could convert the massive amounts of data in the Library of Congress Classification schedules in a reasonable amount of time. Now that the endeavor is largely completed (with the exception of full review), the challenge becomes the effective use of the data. There is great potential in utilizing the wealth of information in the online schedules effectively and efficiently. Designers of future retrieval systems for classification records need to utilize online classification as a powerful tool for subject access, the maintenance of classification schedules, machine-assisted classification, and validation of assignment of numbers in bibliographic records. In our age of increasing use of networked information resources, classification data can provide an alternative approach for retrieval of information, using a subject oriented approach, and thus ensuring that language is not a barrier to finding needed information.

References
Library of Congress Classification: Shelving Device for Collections or Organization of Knowledge Fields?

Abstract: In the literature on cataloging and classification published over the past 70 years, the Library of Congress Classification (LCC) has been described as a generally acceptable "library shelving device" and "shelf browsing mechanism" (Chan, 1990, 15). Indeed, the various classifiers never claimed that they developed LCC as a scientific system but rather as an utilitarian tool. Likewise, its critics rarely ever concluded that it is, in fact, a knowledge-based logical system, in which the records of a literate culture are organized: the various manifestations of recognized knowledge fields, corresponding to their scientific framework at a given point in history. With inception of the last LC Class, K: Law, whose design was no longer retrospectively tied to existing holdings of the library, classification has further evolved into a coherent concept classification. By freeing LCC from the constraints of its earlier function as shelf locator for American libraries (Chan, 1996, 9), its potential as a unique retrieval and navigation tool for electronically-stored bibliographic information will be realized.

1. The Genesis of the Library of Congress Classification

The LCC should not be contemplated as isolated from the general history or intellectual climate in the country, since the political history explains what is written and why, while the intellectual/cultural history explains what is collected and why, and is organized the way it presents itself today. It was no coincidence that the classes for Political History, Political Sciences, and Library Science were the first ones to be developed at LC.

1.1 Collection and Organization during the First Period ca. 1801-1861

In the United States, the periods of political history and historiography coincide with the first growth period of LC collections, topped with the introduction of a new general catalog in 1861.

After the war for colonial independence, historical accounts show the colonies' progression from colonial independence to a nation. For that, one needed a common politico/legal system, but mostly: the creation of a common heritage, a common national history. The organization of a common bond were the efforts of a few men. Histories which they wrote subordinated the role of the individual colony or state to the story of the nation: recording and explaining the past of the country as a whole. The history of the Revolution was also to provide its justification: restoration of lost liberty and lost rights (Van Tassel, 1960, 32-34).

The years particularly interesting for the type of collections then forming at LC, were ca. 1815 to 1860, dominated by the popular historians; writing was guided by the desire to see the U.S. acquire "American character", to teach patriotism and to educate; the latter interest was stimulated by the flood of immigrants (to 1848) from the "old country", bringing advanced education which led to: building of schools, a big market for the flood of text books, and development of a curriculum of which American history was a central part (Van Tassel, 1960, 87-89). At the same time, one encounters the steadily rising number of local historical societies, including those in the new territories (frontier) which provided the historiographer with collections
of personal papers; accounts of lawyers, business men, pioneer settlers, colonial offices, etc.; and
treatises on American institutions, including the frontier. Another phenomenon was the so-called
Documania, the scramble for documentation. Collections that are building include government
documents, treaties, laws, and legislative journals, official/diplomatic correspondence, all “Sources
of history to its truth”. In all this, the original colonies competed for acquisitioning of colonial
documents from abroad (Van Tassel, 1960, 107).

The years from about 1830 on may be characterized as the years of the Romantic
Nationalist. His mission was to spread the saga of freedom and hope of mankind for a better
world. This period coincides with the younger (Heidelberg) Romantik in Germany: the academic
youth movement during the 1848 Revolution; the growth of German nationalism (Vaterland) and
early constitutionalism in German states; the revolutionary civic response to absolutism (as
someone observed: George Bancroft's translation of democratic faith into German romanti-
cism)(Van Tassel, 1960, 118-120); and, on the academic side, the cresting of the "Historische
Schule".

1.2 The First Growth Period of the LC Collections
By 1861, with issuance of a new general catalogue, the collection had grown from 152 works
to 79,214 volumes. We know from various accounts, such as annual reports, memoranda, and
laws, what was collected, and from criticism of the collections of what was not collected. From
1808 on, collecting concentrated on laws and Congressional papers; in 1815, Jefferson's collection
was acquired; since 1817, LC received copyright deposits; in 1836, the Secretary of War (in an
address to the American Historical Society) advocated expansion of LC collections to all subjects
of human learning "...to elevate it to an equality with those great repositories of knowledge which
are among the proudest ornaments of modern Europe" (Cole, 1979, 16); since 1837, the Joint
Library Committee supported international exchange of public documents; and, in 1853, Congress
rescinded the law for international documents exchange (Cole, 1979, 23).

1.3 Early Organization of the Collections
Before Congress moved to Washington and established its' Library in 1800, Congress used
the collection of the Library Company of Philadelphia. In Washington, a catalogue for the 900
plus works was issued in 1802, followed in 1808 by the third catalogue which, besides size,
introduced for the first time forms as plans, state laws, journals of the House, House reports,
executive papers, and gazettes. By 1812, the first classed catalogue was issued for the growing
collections. It was the first subject approach according to the catalogue of the Library Company
of Philadelphia which was used by Congress before the move of the government to Washington.
The Philadelphia scheme was modelled after Francis Bacon’s System of Knowledge Classification
(1605, modified by Jean le Rond d’Alambert, 1751 (Chan, 1990, 3-6). Before applying it to the
approximately 3,076 volumes still subarranged by size, LC reduced the number of original classes
from 31 to 18 (Cole, 1979, 6-10). Thomas Jefferson’s classification scheme which had been
introduced in 1815, was applied at LC in expanded form until the new classification was
developed.

1.4 The Second Period of LC Collection Building and Organization, 1861-1910
In this time frame two different approaches towards the nation’s history occur that deeply
impact LC’s collection and classificatory policies. Critical historical exploration (ca. 1866-1884)
established after the Civil War the national past as the basis for the reunion, which in turn became the justification for the Civil War. The scientific (seminar) method of study and writing of history came back to the US around 1884 with scholars trained in German seminars, where “government and international law” formed part of the history curriculum. This so-called new historical movement viewed history as a study and record of social evolution. The best of these scholars, viewing themselves as political scientists, would establish “Departments of History and Political Science” at American Universities. It is interesting to observe that the American Social Science Association lent its authority to the foundation of a new organization: the American Historical Association (Van Tassel, 1960, 171-176).

The expansion of the LC collections in this period was based on congressional requests for deposit of all documents by the states governors (1866), the formalized exchange of foreign government documents (1867-1875), and establishment of major collections either by bequest, gift, or congressional appropriation, among them the Chinese, Turkish, and Lincolniana collections (Cole, 1979, 30-42).

These historic accounts must be seen against the general historico/cultural background of the second half of the century, the citizens’ century: Humboldt’s declaration of Freedom of Learning and Teaching in Germany, accompanied by major achievements in the arts and sciences; and archeological endeavors of the German and the English in the Middle East, deepening the interest in antique/Hellenic studies and Roman law. Under Bismarck’s leadership, the civilian government had taken on the mandate for education and welfare. A critical factor was the wealth-paired with the German Bildungsideal—of the new industrial upper and upper-middle classes, the bourgeoisie, which were to lay the foundation for public collections, mostly museums and libraries, as well as botanical and zoological gardens, open to all citizens. This was the time when the best of America’s industrial upper class had their agents for buying and acquisitioning in Europe: the DuPonds, Frick, Vanderbilt, Morgan, and the Pittsburghers, to compete with the Continent. Most importantly, LC also had a permanent purchase agent in Europe (Cole, 1979, 25). And in 1892, the President approved a Congressional resolution to finally open the Library of Congress to the public.

It was around this time that the Library of Congress—with collections up to the one million mark—had undergone a major reorganizations into several departments, and Herbert Putnam (Boston Public Library) and Melvil Dewey (New York State Library) testified before the Joint Library Committee on behalf of the American Library Association emphasizing the need for a new classification system (Cole, 1979, 51). In 1897, with the new Library of Congress building ready for occupancy, Charles Martel (new Superintendent of the Catalogue Department) began preparation for the intended reclassification and exploration of available schemes, joined by J.C. Hanson, previous Superintendent of the Catalogue Department (Cole, 1979, 59-60). In 1898, during these preparations, Librarian Young proclaimed LC’s mission: to collect “whatever illustrates American History . . . varied forms of American Growth, Theology, Superstition, Commonwealth, Building, Jurisprudence, Peace and War” (Cole, 1979, 64).

1.5 Development of the Library of Congress Classification (LCC) to 1948

After his appointment in 1899 as Librarian of Congress, Herbert Putnam focused on the development of the collections and of a new classification: according to his first statement to Congress, he found the collections to be defective and the classification meager, rigid and inelastic (Cole, 1979, 67).

The Dewey Decimal Classification was ruled out in its' present form.

The classification Schema of the University of Halle (Germany), originally studied, was also
ruled out because it seemed "too strongly oriented on traditional German philosophical thought".

Charles Ami Cutter's Expansive Classification was selected as the prototype, however, with modifications in the notation structure, using a combination of letters and numbers (Chan, 1990, 21). The comparison between the various outlines for the LC Classification created up to 1904 and Cutter's classification, show that Cutter's main classes do not have any relationship to Bacon's or d'Alambert's systems, nor to the Jeffersonian outline, all Europe-oriented in their comprehension and division of knowledge into the principal three classes, history, philosophy and fine arts (poesy). Cutter already had separated Science, Mathematics (Pythagoras, Heraklit, Anaximander, etc.), and the Law (Jus naturae et gentium) from Philosophy. Hanson went beyond that: he broke up Class A (Philosophy and Religion) and, introducing double letters, created Classes B-BJ for Philosophy, and Classes BL-BX for Religion and Theology. Classes J-JX in the early version of LCC are already the Classes for Political Science. Class K: Law, had not been retained as a distinct Class but was incorporated into Class H, then Political Science. This was a troubling approach for later developments.

First to be drafted was Class Z (Bibliography and Library Science, 1898) for critical materials in preparation of the massive reclassification (Cole, 1979, 64), but first to be published (1901) were the Classes for the most extensive collections, History Classes E-F (History and Geography of the United States), followed by the draft of D (History of the Old World). Political Science Classes J to JX, in company of Naval and Military science (V and U) and some other classes, were published in 1910. This shows clearly two things: the follow-through with the mission statement of Librarian Young (American orientation) under Putnam (LC as "a bureau of information for Americana" [Cole, 1979, 70]); and the prevailing philosophical comprehension of history as an all-inclusive field in subject matters because historical studies have, in fact, provided answers to contemporary calls, closely tracking political, social, and economic developments. Thus, Classes E-F included boundary questions and treaties, both relating to the narrower subject of a state's territory and sovereignty; the recorded manifestations of U.S. westward expansion and territorial dominance over the Indian territories; works dealing with war and peace and the peace treaties as well; a whole regional development leading eventually to the formation of the OAS, a subject belonging by definition to International law; and geography. It also shows the encyclopedic character of the LCC from the beginning: to round out the schedules with information for the user, detailed encyclopedic notes introduced and illuminated each major chapter. This practice was retained for the 2nd edition (1913). By 1958 (3rd edition), all these notes were removed.

Legal history was never recognized as a discipline per se but formed part of general history. Therefore, eminent historic-legal sources were classed in Class D. Classes J-JX: Political Science, had absorbed official gazettes (a primary source of the law), legislative papers, texts of constitutions of the world together with constitutional history; and, in accord with the understanding of the time, international law was welded together with international relations, instructing the cataloger: in case of doubt, prefer D-F. Class K: Law, although in the final outline of the LCC reflected as a class, was never developed. Instead, many of the component classes of the LCC completed by 1948, had absorbed legal materials that one would have expected to serve as the basis for Class K. Law was considered a congeries of aspects of other disciplines and not a discipline in itself; in fact, it had become a form of other subjects (in particular Classes H, L and R) (Goldberg, 1986, 329).

Evaluation of the new classification proves that, despite various interpretations of statements uttered by library officials, LCC was from its inception a composite system: organization of a knowledge field by design, and library organization by application to the shelf. Richard S. Angel, Chief of the Subject Cataloging Division, in 1964 clearly distinguishes between LCC as a "complete system, embracing all areas of human knowledge, the various components of this
universe of knowledge having been allocated to the various schedules corresponding to welldefined areas and concepts by which the separate fields are taught and expounded, and on which developmental research is based", versus the provision of an "orderly arrangement of volumes which makes access to the collections useful and meaningful to qualified students, scholars, and staff" (Angel, 1964, 353-354).

This was, however, merely a restatement of the observation of F.H. Wagman (Director of the LC Processing Department), already made in 1949 at the AALL Annual Convention, that "the Library of Congress must maintain an encyclopedic classification of knowledge" (Goldberg, 1986, 332). Translated into theoretical terms and classificatory structure, this means: systematic-hierarchic chain of concepts or textual elaborations of a knowledge field or topic condensed into captions and arranged from the most general (General works) and broadest to the most specific and refined term or definition in descending order, a logical, deductive method grounded in late 18th century European philosophy (Wieacker, 1967, 193; the principal representant is Christian Wolff, the head of German Enlightenment philosophy). Applied to the LCC it means, a systematic arrangement of related concepts or subjects, expressing by way of indentation their relationship in the hierarchical structure to each other.

The adopted alphanumeric notation system for the LCC is a composite system utilizing letters and Arabic numerals. Main classes are denoted by a single or double capital letter, for classes D and K also triple letters denote subclasses. Within each main class or subclass, the integral numbers 1-9999 are used for subject division, with generous amount of free numbers for future expansion. This Class number assigned to each caption is the numeric expression (code) of a phenomenon, fact or intellectual information defined by natural language or subject-typical terminology in the caption. In other words: it is coded information denoting the order of subjects in the schedule. It was originally decided not to use decimal numbers, later, the decimal extensions of integral numbers were introduced for new subjects. The addition of a "work specific" Cutter number for the book in hand (the "book Cutter") on a subject outlined in the schedule converts the systematic class number to the call number (or shelf locator). After the first set of letters and numbers, the second set, the Cutter numbers, preceded by a period, follows. Serving here, as example, are the Rules of procedure of the European Parliament:

| Class number | KJE       | [=Community Law] |
|             | 5390      | [=Integral number for European Parliament] |
| Subarrangement | .A18     | [=Form Division IX for Rules of procedure] |
| Book number  | .E97      | [=Main entry] |

Extensive referencing from the systematic location of fact, event or topic to the actual location for bibliographic reasons is introduced and widely applied from the first Class on.

2. Final Stage of Library of Congress Classification Development. Interdisciplinary and Comparative Classification

The Annual report of Librarian MacLeish, in 1940, states that the Library's collections are "preeminent in American history, bibliography, library science, publications of learned societies, economics, political science and social science", but "weak in specific fields of European literature, social sciences, history, education, modern anthropology, and most technology" (Cole, 1979, 110). This marks the beginning for new collection and selection policies, the new canons of public services, followed through by Librarian Luther H. Evans. His Library of Congress "Mission in Europe" for obtaining "multiple copies of European publications for the war period" for
distribution to American libraries and research institutions accounted for an unprecedented expansion of LC's collection, due to the massive inflow of foreign, predominantly legal materials between 1945 and 1949 (Cole, 1979, 121). Of special importance were German laws and all related materials for the study of Germany's past role in the European theater, now the Territory under Allied Occupation. This would finally lead to development of the last component Class of the LCC: Class K for the law.

2.1 Policies and Principles for Law Classification Development

In 1949, the LC Classification Committee recommended to the Librarian to develop the law classification scheme based on the methods and principles that have governed the construction of all other schemes.1 In May of the same year, during the Annual Convention of the American Association of Law Libraries (AALL), the guiding principles for Classification of Law were worked out, later to be published as the *Interim Report of June 10*, which has remained the governing document for the development of Class K, accompanied by a first outline of the entire Class. The report defined for the first time the types and categories of what constitutes "law materials" and has since set the demarcation line, retrospectively and forward, between Class K and other Classes. However, opinions were greatly divided over several issues: (1) should the schedule be based on the actual library holdings; or (2) should it be developed on a strictly theoretical basis? The latter was ruled out by the Committee on Development of Class K, since the Library had considerable experience in classifying existing collections. (3) From the outset, it was not certain, whether LC would ever make a commitment to reclassify its legal materials.

The development of the first K class was caught up in this point-counterpoint argument for 20 years, both outside and inside LC. The resulting shift in development patterns was due to the shift in classification policy. Firstly, the planned creation of a "virtual law collection" on which the development of Class K would be based, i.e., a shelf list consisting of new acquisitions and bibliographic data, extracted from the shelf list of Classes A-Z for the older materials buried in such classes, was abandoned as "dragnet operation". Subsequently, for the first time in classification development at LC, reclassification of an entire collection, the retrospective conversion of law materials, was ruled out. This policy is since restated in each published K schedule. Secondly, highest in the order because it governs the structure of Class K, was the *jurisdictionality* policy, which would provide the outline for all its subclasses. It was recognized that the differences in legal systems, nomenclature, public policies and intellectual tradition from one jurisdiction to another would not allow for superimposing of concepts and nomenclature valid only for one jurisdiction onto another. Thus, arrangement by jurisdiction takes precedence over other criteria such as subject. Jurisdiction, therefore, will form the principal hierarchy (Goldberg, 1986, 332-334). Thirdly, already with the first group of classes for the Common law countries, classificatory technique had resorted to model schedule development (the first schedule, Class KF: Law of the US was the model for the other Common law schedules and tables), however, only in approximate, broad use of patterns without a common number base, and not strictly symmetric. With creation of KK-KKC: Law of Germany, the model for Civil law jurisdictions, the derivation technique was introduced, the use of one schedule as the number, pattern, and, as far as possible, terminology pool for creation of a related Class, KKA Socialist Law of East Germany. For that, comparative study would determine common principles underlying both systems; as a result, harmonized/analogous class detail based on the legal doctrine shared by both jurisdictions could be created. As an example from constitutional law:
Fourth, the *regionalism principle* was an outgrowth of a further restriction in development of the class: not all jurisdictions could be accommodated with their own classification. Thus, another formula had to be invented to create some organization for the rapidly growing foreign law collections.

Customary, a region has been defined (geographically) as an area in which historical, religious, and socio-economic or ethnic similarities, as reflected in laws in the area; and where cultural, scientific, and economic interests have led to regional integration and organization. In both cases, further refined model schedules and symmetric uniform tables, applicable to the largest number of jurisdictions in that region, were the product of comparative pattern and concept analysis.

This principle had been adopted for the first time during development of the schedules KJ-KKZ (Law of Europe) (Goldberg, 1987, 69-72) and applied since to all regions of the world (Europe, Asia, Africa, Pacifica). Harmonization and approximation technique was applied to develop a jurisdictional schedule for regional organizations as well, first for the European community, and—further refined—for the Antarctic Regime and other international organizations (Library of Congress, 1993, Introduction; Goldberg, 1993, 88-133). By these methods, second and third generations of schedules or tables could be derived from one model.

### 2.2 Perimeter of Classes. Adjustments. Revisions

Implementation of the vast Class K, stage by stage, has demonstrated that classification is the conceptual structure and complex policy framework for development and maintenance of large collections. Classification sets not only the perimeters within a particular class for its subclasses, but delineates the demarcation to other classes, especially to neighboring classes as H (Social Science) and J (Political Science). It is in the overlap areas, the "grey zones" of classification, that policies emerge as consequence—or prelude—of periodic major changes as new knowledge fields emerge or established ones mature; such occurrences have been handled not only by numeric expansion in a particular classification for new or specific aspects, or by refinement of the scope of a class, but by major shifts of whole subclasses (cancellation and merger of one subclass into another) and extensive re-developments. Although LC never encouraged the use of LCC as a general American classification, it could not prevent that LCC in fact became the preferred library classification. The voiced concerns by libraries which adopted LCC for "stable shelf arrangements", prompted Library officials repeatedly to restate: that adopting libraries must sufficiently understand those aspects of the LCC itself, its development, revisions and application (Welsh,
The example of Class K also demonstrated that only after freeing the development of the class from the strictures of reclassification of existing collections or conservation of existing number structures, could the Class mature to a system for organization of a knowledge field. The freedom to borrow principles developed by legal scientists—the high level of abstraction and the working method of formal logic—has allowed for a strong hierarchical order of concepts and subjects. Techniques, as analytical exegesis, and tools for comparative study, as approximation, harmonization, analogous interpretation, and synthetic construction (guided by references to parallel hierarchies), were kindred methods adopted for the development of Class K.

The detailed enumeration of topics and definitions, represented in the captions in scientific language is another important factor: the more advanced or experienced the expected user is thought to be, the more detail is commanded by the index (Moys et al., 1993, 2, 19).

2.3 The Electronic Version of the Library of Congress Classification

Several factors in the most recent history have made new in-depth study of the LCC necessary and have indeed lead to a better understanding and renewed interest in this classification system: the conversion to the USMARC Format for Classification; the needed standardization for indexing procedures before the merger of indexes to the classes A-Z into one integrated LCC index; and identification and harmonization of common underlying patterns in related schedules or groups of schedules in preparation for, or during, revision of the LCC.

Most important during the conversion of LCC to the electronic format was the evaluation of existing or projected classificatory strategies, including model creation for common divisions (form, geographic, or period) and concept/subject divisions, based on experience gained by development of the newest class, Class K. Pattern adjustments, the harmonization of old and new or parallel patterns and pattern conformity, realignment of hierarchies, standardization and updating of terminology, and generous introduction of LC subject headings into the captions are all critical and usually do not effect number structures of old existing collections.

It is interesting to observe that the last two K subclasses that close gaps in the LCC, will rectify classification practices rooted in the ideological perception of the "history complex" and policies dating back to 1901. The two new schedules: KZ (Law of Nations) and JZ (International Relations) will—upon implementation at LC—substitute for the schedule JX (International Law. Foreign Relations. Diplomacy), the last subclass of J: Political Science. JX, first implemented in 1910, has been devoid of many pertinent subjects that have since been incorporated in the history schedules D-F.

Several classification techniques which allow for comparative classification and approximation either of two classes or of particular subject fields (subdivisions) in related classes have recently been tested during the development of two new classes, JZ and KZ. To finalize the design of the new schedules, a thorough JX collection survey based on subject/pattern analysis of the JX schedule was produced. Simultaneously, an extensive shelf list examination of the History classes E-F (America), and Class D (History General and Old World) was conducted to recover subjects for incorporation in the new schedules, while numbers in the schedules D-F will be closed. The subjects stemming from materials classed originally under the old JX, and legal subjects in schedules D-F to some degree, are now distributed between the two new classes.

The two schedules are devised in subject arrangement and number structure, so as to compliment each other, allowing for discretionary placement of works, especially document collections, in either class depending on an individual library's point of emphasis or collection
policy. The old JX can serve as a concordance (and at LC: shelf locator, since it is not envisioned at present, to re-label and move the collection) while the new numbers—after they have been linked to the original JX numbers—will eventually allow for systematic search and retrieval.

3. Conclusion

Despite such statements made by Librarians of Congress (in particular, Herbert Putnam, at the inception of the LCC, and Luther Evans, during the planning state of the Law Classification) that the LCC is not intended to serve as a general classification for American libraries, the LCC, in fact, has become the preferred classification for libraries and educational institutions in the United States and for many institutions abroad. Their voiced concerns for stability of shelf arrangements by classification has, in the past, markedly inhibited the inventive remoulding of the LCC. Ultimately, all efforts will have to concentrate on the envisioned function of the electronic LCC as an on-line retrieval tool. For on-line browsing and navigation of electronically stored information, including the segregation of whole portions of one class and transfer to another, a knowledge-based, field-specific structure of the classification is of utmost importance. So, also, is the separation from the shelving function.

Notes

1. In Spring 1949, the Librarian of Congress appointed a Committee on the Classification Schedule for Law to study theoretical and administrative problems. At the same time, AALL appointed a Committee to Cooperate with LC on Law Classification. In May 1949, the two committees developed the principles for the Classification of Law, to be announced during the 42nd Annual Convention of AALL of the same year, and henceforth known as the "Interim Report of June 10, 1949."

References


Snunit Shoham and Moshe Yitzhaki
Department of Information Science & Librarianship, Bar-Ilan University

The Impact of Cultural and Technological Changes on Titles Content and their Use in the Process of Information Retrieval

Abstract: The use of the title as a source of information about books in library catalogs and as a device for retrieval has undergone many changes over time. The paper touches on its historical development, described the impact of technological changes, and analyzed the impact of the computers on the informativity of titles. Also, the differences between Western and Eastern cultures regarding use of title is described.

1. Historical Perspective

The use of the title as a source of information about books in library catalog and as a device for retrieval has undergone many changes over time, and titles were also treated differently by different cultures and societies.

The title in ancient times, was actually the book’s opening words (Avrin, 1991). In many Babylonian libraries there was title information on each tablet together with other information such as the name of the owner (e.g., Amarna tablets of 1400 B.C.); name of the scribe (e.g., tablets excavated in the capital city of the Hittites, dated about 1300 B.C); the tablet number and the first words of the following tablets (e.g., in tablets excavated from the library/archive of Ashurbanipal, in the city of Nineveh, dated about 630 B.C.) (Strout, 1956).

In large ancient libraries, such as that of Ashurbanipal, the clay tablets were kept in earthen jars and the jars were kept in orderly rows on shelves. Each tablet bore an identification tag, indicating the jar, shelf, and room in which it was to be found. On the walls of each room, beside the door was a list of works to be found in it; and something like a subject catalog or descriptive bibliography has been found on tablets, kept near the door of each room. These tablets include title of each work, number of tablet for each work, number of lines, opening words, important subdivision, and locational symbol (Johnson and Harris, 1956).

In Mesopotamian libraries, a colophon is often used in the last column. The word colophon derives from the Ionian city of that name (Glaister, 1979), but the colophon already flourished in ancient Mesopotamia and Egypt. The colophon gave information details about the books such as title (which consisted of the opening words), name of scribe, and his patron; seldom was the author named (Avrin, 1991).

In the literature found in Mesopotamian libraries, some lists of titles were written on walls of temples or pyramids, though the Egyptian books themselves have not survived. The Egyptians would sometimes write the title of the book, a summary of its contents, and, the name of the author on the reverse side or the cover of the scroll.

Throughout Greek literature there are references to book collections. However, most of our knowledge about the Greek’s views of books is derived from the Pinakes compiled by Callimachus in the Alexandria library in about 250 B.C. The Pinakes, which is some sort of catalog or possibly a bibliographical work, is arranged according to incipits—the first word or words of the text. In it the number of lines in each work, the opening words, and some biographical data on the authors were recorded.

The main purpose of the Greek colophon in the Hellenistic and Roman periods seems to have
been to establish the authenticity of the text copied. The counting of the lines originated as an assurance to the reader or purchaser of the scroll that the text was copied properly in its entirely, and it also determined the scribe's fee (Avrin, 1991).

The cataloging and bibliographical procedures of the period were began with a few general "subject" categories. The scholar, after checking the general subject, next looked for the author (title was secondary). Indeed, the concept of the author entry originated with the Greeks; it never appears in any works that have survived from earlier civilizations of the East. It is probably rooted in the democratic belief in the importance of the individual (Strout, 1956).

The evolution of books from papyrus surface to parchment and from roll to codex began in the first centuries and was not entirely established until the fourth century. In the European Middle Ages books took the form of the parchment codex, and were made by monasteries. The scribe frequently began the text with the "incipit" ("here began") and ended it with "explicit" ("explicitus," a tradition from the scroll period) or "finit". In manuscript books, a concluding statement, a colophon, indicated the title of the work, the name of the copyist, date and place of copying, a blessing for the patron or client, and threats of excommunication to unauthorized copiers.

Throughout the Middle Ages, cataloging was limited to an unorganized inventory list, probably representing shelf arrangement, which may have been done according to size or chronology. In fact these were merely lists of books, some arranged by author, some by title, and others by catchword from the title or first line, or sometimes a combination of all three. Some of these lists were originally kept on strips of parchment tacked to the side of the book chest; others were kept in codex form (Johnson and Harris, 1976).

Sometimes, among several works that were bound together only the title of the first work was mentioned. Most works were essentially anonymous, and the book bore the name of the scribe who copied it and not its author. Indeed, throughout the Middle Ages the ancient world's indifference to the proper naming of books and to their authors was sustained (Steinbrg, 1974). Many colophons did not contain the book's titles. Sometimes the name of the copied book appeared on the book's cover (De Vinne, 1972); and the copyist saw himself as exempt from repeating the title because the parchment and paper were expensive. Sometimes the opening words were written in red ink, or, beginning in the sixth century, capital letters rather than the book's regular letters were used to highlight the first words of the book (Riegler, 1995).

One of the earliest listings of a medieval library is dated in the eighth century. It consists of a list of brief titles, with authors added to a few of them (Strout, 1956). One of the first lists that can be designated as a catalog is from St. Martin's Priory at Dover (1389). It is divided into three sections, two of which are arranged by call numbers and give information on books including a short title. The third section is a catalog of analytical entries (which is a genuine innovation) alphabetically listed, some under author, some under title followed by author, some beginning with such words as "book," "part," or "codex".

The first printers followed the form of the manuscript book, and continued to use the colophon. The earliest known printed colophon is from 1457 in a book titled "Psalmorum codex" and published by Johannes Fust and Peter Schoffer of Mainz (Clair, 1976). Peter Schoffer of Mainz was also the first to use a label title in Pope Pius's Bull against the Turks, which he printed in 1463. It consisted of two short lines:

"Bulla cruciata sanctissimi domini nostri Papa cotra turchos"
(cited by Clair, 1976, p.115).

The first label titles took the form of a brief mention of title and author's name, placed at the
top of an otherwise blank protecting leaf at the beginning of a book. These early label titles probably served to prevent the first printed leaf from becoming soiled while lying about in the print shop before being given to the binder (Steinberg, 1974). However, printers soon realized the inherent potential of a full title. Title-page was first used by Erhard Ratdolt, of Augsburg and Venice, in his edition of Johannes Regiomontanus's astronomical and astrological calendar of 1476. He included a page that proceeded the calendar and gave some details about the book. It was set in a woodcut frame that clearly distinguished it from the text.

In 1500 Wolfgang Stockel, in Leipzig, was the first to issue a book with complete title page, giving subject-title, name of publisher, name of printer, date and place of printing (Clair, 1976). For producers, the title page offered more than the technical advantage of a protective cover; it was also a cheap and effective means of advertising the book. Indeed, the first books to use a title page were all new publications that needed some introduction to the public (Steinberg, 1974).

Title pages became increasingly common, by 1500 the title page had established itself; thus facilitating the production of book lists and catalogs while acting as advertisements in themselves (Eisenstein, 1983), and they were decorated and used extensive wording.

During the Renaissance, works again came to be identified with their authors (a concept which has originated with the Greeks). Titles, therefore, were neglected in the single-entry catalogs. The catalog or index was arranged according to author, often using the Christian name rather than the surname. The title was recorded in the catalog, but not as an access point.

In a bibliography arranged in chronological order by the German bibliographer and librarian, Johann Tritheim in the late fifteenth century, an alphabetical author index was included (Strout, 1956). A classified catalog of Syon Monastery, Isleworth, England, from the early 16th century also included an alphabetical author index.

Beginning in the mid-16th century, bibliographers rather than librarians took the lead in improving the catalog. Florian Trefler published at Augsburg a treatise on the keeping of a library (1560). He advocated a five-part catalog consisting of alphabetical author catalog, a shelf list, a classified index to analytics, an alphabetical index to the classified index, and a list of books that were not kept with the main collection (Strout, 1956), although he did not include a title catalog.

In 1595 Andrew Maunsell, an English bookseller, wrote rules for entry in the preface to his catalog of English printed books. He used surname rather than Christian name. He placed anonymous works either by title or else by the subject matter, and he was actually the first to set up the principle of uniform title - he established uniform entry for the Bible. This marked the first recognition that the title can also be used for retrieval (Strout, 1956).

In the 17th and 18th centuries, the concept of the catalog changed. The catalog became a means of finding books rather than inventory. The accepted retrieval entry was that for the author, and subject catalogs provided an author index. This was the case with the catalog of Oxford University, which was prepared under the supervision of the library director, Sir Thomas Bodley early in the 17th century.

The title was usually not regarded as a retrieval device, but it was recorded in the catalog records. The Frenchman Frederic Rostgard asserted in regard to catalogs that the word order of the titles as found on the title page should be preserved.

In the French government's code of 1791, libraries were directed to use card catalogs (because of the wartime shortage). The title page had to be transcribed on the card, and the author's surname underlined as a filing word. If there was no author, the key word in the title had to be underlined (Strout, 1956).

It is interesting that if during the 17th and 18th centuries, an anonymous work did not have a title, a word or words from the document's subject were used (Domanovsky, 1974). That is - there was no clear distinction between an author-title catalog and a subject catalog.
The 19th century was a period of much argument over the relative merits of classified and
dictionary catalogs. The codes of that century emphasize alphabetical arrangement of the catalog,
and usually the author is the main (if not the only) entry. Panizzi, in his 91 rules (British Museum,
1841), preferred, when there was no author, to use the name or the event or person that the book
was about (Rule XXXIII), or the name of any society or body, or place (Rule XXXIV), or of the
editor or even the translator (Rule XXXV); only in cases where none of these rules could be applied,
he suggested the use of the first substantive word in the title (Rule XXXVIII). He also recommended
the use of a uniform title for the Bible (Rule LXXXIX), acts, memoirs, journals (Rule LXXX), almanacs,
and calendars (Rule LXXXIII).

Jewett (1853) referred to title entry in Rule XXIII of his code, according to which translations
of works of unknown authorship were to be entered, like other anonymous works, under the first
word of the original title.

Cutter (1891) used the title entry in anonymous works (Rule 68) and in all works of prose
fiction (Rule 75). In cases where the author's name was known, he suggested a cross-reference
from the title to the author (Rule 70). For the first time, a cataloging code stipulated that the title
always be used as a retrieval access.

In the German and Anglo-American codes of the early 20th century, the use of uniform titles
(together with the use of uniform names) became compulsory (Domanovsky, 1974). This constituted
an important step in the use of titles as a tool for retrieval.

The cataloging codes of the 20th century gave an added entry by the title. However, in the
late 19th century and early 20th century, library catalogs gave only limited access, with only few
title entries (Brunt, 1992). Already in the 1950s, Lubetzky (1953) argued that many common titles-
such as "collected works," "essays," "letters," "memoirs," "Poetry collection," "journal," and so
on - were inadequate clues for the identification and location of these items.

Another of Cutter's innovations was the provision of a subject access point for documents.
Each document was analyzed for its content and assigned a specific subject, to form a subject catalog.

In the 17th and 18th centuries a word expressing the book's subject was sometimes used as
a title for anonymous works, and in the 19th century a catchword from the title was sometimes
used to indicate a book's subject (Harris, 1970). By the late 19th century title words were recognized
to be inadequate for subject access, largely because of synonymy and nonindicative titles. Cutter
introduced a clear distinction between title and subject approaches, which he proposed to be expressed
in alphabetical words; Dewey proposed that they be expressed in decimal numbers.

2. An Increase in Title Informativity

In the mid-20th century, the informativity of documents' titles began to increase (Buxton and
Meadows, 1977; Yitzhaki, 1994). This was manifested in two ways: (a) the replacement of trivial
or literary words by substantive and more meaningful ones, and (b) an increase in the length of titles,
thus including a larger number of substantive words. Many studies have dealt with this issue; most,
however, focused on journal articles' titles and only a few on book titles. This may be partly explained
by the fact that whereas from ancient times up to the late 19th century the book was the main source
of knowledge, since then the journal article has become the most important channel for scholarly
communication.

Some studies left it to persons to judge the informativity of titles (e.g., Peritz, 1984). This
subjective approach is based on classifying the titles checked into two distinctive categories: informative
and noninformative. A title is considered informative if it conveys at least some general idea of
the paper's context. Other scholars relied on counting the number of substantive words in the title
(e.g., Tocatlian, 1970; Buxton and Meadows, 1977; Yitzhaki, 1992, 1994, 1995).
Diodato (1982) examined the occurrence of title words in the abstracts, first and last paragraphs, and cited titles of research papers in chemistry, economy, history, mathematics, and philosophy in the 1960s and 1970s. The best reflection of title-word occurrence was in the abstracts, followed by first paragraphs, last paragraphs, and cited titles, respectively. Longer-than-average titles demonstrated a higher frequency of title word occurrence in first and last paragraphs than did titles in general. These findings strengthen the assumption that longer titles have a larger number of useful words for document retrieval than do shorter titles.

Montgomery and Swanson (1962), Bottle (1970), Miller (1971), and Frost (1989), on the other hand, analyzed the matching of keywords or descriptors from the index entry against the title keywords of the corresponding documents. Ghosh (1974) employed a different approach, examining each title in a specific subject bibliography to see whether the word prostaglandin(s) or its synonymous terms existed in the title.

The increase in title-informativity findings probably indicates a growing awareness among editors and authors of the importance of titles in the process of information retrieval. This trend has progressed at different rates in various fields; generally speaking, the highest increases in title informativity were found in the sciences, with the social sciences and humanities showing lower rates (Yitzhaki, 1992).

The increasing number of substantive words per title over a wide range of subjects before 1960s may be have resulted from the increase in the number of articles published, thus easing the scanning of lists of titles in journals for selecting papers to read (Bird and Night, 1975). The key-word-in-context (KWIC) permuted-title index was introduced by Luhn in 1958 as a relatively inexpensive means of building a dissemination index whose task is to prompt notification to new material. Luhn pondered a temporary bridge between the content of the current literature and readers, awaiting the completion of the more slowly prepared conventional indexes (Luhn, 1960).

Although the trend toward more informative titles has preceded the introduction of KWIC and KWOC indexes (e.g., Buxton and Meadows, 1977; Yitzhaki, 1992), these indexes have undoubtedly contributed much to the growing awareness of the importance of informativity of titles. Indeed, one may associate the "wave" of increase in title informativity as observed for example by Yitzhaki (1992) between 1960 and 1970 to the KWIC and KWOC indexes, but also to the "current contents" that were initiated by Garfield in 1961 for chemical and physical sciences.

The further increase in title informativity that occurred between the 1970s and 1980s and then between the 1980s and the 1990s was probably influenced by the construction of online bibliographic databases and the introduction of OPAC (Online Public Access Catalog) into the libraries. With the advent of computers, the search for words from titles has become one of the main modes of subject search.

It is interesting to note that the increase in title informativity also contributed to the improvement of subject indexes. Studies of indexers have shown that they rely a great deal on titles for their subject analysis (Bottle and Feibish, 1970; Feinberg, 1973). The indexing in PRE-MED database, which provides access to key journals in clinical medicine, nursing, and hospital administration, is based on the title of the article as it appears in the table of contents of the journal (Pitemick, 1985). Thus Pitemick suggests using terms taken from titles in addition to other subject terms. He especially stresses the usefulness of terms taken from titles in new subject areas, for which generally accepted terms have not yet been established.

3. Cultures

The identification of a work with its author or with its title is to a large extent related to the cultural ambience. The Greek's identification of a work with its author reflected their democratic
orientation and their esteem for the individual. The resumption of that identification since the end of the Middle Ages stemmed from a similar orientation in the new era. In oriental cultures, which are less individualist, the traditional entry for a book was its title.

Until the publication of the first edition of the Nippon Cataloging Rules (1943), a title main-entry system had been used in almost all of the cataloging codes in Japan (Takawashi et al., 1989). According to Takawashi's explanation, this can be attributed to the Japanese attitude toward books; the Japanese tend to consider a work as existing independently apart from its author. Older Chinese and Japanese books continue to be entered under title.

A similar situation prevailed with Hebrew books. Most Hebrew books published up to the 20th century were in the field of Talmudic research, in its widest sense, that includes also those dealing with the Hebrew language. An analysis of their titles has revealed that the vast majority consisted of two-words expressions, usually taken literally from one of the books of the Bible (Haberman, 1968). Even a short biblical book such as Canticles (The Song of Songs) has been a source for no less than 50 titles over the past 1500 years (from the completion of the Talmud to our day): for example, "Locked garden" or "Nuts garden" (Riegler, 1993).

A more in depth examination of Hebrew book titles has shown that only in some cases the title is informative, which means that there is any connection between the title and the content of the book. It seems that in most cases when authors had to decide what title they would give a book that they had worked on possibly for many years, having an informative title that would enable information retrieval about the content was not major consideration. It seems that usually the author chose certain biblical phrase as a title mainly because this phrase included his name, whether explicitly or implicitly -- for example "Jacob's Tent" or "Yoshua's faces." Authors regarded this as a suitable way of perpetuating their names, and it is indeed a well-known historical fact that the actual names of authors of Talmudical books tended to be forgotten among the Jewish people, whereas the book titles became widespread surrogates, for these names.

4. Summary

The concept of the title has undergone many changes over the course of human, social and cultural history. In ancient civilizations books did not have a distinctive title; instead the opening words were used to designate the document. In some cultures, such as the Hebrew one, the book's title was chosen without any relation to its content, but instead constituted of a biblical phrase, sometimes with the author's first name implicit in it.

The advent of printing led to the appearance of the title page and the use of titles that represented a book's content in order to advertise and sell the book. As self-serving publicists, early printers issued book lists. They would put their firm's name, emblem and shop address on the front page of their books. Their use of title pages involved a reversal of scribal procedures: they would put the name of their firm first, whereas the scribal colophon had come last. The early printers also extended their new promotional techniques to the authors and artists whose work they published.

Since then the title has constituted a source of information about the book (or article). This process has accelerated during the 20th century, to a large extent because of computerization, with titles becoming more and more informative.

Throughout much of history, there was no clear distinction between a document's subject and its title. There were periods when terms related to the subject appeared in the author/title catalog; in other periods a catchword from the title was used as a subject. However, beginning with the appearance of theorists of knowledge organization in the late 19th century, a clear distinction was made between author catalog, title catalog and subject catalog, with each designated to fulfill a different role in the retrieval of information.
More recently, new retrieval methods that were developed for the computerized systems again caused an overlap between titles and subjects, and we are currently witnessing blurring of the distinction between title and subject approaches.

Today the title is a very important element of any scientific or scholarly book or article. Its primary function is to draw readers' attention to a book or article and indicate its content in a short glimpse, thus influencing its initial selection or rejection. Titles are also an important source for subject searches in most of the automated catalogs and databases.

Authors should be aware of this situation and should provide their writings with informative and indicative titles that will contain as many substantive keywords as possible as opposed to trivial words.

References


Frost, Carolyn. (1989). Titles words as entry vocabulary to LCSH: correlation between assigned LCSH terms and derived terms from titles in bibliographic records with implications for subject access in online catalogs. Cataloging and Classification Quarterly, 10, 165-179.


Towards A Unified Medical Language in a Diverse Cultural Environment

Abstract: Presents compatibility issues which have surfaced as a result of the attempts to establish unified indexing vocabularies. Based on an analysis and comparison of MeSH and four indexing vocabularies for Chinese medical concepts, as well as an investigation of the indexing practice in MEDLINE, the paper shows the problems related to concepts and terminology which have strong cultural backgrounds. Existing approaches such as a metathesaurus (which transcends a set of thesauri), microthesauri (which treat specialized thesauri as the satellites of a superstructure), and “anchor” treatments are considered valuable to the solution while a visualized thesaurus is discussed as a good and necessary indexing and searching aid.

1. Introduction

One important problem to deal with when establishing unified indexing vocabularies is concepts which have strong cultural backgrounds. The impact of cultural diversity has been reflected by the tools for organizing information and knowledge as demonstrated by the many thesauri and classification schemes used in the world. In recent years, the National Library of Medicine has been engaged in a long term project to develop a Unified Medical Language System (UMLS) that will retrieve and integrate information from a variety of information resources. It is one of the pioneers in endeavors to facilitate international exchanges through efficient mechanisms.

In the medical world, one fact that can never be ignored is the abundance and richness of information documenting the uses of natural substances, plants, chemicals, and animal products in Chinese medicine. Medical and pharmaceutical literature, either commissioned by the government or written privately as scholarly exercises, has existed in China for more than 4000 years (Huang, 1993). Chinese medical science has been a significant component of the Chinese culture and a well-established branch in medical science. Much of this knowledge is still in wide usage, neither obsolete nor forgotten. Its influence can be seen all over the world. On the other hand, Western successes in medicinal chemistry, structure-activity studies, and generic modification have opened a promising route to the development of better and more effective drugs from Chinese herbs. Recently, the National Science Foundation of China launched its 1996-2000 five-year project which includes 17 multimillion projects. Listed in number two is a chemical and biological research project for 12 species which are widely used in herbal medicine (Chinese Science News, 1996). All of these have resulted in research and practice and have produced a great amount of information. This trend will continue in the future.

The research presented here is based on a comparative study of indexing vocabularies for Chinese medical concepts. The researcher used several thesauri and classification schemes as the objects of the study including MeSH, a MeSH-based Chinese Thesaurus for Medical Science (1985), the Military Medical Thesaurus (1993, in Chinese), the R class (Medical Science and Medicine) in the Chinese national standard classification, Chinese Libraries Classification (1988), and the Subject Headings of Chinese Medicine (1987).
2. Coverage of Chinese Traditional Medicine Concepts in Five Indexing Vocabularies

For years, Medical Subject Headings (MeSH) has been used to handle information from all over the world by the National Library of Medicine. As the most comprehensive and unique medical indexing vocabulary, MeSH has also been translated, adapted, or directly used by many nations of the world. In the recent Unified Medical Language System (UMLS) project, MeSH is playing a key role again. Although MeSH developers have tried to comprehensively cover as much as possible all of the necessary retrievable subject terms reflecting a collective work in medical science, it is unlikely that MeSH is as exhaustive as to include enough local and cultural-based medical concepts, traditional medical treatments, and alternative medicines used in various nations and communities. In MeSH, specific subject headings for Chinese traditional medicine are very limited. Mainly they are: MEDICINE, CHINESE TRADITIONAL; ACUPUNCTURE (plus 5 specific terms); DRUGS, CHINESE HERBAL; YANG DEFICIENCY; YIN DEFICIENCY; and YIN-YANG. Although there are other subject headings that can be shared or used when indexing Chinese medical literature, it is still very limited since this branch of medical science has formed its unique concepts and terminology through thousand years.

Chinese Thesaurus for Medical Science is an indexing vocabulary compiled based on MeSH and several other Chinese thesauri and was published by the Information Center of China Academy of Medical Science in 1985. It is mainly a list of medical terms and synonyms without cross references. Among the total 10462 entries, about 80% are translated MeSH terms, and 17.36% (1816 entries) are for Chinese traditional medicine. The exhaustivity of Chinese medicine terms is great. However the lack of English or Latin equivalents and cross references has suggested that indexers are expected to have subject expertise at a high level.

The Military Medical Thesaurus (in Chinese) published in 1993, includes 23387 entries, among which 20662 are preferred and 2725 are non-preferred terms. Each entry includes one or more English and Latin equivalent terms for the descriptor and extensive cross references. Two major categories are dedicated to the Chinese traditional medicine: 0803 “Herbal Medicine and Chinese Medicine”, and 09 “Chinese Traditional Medical Science”. The former one contains 627 preferred terms and 26 non-preferred terms. The latter one contains 198 preferred terms and 20 non-preferred terms. Thus Chinese medicine and medical science related descriptors are approximately 4% (825 terms) of the total preferred terms in this thesaurus. Category 0803 could be valued as having a moderate level of specificity in the vocabularies, and Category 09 could be valued as having a minimum level of specificity. There is no indication in this thesaurus for the corresponding MeSH terms, though from a quick glance one can tell that the Western medical science and medicine topics have a great degree of mapping with MeSH terms.

A schedule of MeSH, Class R of the Chinese Libraries Classification, and the Subject Headings of Chinese Medicine was published in 1991. The Chinese Libraries Classification (CLC) is the national standard and has been used nationwide since the 1960s. Class R2 of the CLC presents a systematic structure for organizing Chinese medical literature. Subject Headings of Chinese Medicine (SHCM, 1987 edition) is by far the most exhaustive indexing vocabulary for Chinese medical science. Using Class R (1988 3rd edition) as a basic structure, terms from three vocabularies are mapped. Over 12000 terms from MeSH (1990 edition) are mapped into the 3700 classes and subclasses in the CLC. For Chinese traditional medicine, a total of 3223 terms from SHCM are listed under the corresponding subclasses in R2, among which 46 terms (1.42%) are found to have direct MeSH equivalents and 136 terms (4.22%) are found to have sharable MeSH terms in certain degrees. Many subclasses have encompassed a significant number of Chinese medical terms. The most impressive groups are “Herbal Medicine (Materials)” (614 terms), “Proved Prescriptions and Single Drug Prescriptions” (347 terms), “Diagnostics” (140 terms), and
“Natural Medicinal Chemistry” (124 terms). Figure 1 lists selected examples according to the structure of R2 “Chinese Medical Science (TCM)”:  

<table>
<thead>
<tr>
<th>CLC Notation</th>
<th>CLC Class Name</th>
<th># of terms included</th>
<th>Matching MeSH terms</th>
<th>(% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>R223</td>
<td>Physiology of TCM</td>
<td>77</td>
<td>6</td>
<td>7.79</td>
</tr>
<tr>
<td>R224.1</td>
<td>Channels and Collaterals</td>
<td>55</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R224.2</td>
<td>Acupuncture points</td>
<td>51</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R228</td>
<td>Pathology of TCM</td>
<td>33</td>
<td>1</td>
<td>3.03</td>
</tr>
<tr>
<td>R241</td>
<td>Diagnostics of TCM</td>
<td>140</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R242</td>
<td>Therapeutics of TCM</td>
<td>24</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R243</td>
<td>Eight therapeutic methods (Herbal-based)</td>
<td>92</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R244</td>
<td>Physical therapy of TCM</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R245</td>
<td>Therapy of acupuncture and moxibustion</td>
<td>99</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R247</td>
<td>Other therapy of TCM (including traditional Chinese breathing exercise, Qi-gong)</td>
<td>42</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R28</td>
<td>Traditional Chinese pharmacology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R282.71</td>
<td>Herbal medicine (materials)</td>
<td>614</td>
<td>1</td>
<td>0.16</td>
</tr>
<tr>
<td>R282.74</td>
<td>Crude drug ofzoological origin</td>
<td>64</td>
<td>1</td>
<td>1.56</td>
</tr>
<tr>
<td>R282.76</td>
<td>Mineral drug (material)</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R283</td>
<td>Chinese medicine processing, preparation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R283.6</td>
<td>Dosage form</td>
<td>58</td>
<td>8</td>
<td>13.79</td>
</tr>
<tr>
<td>R284.1</td>
<td>Natural medicinal chemistry</td>
<td>124</td>
<td>23</td>
<td>18.55</td>
</tr>
<tr>
<td>R287</td>
<td>Chinese materia medica</td>
<td>69</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R289</td>
<td>Prescriptions of TCM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R289.5</td>
<td>Proved prescriptions and Single drug prescription</td>
<td>347</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 1: Number of SHCM and MeSH terms mapped into R2 “Chinese Medical Science (TCM)” of CRC

The comparison of the number of terms and types of terms covered by the five vocabularies presented above indicates that MeSH has a very limited range to represent the knowledge of Chinese medical science and has very limited access to this whole area.

3. The Indexing Practice for Chinese Herbal Medicine Document in MEDLINE

3.1 MeSH Heading-based Indexing and Searching

One might doubt at this point whether there is a need to include more specific subject headings or descriptors in a Western language-based and Western service-oriented bibliographic database. To discuss this issue, I would like to use herbal medicine, one of the most distinguished components of Chinese medical science, as my example to examine the indexing practice in MEDLINE.

In China and several Asian countries, medicinal plants play an inherent and prominent role in the general health service and are by no means secondary in functions to synthetic drugs and antibiotics. In China, the quantity of medicinal plants used for direct decoction in the traditional Chinese doctor’s prescription and as ingredients in the official medicine is huge and the current annual demand is reported to be 700,000 tons. A nation-wide survey under WHO investigated thousands of species existing in China in the 1980s and recorded their region, distribution, ecological features, resource and their therapeutic effectiveness. Total number of plant species involved in this survey is 5136 (Xiao, 1988).
The treatment of Chinese herbal medicine is pretty simple in MeSH in which a countable number of subject headings are available for indexing and searching Chinese herbal medicine. A close exam of indexing practice in DIALOG File 154, MEDLINE (1985 to present) that was conducted on March 25 1996 showed following data:

<table>
<thead>
<tr>
<th>Items</th>
<th>Type</th>
<th>RT</th>
<th>Index-term</th>
</tr>
</thead>
<tbody>
<tr>
<td>4190</td>
<td>9</td>
<td></td>
<td>DRUGS, CHINESE HERBAL</td>
</tr>
<tr>
<td>4190</td>
<td>X</td>
<td>D24.310.674.350.</td>
<td>(DRUGS, CHINESE HERBAL)</td>
</tr>
<tr>
<td>4190</td>
<td>X</td>
<td>D26.394.350 (DRUGS, CHINESE HERBAL)</td>
<td></td>
</tr>
<tr>
<td>4190</td>
<td>X</td>
<td>D26.714.350 (DRUGS, CHINESE HERBAL)</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>X</td>
<td>1</td>
<td>CHINESE HERBAL DRUGS</td>
</tr>
<tr>
<td>0</td>
<td>X</td>
<td>1</td>
<td>PLANT EXTRACTS, CHINESE</td>
</tr>
<tr>
<td>2332</td>
<td>R</td>
<td>13</td>
<td>MEDICINE, CHINESE TRADITIONAL</td>
</tr>
<tr>
<td>468</td>
<td>R</td>
<td>8</td>
<td>MEDICINE, HERBAL</td>
</tr>
<tr>
<td>108535</td>
<td>B</td>
<td>56</td>
<td>DRUGS</td>
</tr>
<tr>
<td>5705</td>
<td>B</td>
<td>16</td>
<td>PLANT EXTRACTS</td>
</tr>
</tbody>
</table>

Figure 2: Related terms of “Drugs, Chinese Herbal” and number of items under these terms in File 154

Figure 3 further displays the distribution of these 4190 items under the heading DRUGS, CHINESE HERBAL, ranked by the frequency of items (not including 5 & below items).

<table>
<thead>
<tr>
<th>Items</th>
<th>Index-term</th>
<th>Items</th>
<th>Index-term</th>
</tr>
</thead>
<tbody>
<tr>
<td>1519</td>
<td>THERAPEUTIC USE</td>
<td>82</td>
<td>ISOLATION &amp; PURIFICAT.</td>
</tr>
<tr>
<td>1511</td>
<td>PHARMACOLOGY</td>
<td>42</td>
<td>HISTORY</td>
</tr>
<tr>
<td>847</td>
<td>CHEMISTRY</td>
<td>38</td>
<td>POISONING</td>
</tr>
<tr>
<td>317</td>
<td>ANALYSIS</td>
<td>36</td>
<td>PHARMACOKINETICS</td>
</tr>
<tr>
<td>254</td>
<td>ADMINISTRATION &amp; DOS.</td>
<td>30</td>
<td>CLASSIFICATION</td>
</tr>
<tr>
<td>117</td>
<td>ADVERSE EFFECTS</td>
<td>24</td>
<td>STANDARDS</td>
</tr>
<tr>
<td>93</td>
<td>TOXICITY</td>
<td>17</td>
<td>METABOLISM</td>
</tr>
</tbody>
</table>

Fig. 3: Subheadings under Drugs, Chinese Herbal in File 154

Obviously the approach of searching by MeSH headings under DRUGS, CHINESE HERBAL has several limitations:

First, one finds that no more specific headings are available under this heading. It is unusual that under a most specific level’s subject heading, there are more than 4000 items.

Second, with the help of subheadings, more specific identification of the subject contents of the document can be achieved. However, in this case, the number of items under one subheading may reach several hundred or even over 1500. It will be very difficult, for example, to identify therapeutic use of particular species or the pharmacology of any herb.

Third, in most cases, Chinese drugs based on natural sources include not only botanical origins, but also animal and mineral origins. For example, Cervus nippon (deer horn), nidus vespae, concha ostreae, gypsum, calomel and many others are popular medical materials. It seems to me that there exists a lack of access to non-herbal natural medicine since such documents are given either a general term under MEDICINE, CHINESE TRADITIONAL or the imperfect term HERBS, CHINESE MEDICINE.

Fourth, some of the plants listed in the ancient Chinese Materia Medica were of foreign origin and even today some of the crude drugs prescribed have to be imported. There is no unique subject heading for those foreign origin herbs. I believe this issue widely concerns many nations and communities. It would be very useful to have a unified system or authority file for the names...
of herbs regardless of their origin.

3.2 Freetext-based Searching for Herbal Medicine

Due to the lack of specificity in subject indexing, searching for information on particular herbal medicine drugs heavily relies on the appearance of herbal names in titles and abstracts. Latin names, English names, and various Romanized Chinese or Japanese names are the possible access points leading to the rich documents about herbal drugs. Although for each herb an equivalent Latin name in biological classification and nomenclature may be found, the widely used names of herb are still in Chinese. In MEDLINE databases, at least 80 percent of the literature reporting Chinese herbal medicine research and practice use Chinese names of herbs in titles and abstracts. The complexity of freetext-based searching is very well demonstrated by the titles in MEDLINE records. As demonstrated by the selected examples in Figure 4, at least six different popular situations have appeared in the titles relating to herbal medicine names (titles with brackets are translated titles in MEDLINE).

The variants of existing Romanization systems have resulted in incompatible rules applied by the Western world. For years, LC records have used Wade-Giles Romanized system. This differs from many other products which employ standardized Pinyin (spell-out) Romanization system. Such products can be seen in the papers in the MEDLINE databases and reports in a majority of English newspapers, journals, conference proceedings, and radio. Accompanying this are different parsing rules. It is not surprising that the situation becomes very complicated in MEDLINE when herb names are Romanized or translated into English.

There are many reasons for MARC records to keep the old Wade-Giles Romanization system and at least a consistency is achieved in this way. In MEDLINE records, however, both Wade-Giles and Pinyin Romanization systems are used. Pinyin system is applied in all the fields including Title, Abstract, Authors, and Institution, except the Source field in which Wade-Giles system has been employed. Although this might be the solution of keeping consistent authority records for sources, the practice can cause further confusion to end-users.

This is not the end of the story. Herbs have a tremendous number of synonyms and most of the names are widely used in practice and are recorded in the literature which adds more complexity into freetext-based searching in MEDLINE. For example, Condonopsis pilosula has other popular names such as: Dang Seng; Shang Dang Ren Seng; Huang Seng; Shi Tou Seng; Zhong Ling Cao. Since 1995, I have been tracing a special column about herb medicine which has appeared once a week in the most popular Chinese newspaper China Daily (Ren Min Ri Bao) overseas edition. Eighty species of herbs from this column were included in my data by March 16, 1996. The number of popular names of a herb ranges from one to 17. These 80 herbs have a total of 426 names. In another word, on an average each herb has about five popular names. This number does not include popular names used by Japanese, Korean, and others, nor the phonetically differed names.

It is difficult to standardize the terms in publications or to provide effective access to any of the bibliographic citations and abstracts when no controlled vocabulary is employed. All of these facts indicate the problems of searching based on freetext.

4. Compatible Approaches: The Options

Existing approaches to establishing compatible indexing and retrieval vocabularies have been widely discussed and reported, especially at the Research Seminar on Compatibility and
Integration of Order Systems organized by ISKO in 1995. Based on the experiences reported in published documentation, I have formed several alternative options which might be helpful to the UMLS.

A metathesaurus transcends a set of thesauri. Apparently UMLS has been targeting on incorporating more specific subject areas instead of more geographical areas. If UMLS is considering to incorporate much more cultural-based medical terminology, it should establish for each such entry some kind of authority file so that equivalencies of a term/concept used in various geographical areas can be controlled and items indexed by these terms can be searched efficiently. The three Chinese medical thesauri discussed previously could be very good for UMLS to consider as its source thesauri. Among which, the Military Medical Thesauri would be the best source because it has English and Latin equivalents for each entry while it has kept a moderate specificity in traditional medicine vocabulary.

Microthesauri treat specialized thesauri as the satellites of a superstructure. In this approach, MeSH could be seen as the superstructure since it has exhaustive coverage of Western medical subject areas and has been widely accepted in China. The three thesauri and classification schemes introduced in the previous section are the most popular ones used in major medical libraries and information centers in China. Just like the treatment of LCSH subject headings in its satellite thesauri, these Chinese thesauri have put special marks for the entries for which MeSH headings can be used.

Roulin introduced the "thesaurus and sub-thesaurus global set" concept. In the set, the thesaurus and sub-thesaurus will be two non-disjointed, overlapping sets of concepts: there will
be concepts that only belong either to the former or to the latter as well as concepts that are part of the intersection of both sets, since the most specific concepts in the reference thesaurus are the anchoring points of parts of hierarchical chains peculiar to the sub-thesaurus (Roulin, 1990, 34-35). This valuable approach could be adopted to extend UMLS vocabulary based on MeSH or metathesaurus. Only a countable number of “anchors” need to be added to the system, and existing Chinese indexing vocabularies could be modified to form a sub-thesaurus of MeSH. It could also be considered that only anchors in MeSH Tree Structure be extended and further specific levels in sub-structure rely on classification scheme and notations. This approach will avoid the complicated process of selecting an authority representation of a concept among synonyms.

Considering Chinese herbal medicine, a visualized thesaurus can be established since there is still much confusion in indexing and searching for such information. A visual thesaurus will solve some basic problems such as various equivalents of herbal names caused by different languages, various expressions of herbal names caused by different Romanization and parsing rules, and synonyms of herbal names used in practice and literature. In my experiment, for each entry, a image of the herb and its Latin name, Western name, and various Chinese names are shown with links to broader and narrower terms, as well as the categories. Ideally, we should have a higher lever, where an entry could be further explored by showing a herb’s therapeutic uses, pharmacognostic characteristics, classification, the Chemistry (properties, structures, saponin contents of different parts or different types), pharamacodynamic actions in different systems, etc. The entry could also be linked to the related decoctions used in Chinese folk medicine. No matter at which level the thesaurus/system will exist, it would be a good and necessary indexing and searching aid. However this would require great cooperation between information professionals and medical professionals, and so far the available complete data about herbs is still very limited. It would be very helpful if UMLS would build a knowledge-base for this.

Traditional medicine is widespread throughout the world. As its name implies, it is part of the tradition of each country and employs practices that have been handed down from generation to generation. Its acceptance by a population is largely conditioned by cultural factors and much of traditional medicine; therefore, it may not be easily transferable from one cultural to another (Akerele, 1988). The responsibilities of a unified medical language system are great and so are the challenges. I hope that through this research, I can contribute to the understanding of knowledge organization in today’s changing world, and contribute to the continuing efforts such as UMLS for facilitating information exchanges across time, language, cultural, discipline, and geographical boundaries.

References
Shaoyi He  
University of North Carolina at Chapel Hill

Concept-based vs. Word-based Measures of Medical Information Transfer via English-Chinese and Chinese-English Translation of Medical Titles

Abstract: This comparative informational analysis is conducted on 200 article titles and their 200 translations from two English medical journals and two Chinese medical journals. The informativeness of the original and translated medical titles are calculated and compared for concept-based and word-based measures of medical information transfer via English-to-Chinese and Chinese-to-English translation. The results have shown convincing evidences that concept-based measure provides a more accurate and reliable evaluation on medical information transfer via English-to-Chinese and Chinese-to-English translation of medical journal article titles.

1. Introduction

Many scholars have studied translation as information transfer between languages (Anderson, 1978; Carpenter, 1988; Ganeshsundarm, 1980; Godman and Vetman, 1990; Ibuki et al., 1983; Kaliyan and Rao, 1993; Large, 1983; Olshtain, 1986; Riley, 1992), and some of them have studied the information transfer through English-Chinese and Chinese-English translation (Kovacs, 1986; Liu, 1983; Zhang et al., 1994). But, relatively fewer scholars have investigated translation of journal article titles as information transfer between English and Chinese, and vice versa (Loh and Kong, 1977; Wang et al., 1991). Since medical journal article titles function as condensed summaries of medical information transferred in the medical journal articles in English and Chinese, studies on translation of medical journal article titles in English and Chinese offer opportunities to examine medical information transfer between English and Chinese. This paper concentrates on the comparison between the concept-based and word-based measures for medical information transfer via English-to-Chinese and Chinese-to-English translations of medical journal article titles.

Words and concepts both play very important roles in information representation and transfer. Generally speaking, a concept could be represented by a word (Barnwell, 1980, 141; Rowley, 1992, 253), two words (Haas and Losee, 1994, 626; Ku, 1987, 146), or a phrase (Addison, 1991, 8; Yin and Felley, 1990, 134). Also, a concept represented by a word in one language could be represented by a word, or two words, or a phrase in another language, depending on different situations, since "a concept is a recognizable unit of meaning in any given language (Larson, 1984:55)". Take English and Chinese for example. The English word *patient* has its Chinese counterpart *rongrende* when it means "having ability to endure without complaining". But, when it is used to refer to "an individual receiving medical treatment", its Chinese counterpart *bing ren* means "sick person". Another English word *memorialize* could be a Chinese phrase *chengdi qingyuan-shu* which means "present a written statement to authorities for or against something". Since the concepts represented by such single words as *patient* and *memorialize* in English are not necessarily represented by single words in Chinese, how to measure informativeness of journal article titles has become an important and interesting issue for information representation and transfer via article title translation between English and...
There are two types of informativeness measure: one is based on word-counting and the other is based on concept-counting. The word-based measure for document titles was first proposed by Richard Diener in 1984. The title informativeness, also known as Informational Value (IV), is defined as "counting the total number of words (Yf) in a title, counting the number of substantive words (Ys), or keywords, in a title, and computing the proportion of substantive words \( Yp = Ys/Yf \)" (Diener, 1984:222). While it was useful for calculating informativeness of English document titles, the word-based measure is unable to account for conceptual representation in medical information transfer between English and Chinese via medical title translation. Thus, such an incapability of the word-based measure calls for another better measure. In this paper, the author proposes a concept-based measure that is called Conceptual Information Value (CIV). It is defined as "counting the total number of words in a title (W), counting the total number of concepts in the same title (C), and computing the proportion of concepts \( CIV = W/C \)", and used to evaluate medical information transfer via medical title translation between English and Chinese, and vice versa.

The goal of this paper is to conduct a comparative informational analysis between the concept-based and word-based measures of medical title informativeness, via English-Chinese and Chinese-English translation of medical titles, to see which measure provides greater accuracy and is more reliable in information representation and transfer.

2. Materials and Methodology

The corpus for this research include two types of translations: one is English-to-Chinese Translations of medical journal article titles and the other is Chinese-to-English translations of medical journal article titles. The 400 original and translated medical article titles in English and Chinese are as follows:

One hundred original and 100 translated article titles are obtained from two medical journals in English and their translated counterparts in Chinese: 1) 50 article titles in English are from JAMA and their counterparts of 50 Chinese translations are from Meiguo Yixuehui Zhai Zhongwen Ban (Chinese Edition of JAMA); and 2) 50 article titles in English are from Archives of Ophthalmology and their counterparts of 50 Chinese translations are from Meiguo Yixuehui Yanke Zhai Zhongwen Ban (Chinese Edition of Archives of Ophthalmology).

One hundred original and 100 translated article titles are obtained from two Chinese medical journals and their translation in English: 1) 50 article titles in Chinese are from Zhonghua Yixue Zhai (National Medical Journal of China) and their counterparts of 50 English translations are from the translated table of contents in English; 2) 50 article titles in Chinese are from Zhonghua Yanke Zhai (Chinese Journal of Ophthalmology) and their counterparts of 50 English translations are from the translated table of contents in English.

The measurement of title informativeness of the original and translated medical journal article titles are compared and analyzed as follows:

1) The average Informational Value (IV) per title of the original medical journal article titles is first calculated and then compared with that of the translated medical journal article titles: the less difference between them, the more accurate the measure, and vice versa.

2) The average Conceptual Information Value (CIV) per title of the original and translated medical journal article titles is first calculated and then compared: the less difference between them, the more accurate the measure, and vice versa.
3. Results

First, the average word-based measure is used to calculate the *information value (IV)* of the original and translated article titles from the two English medical journals and their Chinese counterparts. The results are as follows:

<table>
<thead>
<tr>
<th>Journal Title</th>
<th>Average IV of Original Titles</th>
<th>Average IV of Translated Titles</th>
<th>Difference of IV in Ori. &amp; Tran. Titles</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAMA</td>
<td>0.742</td>
<td>0.828</td>
<td>0.086</td>
</tr>
<tr>
<td>Archives of Ophthalmology</td>
<td>0.753</td>
<td>0.829</td>
<td>0.076</td>
</tr>
<tr>
<td>JAMA &amp; Archives of Ophthalmology</td>
<td>0.748</td>
<td>0.829</td>
<td>0.081</td>
</tr>
</tbody>
</table>

Table 1: IV in English-Chinese translation of medical journal article titles

Second, the average *information value (IV)* of the original and translated article titles from the two Chinese medical journals is calculated with the word-based measure. The following are the results:

<table>
<thead>
<tr>
<th>Journal Title</th>
<th>Average IV in Original Titles</th>
<th>Average IV in Translated Titles</th>
<th>Difference of IV in Ori. &amp; Tran. Titles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhonghua Yixue Zazhi</td>
<td>0.807</td>
<td>0.732</td>
<td>0.075</td>
</tr>
<tr>
<td>Zhonghua Yanke Zazhi</td>
<td>0.871</td>
<td>0.734</td>
<td>0.137</td>
</tr>
<tr>
<td>Zhonghua Yixue Zazhi &amp; Zhonghua Yanke Zazhi</td>
<td>0.839</td>
<td>0.733</td>
<td>0.106</td>
</tr>
</tbody>
</table>

Table 2: IV in Chinese-English translation of medical journal article titles

Third, the average *Conceptual information value (CIV)* of the original and translated article titles in the two medical journals in English and their counterparts in Chinese is calculated with the concept-based measure. The results are shown as follows:

<table>
<thead>
<tr>
<th>Journal Title</th>
<th>Average CIV in Original Titles</th>
<th>Average CIV in Translated Titles</th>
<th>Difference of CIV in Ori. &amp; Tran. Titles</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAMA</td>
<td>0.434</td>
<td>0.432</td>
<td>0.002</td>
</tr>
<tr>
<td>Archives of Ophthalmology</td>
<td>0.455</td>
<td>0.451</td>
<td>0.004</td>
</tr>
<tr>
<td>JAMA &amp; Archives of Ophthalmology</td>
<td>0.445</td>
<td>0.442</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Table 3: CIV in English-Chinese translation of medical journal article titles

Fourth, the average *conceptual information value (CIV)* of the original and translated article titles from the two Chinese medical journals is calculated with the concept-based
measure. The results are given below:

<table>
<thead>
<tr>
<th>Journal Title</th>
<th>Average CIV in Original Titles</th>
<th>Average CIV in Translated Titles</th>
<th>Difference of CIV in Ori. &amp; Tran. Titles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhonghua Yixue Zazhi</td>
<td>0.425</td>
<td>0.412</td>
<td>0.013</td>
</tr>
<tr>
<td>Zhonghua Yanke Zazhi</td>
<td>0.403</td>
<td>0.410</td>
<td>0.007</td>
</tr>
<tr>
<td>Zhonghua Yixue Zazhi &amp; Zazhi</td>
<td>0.414</td>
<td>0.411</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Table 4: CIV in Chinese-English translation of medical journal article titles

4. Discussion

From the tables 1 to 4 which are shown above, we can see clearly the differences between the word-based measure of information value (IV) and the concept-based measure of conceptual information value (CIV). Quite obviously, the concept-based measure of conceptual information value (CIV) has yielded much better results for both English-Chinese and Chinese-English translation of medical journal article titles.

Let us first look at the word-based measure of information value (IV) in both English-Chinese and Chinese-English translation of medical journal article titles. On the one hand, for the word-based measure of the information value (IV) in the English-Chinese translation, the information value (IV) of the translated article titles in Chinese is about 8.1% more than that of the original article titles in English. That is to say, the translated article titles in Chinese have more substantive words than their English counterparts. On the other hand, for the word-based measure of the information value (IV) in the Chinese-English translation, the information value (IV) of the translated article titles in English is about 10.6% less than that of the original article titles in Chinese. In other words, the translated article titles in English have fewer substantive words than their Chinese counterparts. With such differences between the original and translated information value (IV), it is not reliable to measure the informativeness of the original and translated article titles just by counting their substantive words in English and Chinese.

As for the concept-based measure of conceptual information value (CIV) in both original and translated medical journal article titles, the results are much better in both English-Chinese and Chinese-English translation. In the translation from English into Chinese, the concept-based measure of conceptual information value (CIV) in the translated article titles in Chinese is only about 0.3% less than that of original titles in English. This means that there is hardly a change in the conceptual information value (CIV) with the concept-based measure for English-Chinese translation of medical journal article titles. Such is the same case with the translation from Chinese into English. The concept-based measure of conceptual information value (CIV) in the translated article titles in English is also about 0.3% less than that of original titles in Chinese. This again means that there is almost no change at all in the conceptual information value (CIV) with the concept-based measure for Chinese-English translation of medical journal article titles. With almost no difference between the original and translated conceptual information value (IV), it is thus reliable to measure the informativeness of the original and translated article titles by
counting the concepts in English and Chinese.

The above results also indicate the advantage of concept-based measure of **conceptual information value (CIV)** in translation as information transfer between English and Chinese because the informativeness of article titles in both English and Chinese is solely represented by concepts rather than words. In addition, the differences between word-counting and concept-counting for title informativeness have shown the latter is more practical in measuring conceptual information transfer between English and Chinese, and vice versa.

5. Conclusions

Based on the comparative informational analysis between the concept-based measure and the word-based measure, this paper shows convincing evidences that concept-based measure is more accurate and thus more reliable for evaluating medical information transfer via English-to-Chinese and Chinese-to-English translation of medical journal article titles. Further studies with a larger research corpus of medical journal article titles in both English and Chinese are needed to validate the results of this exploratory comparative analysis. The methodology used in this study to compare word-based and concept-based measures of medical article titles in English and Chinese should then be applied to article titles in different fields other than medicine and different languages other than English and Chinese.

References


Abstract: The main focus of this paper is to relate the use of controlled, structured vocabularies to the ability to standardize the definition and form of a link. In the simplest of terms, if there is no well defined and consistent knowledgeable of concepts there can be no well defined and consistent application of inter-concept links. This paper discusses the various limitations of current efforts in both information science and computer science/artificial intelligence to handle interconcept, non-hierarchic links. The research uses an already structured knowledge base which provides several key attributes: a context for the terminology, agreement on concept/term definition, and standardization in form. Second, this research uses concepts apart from their use in sentences and thus, while derived from a discipline, is not tied to any particular body of literature. That is, there are no specific corpora from which the terms were drawn. Third, it regularizes the use of associative relations in several ways:

- providing a definition for the link
- providing rules for the use of links
- providing directional information for the link
- providing known end points of the link.

This research posits that without a highly structured vocabulary as the basis for constructing links, they cannot be effectively and efficiently developed. The vocabulary used in this exercise is the Art and Architecture Thesaurus (AAT, 1990)

1. Introduction

Individuals seeking information enter an information retrieval system at a specific point. Most often they choose a term or set of terms they feel represents the subject or the information they are interested in and they begin exploring the content of the system from that entry point. Often the information they are able to retrieve is incomplete, or inadequate to their needs, despite the richness or completeness of the actual information contained in the system. The question then arises, how does one effectively and efficiently move the user through the system to the information that will meet his/her needs? One answer is more effective use of a tool already in existence—the associative, inter-concept link. Such links are non-hierarchic pathways connecting concepts and leading the user to information of potential use.

Current systems link information using techniques that: are inconsistent, both within a given system and between systems; have no rules, or minimal rules, governing the use of the linking techniques; result in idiosyncratic application and incomplete pathways through the information system. This paper, based on my dissertation research, indicates that the effective deployment of associative, inter-concept links can only be done within a hierarchically structured vocabulary. This research also suggests that standardizing the definition and form of this existing tool will allow users to have considerably improved chances of finding information relevant to their needs.

2. Background to Problem

At the close of World War II, Vannevar Bush, Director of the Federal Office of Scientific Research and Development, challenged the thousands of scientists whose efforts had been directed toward the war to continue their work, but for peace. In his famous article, "As We May Think," Bush (1945) laid out a broad range of visions, fantastic for the time, that he felt...
were attainable extensions of discoveries to date. One aspect of Bush's vision was a system, the Memex, that would allow individuals to use their own associative links to tie concepts and documents together in long chains with multiple branches. Any document could belong to many chains in his "associative indexing" scheme and each scheme would be unique to the individual who developed it. Using tools such as hypertext links, researchers in the area of information processing and retrieval are only beginning to reach the aspirations he voiced.

Bush's idea brings together what we now call *semantic relations*, or links between concepts that are based on inherent properties of those concepts. Bush would have the links be user-driven, that is, the creator of the link defines what relations are useful for that individual and creates an individuated information network. While this is an interesting idea for private files, and may well reflect the way individuals organize material, his vision does little to bring overall organization to the mass of information which, as Bush predicted, has multiplied at an alarming rate since 1945, the year of his paper.

Since at least the late 1960's, creators of subject heading lists and thesauri have avidly pursued the use of associations, or relationships between concepts, contained in documents. The failing has been two fold: lack of agreement on defining associative, inter-concept links and lack of rules for constructing such links. In both information science and computer science the few rules that exist are loosely structured and usually driven by a particular knowledge structuring tool. The indexer of documents or the knowledge engineer building an expert system decides what needs to be linked and, within the confines of the system, how to do it. At issue is the world view of one person against that of another. Any meaningful lexical-semantic relation that could provide some basis for a semantic structure to the information is frequently lost, or at best subordinated in such systems to the view of a body of indexers or knowledge base builders.

Work over the past few decades has sought to more precisely define the relationship that is implied when concepts are linked. Research on associative links has been simultaneously conducted in computer science, psychology, anthropology, and information science often with little awareness by one of the other. The problems faced by all of these communities focus on the need to structure knowledge for the sake of applying it to a problem—either human inquiry or machine-based reasoning.

Work on associative links has special import in information science and in computer science where the same problems, namely a lack of definition and a lack of rules for linking concepts, occur. There are, however, major differences between information science and computer science systems. One is scale. Because information retrieval systems are massive, that is, they contain 15 to 20 million records accessible by a subject vocabulary of 60 to 100 thousand concepts (either single terms or term phrases), no individual can know or manipulate even a subset, to say nothing of the totality. The information seeker requires a systematic way to maneuver through such systems. By contrast, the typical reasoning system in computer science may have 600 to 1000 concepts operating in a highly constrained domain. One or two orders of magnitude more concepts are found in the commonly used information retrieval systems. Second, information retrieval systems have an immediate need to sort through huge quantities of information whereas computer science applications such as expert systems call for automated, general reasoning.

Both environments, information science and computer science, recognize a need for context, that is, for placing a concept or fact within a framework that provides either or both the system and the user with information needed to direct a search through an information base—be it a knowledge base or a bibliographic database. Stated in its simplest form, the more inclusive and complete the context of the information, the more accurate the information
retrieval. If we can enhance the schemes we use to code information in a way that includes term/concept relationships and context identifiers, we will provide the means for more complete and more accurate retrieval of information. We know that relations and roles are important in the query formulation process. (Bates, 1986). One asks about rivers in relation to their role in transportation or as sources of power, carriers of pollution, etc. If a user can employ more of the knowledge she brings to the information searching process, the relevance of the retrieved information should increase, including fewer false hits. In today's retrieval systems we consciously strip away "extraneous" information, such as concept relationships, brought by the user to the search, and instead require users to work around the problems our system's limitations impose.

In examining the problem of the traditional related term structure employed in vocabularies we see three clear problems. The first focuses on the usual convention for linking concepts in subject heading lists and thesauri—namely, the "see also" cross reference. Because there are no rules by which "see also" connections are made short of Soergel's (1974) admonishment that, paraphrasing, concept A is related to concept B if an indexer thinks so. The relationships codified by such links represent, at best, the world view of the individual making them. Second, it is impossible to conceive of creating a set of rules for "see also" references when they are utilized in vocabularies or word lists that lack any substantive notion of a structure aside from alphabetization. In those few cases where some structure is offered, it often results in lists with mixed membership, that is, the list may include synonyms, broader terms, and narrower terms, as well as a variety of indistinguishable relationships thrown together. Third, the lack of systematic rules causes several problems. There is an inconsistent and incomplete application of links, and an incompatibility among the knowledge structures that are created. Finally, all of this ends up at the feet of the user who is unable to reliably and consistently retrieve information; misses information or retrieves too much irrelevant information because of incomplete linking; and has no sense of why one concept is linked to another making retrieval results confusing or incomprehensible.

3. Attributes of Controlled, Hierarchically Structured Vocabularies

Controlled, hierarchically structured vocabularies offer a concrete basis for the application of associative links. In such vocabularies several assumptions can be made that facilitate the applications of links. First, hierarchies are formed by the application of genus/species or generalization/specialization vertical links. This results in a tree structure that is generally well understood and highly stable. Whole/part relationships can either be used hierarchically or as quasi-associative links. For purposes of inheritance it is preferable not to mix genus/species and whole/part relations in a single hierarchy. In either case, if whole/part is used as a structuring device or as a link, there must be a clear understanding of what is covered by the whole/part relationship since it is not a simple concept (Evens, 1988).

Second, the totality of terms subject to linking is known. When implementing a link within a controlled, structured vocabulary one has a sense of the totality of the domain involved because care is normally taken to represent a domain with reasonable completeness. Unlike extracting whatever terms/concepts appear in a particular corpus and structuring them, constructing a thesaurus is a more deliberate process by which gaps in the vocabulary are filled in.

Third, the hierarchic relationships (is-a or genus/species), which provide structure to the vocabulary, also provide the basis for agreement on the meaning of terms. The definition of a concept is identified by, and in the case of homonyms is determined by, the structure at the point
within the vocabulary that the concept appears. For example:

an ice station is a kind of research station is a kind of observatory which is a kind of research building.

These definitions are ascertained simply by reading upward in the hierarchy and understanding the significance of the indentions used in the structure. It is, in other words, a verbal "translation" of the following hierarchy:

research building
  observatory
    research station
      ice station

Agreement on concept definition is an important element in the accuracy of linking. Without it, for example, homographs cause problems in understanding what concept is actually being linked.

Fourth, mono-hierarchies provide an opportunity for categorization of terms within broad categories, separating objects, actions, abstract concepts, etc., and lay the ground work for link rules. Within an alphabetic list all manner of terms are grouped indiscriminately, ordinarily with no relationship to terms that fall to either side of them. When trying, in an organized fashion, to embed links into an alphabetically organized vocabulary, the essentially random nature of the alphabetic listing is a severe hindrance. Not only are like concepts not grouped together, but there is no separation of terminology/concepts into categories such as objects or agents, which categorization significantly facilitates accurate link rule-making.

Fifth, concepts are already categorized on a broad scale by membership in a hierarchy, and on a more refined scale by the is-a nature of the vertical links. As a result, links can be conceived by categories rather than by individual concepts. One does not have to think of each instance of a link on a term by term basis. This makes the task easier because many "test cases" exist in one place in the vocabulary. In addition, the structure of the vocabulary avoids the problem of overlooking or not remembering some terms that should be linked.

Within this context of a relatively complete, hierarchically structured vocabulary one can begin to apply links.

4. Discovery and Application of Links

The method used to define a link and to reach agreement on what kinds of concepts (agents, objects, actions, etc.) it can link is not one of definition but one of procedure. That is, there appears to be no definitional way to identify links and their properties, rather, there is a process directed toward that end. This involves work by a team of content experts and lexicographers considering proposals, challenging the validity of the proposals, and engaging in conflict resolution until agreement is reached and challenges are exhausted. This is, by no means, the usual or ordinary method of creating links in either bibliographic files (card catalogs, indexes, etc.) or computer science systems (expert systems.) The norm can be characterized by the random and individual efforts of system builders working either in isolation or with some modest check-and-balance collaboration. In the case of the AAT the lead editor gathered information on links that had been proposed by other researchers. The editorial group then began a long and occasionally argumentative process of determining if a particular link was appropriate for the AAT and where and how it could be used. There was an enormous give and
take process. In the end a link was accepted, its definition set, and the rules of its use were codified. An example of a link definition and application criteria is given below.

Inter-concept links operate on the horizontal plane and join two distinct concepts, or groups of concepts, by means of a specific and labeled connection. The connection not only has a name, but it has the property of directionality. An interconcept link may be unidirectional, symmetric, or complemented by a corresponding link operating in the opposite direction. Within that whole, it is easier to determine if a link is valid when all currently known concepts are delineated in the vocabulary and grouped together in a hierarchic structure. It is, however, important to note that the vocabulary can be expanded to meet new needs, but is, at any point in time, a known entity. The expansion process becomes tied to the link structure as new additions to the vocabulary must fit, not only within the hierarchic structure but also within the link structure that has been overlaid on the vocabulary. If a concept/term is suggested and its placement in the hierarchic structure violates the relationship to sibling terms, for example, the term and its placement must be examined for accuracy. Likewise, when adding a term to the hierarchy the term must also work with any interconcept links that may exist between the group of terms of which it is a structural member and those terms to which the link points. If the new terms does not work with either or both the hierarchy and the links, then the term's placement is in question. There is, then, a kind of check and balance operation between link building and vocabulary structuring.

The more ordering of concepts that is provided in a knowledge base the easier it is to formulate rules about linking those concepts. In an alphabetic list all terms are equal and not even the basic genus/species information is provided. Although some subject heading tools provide partial ordering by giving a cluster of broader and narrower terms that surround a concept, these provide an inadequate base on which to build interconcept links. With an inadequate structure there can be no assurance that link building will be done consistently and accurately. One can imagine re-writing the link rules to accommodate a purely alphabetic list of concepts but the result would place an incredible burden on the implementer. For example, the following changes might be made to AAT link 3B:

\begin{quote}
\textit{AAT} Rule: To or from terms in the Associated Concepts facet and the resulting or causative actions (terms in the Activities facet).

\textit{Generic Rule:} To or from terms expressing abstract concepts, qualities, and phenomena that relate directly to the study and execution of art and architecture, or theoretical and critical concerns, ideologies, attitudes, and social or cultural movements AND terms expressing areas of endeavor, physical and mental actions, discrete occurrences, and methods employed toward a certain end that arc resulting or causative actions.
\end{quote}

First, it is difficult to imagine from a list of thousands of terms exactly which terms may fit the description of those qualifying for each half of the 3B link. Second, the generic rule leaves the judgment entirely in the hands of the person applying the rules. The combination of these two factors makes it impossible to assure consistency and accuracy in interpretation, the exact fault of previous systems. Is adaptive reuse an abstract concept or is it a method employed toward an end? Are renovation or remodeling abstract concepts or are they physical actions? Can they be linked by 3B? If one is working from the structure of the \textit{AAT} the answer is straightforward—adaptive reuse is found in the Abstract Concepts facet and renovation and remodeling are from the Activities facet, so the 3B link is entirely appropriate. Without that structure there is near-chaos.
5. Elements of a Link Formalism

Notation for a link should accomplish the following:
- indicate the direction of the link
- provide an identification of the link type
- identify what concepts are joined, i.e., single concepts or groups of concepts

In a mono-hierarchically structured vocabulary such as the AAT there are two levels of link formulation:
- a formalism that enables links to be made between specific groups of terms by linking the relevant broader terms found in two hierarchies. Such a formalism can be called an authorization formalism.
- a formalism that links specific, individual concepts. Such a formalism can be called an instantiation formalism as it instantiates a particular link.

The form of both formalisms, given in the figure below, is the same. The difference is that an authorization formalism has a Node 2 which may have sibling terms, whereas an instantiation formalism has a Node 2 which is a single term with no siblings.

<table>
<thead>
<tr>
<th>Slot Name</th>
<th>Filler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link Type Number</td>
<td>Numeric Designator</td>
</tr>
<tr>
<td>Link Name</td>
<td>Label of the link</td>
</tr>
<tr>
<td>Form</td>
<td>Alphabetic Designator (Link 4 has a null designator)</td>
</tr>
<tr>
<td>Node 1:</td>
<td>Hierarchy and Term or Unique Term Address</td>
</tr>
<tr>
<td>Node 2:</td>
<td>Hierarchy and Term or Unique Term Address</td>
</tr>
<tr>
<td>Reciprocity:</td>
<td>Y/N</td>
</tr>
<tr>
<td>Corollary Link:</td>
<td>Link Type Number and Form Designator</td>
</tr>
</tbody>
</table>

Formalism for Enabling and Instantiation Links

For the sake of the user, as well as the accuracy of the links, there can be no undifferentiated or imprecise links. All links must be governed by rules, be of a specific type, have a directionality, and link-specific end points. Without these features the resulting information system, whether in information science or computer science, will suffer the full range of problems described earlier in this paper.

6. Summary

Building interconcept links can be done in a rational, rule-based manner rather than haphazardly according to the world view of an individual, however well-informed or well-intentioned that person may be. The AAT editors were able to formulate a list of relevant links and implement them in the thesaurus adhering to rules they established. Users of this knowledge base will not only know that two concepts are related, but what that relationship is. There is no other instance of this in the information science community apart from a specific application, namely Humphrey's frame-based MedIndEx system (Humphrey, 1989). The earlier efforts in information science to create broad systems either remain unused (Gardin, 1965) or have resulted in efforts that are hopelessly lost in the inconsistencies of an application (LCSH, 1990).

I have suggested that the nature of the knowledge base, particularly its structure, or lack thereof, is a critical factor in the ability of users to develop and implement interconcept links. There is a reciprocal aspect, namely the process of developing and implementing links serves as a check on the structure of the knowledge base. Given the fact, pointed out in an earlier discussion, that there is no absolutely correct way to structure knowledge, any tool that helps check and tighten the structure of a knowledge base should be viewed as a welcome addition to the arsenal of current methods of
operation. Link building serves such a function.

References
Development of a Relational Thesaurus

Abstract: Various shortcomings typically attend thesaurial relationships: failure to support extended relevance relationships; lack of effort in identifying a common relational inventory across types of retrieval systems; limitation to binary relationships; inattention to relationships built into the meaning of lexical units. To counteract these failings, a preliminary inventory of relational structures underlying the ca. 1250 most frequently occurring English verbs is presented. The inventory is compact and corresponds to a combination of semantic role-based verb types, as identified by Chafe (1970), and image schemata, as identified by Johnson (1987). The nature of hierarchical relationships among relational structures within the inventory is surveyed.

I. Relationships in Thesauri: Desiderata

Thesauri have two components: a set of expressions that refer to conceptual entities—some being authorized descriptors, others being lead-in terms—and the relationships among them. The basic premise of this paper is that, since relationships are the glue that holds everything together, transforming isolated parts into coherent wholes, we need to exercise a degree of control over relationships similar to the terminological and structural control we exercise over the conceptual entities they bind together.

Relationships play various roles in thesauri. Some relationships show the conceptual equivalence between lead-in terms and their corresponding authorized descriptors (e.g., USE/USE FOR). Other relationships define the nature of the semantic link between two authorized descriptors (e.g., BT/NT). Yet other relationships are used with authorized descriptors to create representations of new and complex conceptual entities (e.g., AND, OR, NOT). The first two types of relationships serve primarily to help indexers and searchers locate the best access points; the third type of relationship enables the thesaurus to expand its range to concepts not explicitly covered by the authorized descriptors. Both functions are critical to the intelligent use of a thesaurus.

The relational system of the typical thesaurus is nearly exhausted by the examples just given. (Indeed, many thesauri leave it up to the retrieval system to supply combinatory, e.g., Boolean, relationships.) The only other common relationship type is the nebulous RT relationship. The end result is that thesauri typically use only one well-defined relationship between base concepts for which authorized descriptors are provided—the hierarchical, broader term/narrower term (BT/NT) relationship—encompassing all others in the ambiguous related term (RT) relationship.

This situation is wanting in several regards. First, a recent study concluded that the relationship between the topic of a user's need and the topic of a document relevant to that need is often not strictly a matching relationship, but may range across a wide variety of relationship types, both paradigmatic and syntagmatic (Green and Bean, 1995). In order to model the relevance relationship more fully in our retrieval systems, we need access to a richer array of relationship types than the limited set now in common use.

Second, our retrieval systems—document retrieval systems, information retrieval systems, database systems, knowledge-based systems—would be both more effective and more efficient if they were based on a common model of what entities need to be accounted for and what
relationships exist between them. The more extended inventories of relationships so far used in database and knowledge-based systems tend to be ad hoc. Much effort could be saved if we had a systematic and structured inventory of relationships available for use across all types of information-retrieval systems.

Third, almost all thesaural relationships that have been proposed are binary. But some relationships are by their very nature n-ary relationships, meaning that they bind together more than two entities to form the whole. To use a common database example, Supplier <supplies> Part <to> Project is a ternary relationship; we need to keep all three entities—Supplier, Part, and Project—bound together in a single relationship, because we could not otherwise correctly associate together the specific Suppliers, Parts, and Projects that belong together. In developing a relational system, we therefore need to be mindful not to limit the expression of relationships to binary relationships.

Fourth, many conceptual entities are complex and express internal relationships. For example, the conceptual entity Buyer has built into it a complex relationship structure that incorporates an object (often referred to as Merchandise), previously possessed by a Seller, that comes into the possession of the Buyer in exchange for something of equivalent value, usually Money. The term buyer thus evokes a complex relationship, which is lexicalized by the single term. If our relational systems are to perform their roles adequately, such internal relationships need to be made explicit.

These comments shed insight into the basic features required of a relational thesaurus, i.e., a structured inventory of relationship types. First, in order to model relevance relationships adequately, the inventory must admit of many relationship types. The relational information needed for tasks other than document retrieval may expand the breadth of a general relational thesaurus even further. Second, a relational thesaurus should not set limits on the degree of relationships expressed (where the degree of a relationship is the number of entities it binds together), but should be guided by the nature of the relationships that need to be expressed. Third, a relational thesaurus needs to be as cognizant of relationships that are conceptually embedded into natural language terms as it is of relationships that are used to form new, complex concepts.

This paper reports on the preliminary results of a research effort now being undertaken, which addresses our need for a general-purpose relational thesaurus.2 The major outcome of the study will be the development of a relational thesaurus adhering to the principles stated. This will entail both identifying a reasonably comprehensive set of conceptual relationship types and determining how to organize those relationship types into a structured inventory. The results presented here include an inventory of the most basic relational structures, which will serve as the core of the larger thesaurus. It also demonstrates some of the organizational principles needed in that thesaurus, based on an analysis of Path/Journey relationships.

2. Methodological Issues in Developing a Relational Thesaurus

A relational thesaurus strives to implement terminological control over relationships, much as conventional thesauri set out to implement terminological control over concepts. Part of the import of this statement is that although the units over which we attempt to exercise control are conceptual in nature, we can only implement this control through the medium of language, which is the most effective means human beings have for the communication of ideas. As language is used routinely to convey relationships, we turn naturally to linguistic data to reveal the array of relationships that are important to us. (Note: Since all human languages are not equivalent, a relational thesaurus based on English data, as this one is, will not necessarily be of universal
applicability.)

English uses 8 parts of speech: nouns, pronouns, verbs, prepositions, conjunctions, adjectives, adverbs, and interjections. These parts of speech correspond closely to basic components of the entity-relationship model (see, for example, Fidel, 1987, ch. 5): nouns and pronouns normally refer to entities; verbs, prepositions, and conjunctions are used to express relationships; adjectives and adverbs convey attributes of entities and relationships, respectively. (Interjections, which are said to correspond to the irrational or pre-rational, stand outside the entity-relationship model.) Prepositions and conjunctions are closed classes, and thus express only a limited range of relationship types used in English; for example, the relationships they express are almost all binary in nature. Verbs, however, constitute an open class, thus covering a wide range of relationship types, both conceptually (they cover all subject domains) and structurally (the relationship types expressed by verbs are not artificially constrained to be of a certain degree).

On the structural level, a verb can be envisioned as the hub of the linguistic unit in which it appears, with spokes leading from it to the various entities it pulls together. In other words, a verb (or verb form) governs the clause (or phrase) in which it occurs. Using a specific verb may force the speaker to mention a specific entity, may render it impossible to mention an entity, or may leave the mention of an entity to the discretion of the speaker. For example, the class of Commercial Exchange events referred to as buying and selling employ a relation between 4 participant entities, as previously noted: a Buyer, a Seller, Merchandise, and Money. If we mention such an event using the verb sell, we are required to refer to the Merchandise, but the Seller, Buyer, and Money are optional, as exemplified by the non-acceptability of (1) (which might be informally acceptable, but only if the Merchandise is identifiable from context) and the acceptability of (2)-(4), if, however, we mention the event using the verb cost, as in (5), we cannot refer to the Seller, except as the prior possessor of the Merchandise (i.e., John's bike).

* (1) John sold to Harry for $60. (Merchandise absent)
(2) The bike was sold to Harry for $60. (Seller absent)
(3) John sold the bike for $60. (Buyer absent)
(4) John sold the bike to Harry. (Money absent)
(5) The bike cost Harry $60. (Seller absent)

Several observations are in order here. First, the relational structure of a verb helps define its syntactic behavior. When the verb sell is used, for example, the Seller tends to appear as the subject of the verb, the Merchandise as the direct object, the Seller as the object of a to-clause (a.k.a. indirect object), and the Money as the object of a for-clause. But these assignments are not invariant (for example, both the Merchandise and the Buyer can show up as the subject of the verb), although the relational structure binding Seller, Buyer, Merchandise, and Money holds constant. Moreover, the syntactic tendencies of sell differ from the syntactic tendencies of other Commercial Exchange verbs (e.g., buy, charge, cost, pay, purchase, spend), but they all refer to the same relational structure.

The situation just surveyed calls for identifying relational structures not for single verbs, but for sets of verbs that refer to a common event, process, etc. In linguistic parlance, such verbs belong to the same semantic field. The initial phase in the process of building a relational thesaurus based on verbs thus entails (1) identifying an appropriate set of verbs, (2) grouping them by semantic field, and (3) determining the basic relational structure of each semantic field.

Verb identification in this study is based on frequency analysis data from the Brown corpus, a collection of 500 texts of approximately 2000 words each, representative of American English. Francis and Kučera (1982, 465-532) give a ranked frequency listing of the English word classes (words sharing stem and part of speech are merged into a single class), accounting for those in about the 85th frequency percentile and above. The 1250-odd verb classes from this list were
chosen as the basis of the study on the grounds that they would provide a sufficiently comprehensive sample to yield both the most general relational structures and some number of more specific structures; from the two, the variety of ways in which relational structures can be related to each other can be surveyed.

The grouping of verbs into semantic fields was approached from three distinct directions. The first approach used a clustering algorithm (Iwayama and Tokunaga, 1995) to look for co-occurrence patterns among the words found in the verbs’ definitions. This approach is based on the assumption that verbs in the same semantic field would tend to mention the participant within their shared relational structure, entities that would be unlikely to be mentioned in the definitions of other verbs (unless they belong to a related semantic field). While the basic premise has some merit, there are many phenomena—notably polysemy, synonymy, and inflection—that adversely affect the attempt to form semantic field clusters based on co-occurrence patterns in definitions. Simple modification of the process—the application of Porter’s (1980) stemming algorithm to control for inflection—proved inadequate. Ultimately, with major modifications incorporated to control for polysemy and synonymy, the approach might prove useful, but it is unclear that the labor commitment would be repaid. The second approach used verb groups from Levin (1993), which reflect certain aspects of verbs’ syntactic behavior. But, as we have seen, differences in syntactic behavior can and do occur between verbs in the same semantic field. Consequently, the approach failed to produce fruitful results. The third approach used verb groupings within the categories of Roget’s thesaurus (1911). Although this was the most subjective of the approaches, it proved the best starting point.

Determining the relational structure of a set of verbs results from identifying and characterizing their shared entity participants (i.e., their arguments), as we have already done with Commercial Exchange verbs. Several types of data sources are available to aid in this task. One type of data source presents synthesized information about verb arguments; examples are the verb frame information given in WordNet, an online lexical reference system,3 and Levin’s (1993) discussions of the semantics of verb classes. Dictionary definitions represent an intermediate type data source; they tend to make explicit mention of a verb’s arguments, but not always comprehensively. A third type of data source—the ultimate source, in that the other two source types rely on it—is corpus data in which the occurrence of various arguments/participants with verbs is exemplified.

3. Inventory of Basic Relational Structures

A concern that arises in building a relational thesaurus based on the argument structure of verbs is that the number of relational structures thus identified could be rather large. Several phenomena cooperate to reduce the potential complexity of the relational thesaurus to a manageable size: (1) The verbs within a common semantic field may share relational structure, as observed with Commercial Exchange verbs. (The multiplicity of verbs provides a means for focusing attention on different aspects of the relational structure.) (2) A major component that is transferred in metaphor is relational structure; the structure of a more concrete domain is borrowed to structure our understanding of a more abstract domain. (3) The specificity of verbs ranges broadly; we have both the extreme generality of go and be and make and the considerably greater specificity of meander and continue and manufacture. But the same basic relational structure of an agent moving from one point to another is present in both go and meander, the same basic relational structure of an object and some associated state is present in both be and continue; and the same basic relational structure of an agent engaging in a process to create an object, optionally using material and/or an instrument, is present in both make and manufacture.
The convergence of these phenomena sharply reduces the number of relational structures to be accounted for.

These factors combine to urge the identification of most basic relational structures as a starting point in the actual development of the thesaurus. Figure 1 introduces such an inventory of basic relational structures. Although small in number, this set of structures appears to account for the relationships underlying nearly all the verbs under examination. These relational structures are generally of two kinds. Approximately a half dozen of the more abstract structures (State, Action, Process, Action-Process, Benefactive, Locative, and Experiential) correspond to verb types identified by Chafe (1970), under the influence of Fillmore's case grammar. Many of the other structures (Link, Path, Container, Center-Periphery, Balance, and, from under Action-Process, Compulsion, Blockage, Removal of restraint, and Counterforce) have a more concrete basis and correspond to image schemata discussed by Johnson (1987). (For simplicity's sake, I have adopted their labels.) The basic inventory thus correlates positively with previous work with the semantics of verbs and with relational structures.

4. Organizational Structure within the Relational Inventory

The organizational structure of the conventional thesaurus is largely, often exclusively, hierarchical: BT/NT links between descriptors account for most of the structure imposed by the thesaurus. By its very nature, the relational thesaurus recognizes additional relationship types, but the hierarchical relationship continues to wield the greatest force as an organizational device in the relational thesaurus.

In the context of entity classes, hierarchy is based on the notion of attributes: a narrower entity class inherits all the attributes of any broader entity class. Relationships are more complex than entities, and it turns out that some of the links readily recognized as hierarchical relationships in the relational context require richer explanation. As the relationships between more general and more specific verbs are examined, two patterns emerge. According to pattern 1, specificity is introduced when the identity of entities participating within the relationship is constrained to any degree (the more constrained, the more specific); it is in this sense that calve is more specific than (give) birth. As an extension of pattern 1, specificity is introduced when attributes of entity participants or of the overall relationship are constrained; it is in this sense that meander is more specific than go. According to pattern 2, specificity is introduced when multiple relational structures combine to form a more complex relational structure; it is in this sense that one meaning of capture is simultaneously more specific than both force and enclose.

An examination of relational structures within the Path branch of the inventory demonstrates the broader sense of hierarchical relationship needed in this context. The basic Path image schema consists of 3 components: (1) a starting point, often called the Source; (2) an ending point, often called the Goal; and (3) a sequence of points, a Route, between the source and the goal, which is the focus of the structure.

The Journey relational structure incorporates the whole of the Path structure and adds an entity participant, a Traveler who traverses the Route between Source and Goal, optionally adding a Vehicle which conveys the Traveler along the path. The traversing itself is motion along the Route, which movement is in turn a specific kind of Action. The Vehicle is a specific kind of Instrument. Instruments are normally considered an optional entity participant in Action-Processes, so here we either need to recognize that Instruments can also play a participant role in Actions or we need to recognize that motion which is aided by a Vehicle is no longer just an Action, but has become an Action-Process. Journey is thus related to Path in a complex fashion: it not only incorporates additional entity participants (Traveler, Vehicle), but also is thereby
<table>
<thead>
<tr>
<th>RELATIONAL STRUCTURE</th>
<th>DESCRIPTION OF BASIC RELATIONAL STRUCTURE</th>
<th>REPRESENTATIVE EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>Entity has value of attribute (e.g., emotion, height, weight)</td>
<td>mourn, tower, weigh</td>
</tr>
<tr>
<td>Action</td>
<td>Entity engages in activity</td>
<td>dance, jump, run</td>
</tr>
<tr>
<td>Process</td>
<td>Change from one state (e.g., size, color, emotion) to another</td>
<td>enlarge, darken, discourage</td>
</tr>
<tr>
<td>Action-Process</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Force</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compulsion</td>
<td>Force vector acts upon entity, to move it along path, toward or away from force</td>
<td>attract, pull, press, push</td>
</tr>
<tr>
<td>Blockage</td>
<td>Entity resists force vector acting upon it Potentially restraining entity removed/disabled</td>
<td>block, interfere, restrain</td>
</tr>
<tr>
<td>Removal of restraint</td>
<td>Force vectors oppose each other</td>
<td>escape, liberate, release, rescue</td>
</tr>
<tr>
<td>Counterforce</td>
<td>Entity causes other entity to come into existence or undergo significant change</td>
<td>oppose, conflict build, form, generate, make</td>
</tr>
<tr>
<td>Production</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benefactive</td>
<td>Entity is beneficiary of action or state</td>
<td>benefit, inherit, own</td>
</tr>
<tr>
<td>Locative</td>
<td>Entity is located relative to a landmark (e.g., above or below it, before or behind it, near to or far from it)</td>
<td>sink, confront, adjoin</td>
</tr>
<tr>
<td>Experiential</td>
<td>Entity experiences mental disposition</td>
<td></td>
</tr>
<tr>
<td>Volition</td>
<td>Entity makes choice</td>
<td>choose, intend, prefer</td>
</tr>
<tr>
<td>Cognition</td>
<td>Entity engages mental faculties</td>
<td>amaze, imagine, think</td>
</tr>
<tr>
<td>Perception</td>
<td>Entity engages perceptual faculties</td>
<td>gaze, listen, taste</td>
</tr>
<tr>
<td>Link</td>
<td>A is connected to B</td>
<td>connect, link, marry</td>
</tr>
<tr>
<td>Path</td>
<td>A sequence of points leads from A to B Entity moves from A to B</td>
<td>arrive, follow, return</td>
</tr>
<tr>
<td>Journey</td>
<td>Possession of entity moves from A to B</td>
<td>go, walk, wander</td>
</tr>
<tr>
<td>Gift</td>
<td>Possession of entity moves from A to B, with compensation</td>
<td>distribute, give, accept, receive</td>
</tr>
<tr>
<td>Commercial Exchange</td>
<td>Message moves from A to B Knowledge moves from A to B</td>
<td>pay, purchase, charge, sell</td>
</tr>
<tr>
<td>Communication</td>
<td></td>
<td>ask, tell, listen, read</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td>instruct, teach, learn, study</td>
</tr>
<tr>
<td>Container</td>
<td>Boundary separates inside from outside</td>
<td>enclose, exclude, fill</td>
</tr>
<tr>
<td>Center-Periphery</td>
<td>Focal point contrasts with surroundings</td>
<td>center, concentrate, focus</td>
</tr>
</tbody>
</table>
Balance
Comparison
Justice

Correspondence between A and B
Correspondence between action and reward/punishment

compare, equal, match
assess, judge

Hierarchy
Subordination
Genus-species
Whole-part

One entity is subordinate to another
One category is wholly included within another
One entity is part of another
govern, rule, submit
generalize, include.
specify
consist, consolidate,
integrate

Fig. I: Preliminary Inventory of Basic Relational Structures

transformed from a static image schematic pattern into a dynamic Action or Action-Process.

The Gift, Commercial Exchange, Communication, and Education relational structures in turn are dependent on Journey, in that each conveys movement of some Traveler from a Source to a Goal. In prototypical Gifts and Commercial Exchanges, a concrete entity undergoes physical movement, as when a present (Traveler) is physically handed by the Giver (Source) to the Recipient (Goal) or when Merchandise is physically transferred from Seller to Buyer as part of the Commercial Exchange. But not all Gifts and Commercial Exchanges involve such physical motion. Consider, for example, the giving of a promise or the selling of a birthright. Similarly, in Communication and Education, it is (the meaning of) a Message and Knowledge, respectively, that are conveyed from Source to Goal. Thus, the movement of the Traveler does not always involve physical motion, but may be metaphorical. Moreover, even with the physical motion of Gifts and Merchandise from Source to Goal, it may be that the real entity transferred is an abstract one, Possession. In this sense, some Goals may also become Beneficiaries. In other words, Gifts and Commercial Exchanges not only have the relational structure of Journeys, but also incorporate the relational structure of Possession. Similarly, the relational structure of Cognition is also built into Education. The upshot is that these relational structures are yet more complexly related to the Journey relational structure than Journey is to Path. First, they specify something of the nature of the Traveler. Second, the movement of the Traveler may be literal or metaphorical. Third, several broader relational structures interact in these more specific relational structures, possibly in ways that go beyond the capacity of multihierarchical inheritance to account for.

5. Summary

A relatively compact set of general relational structures form the basis of the large number of relationships encodable in English. The interconnections among the larger inventory of relationships promise to be very complex and stretch our current understanding of the nature of the hierarchical relationship. Extensive further investigation into both the identification and characterization of relational structures and the organizational structure of those relationships is called for. Better control over the expression of relationships portends higher precision of retrieval in cases where searches involve multiple entity participants that could occur in multiple relational configurations; higher recall could be achieved in cases where the relational structure itself is the object of search.
Notes
1. The purview of the paper is actually somewhat larger than relationships: it extends to all predicates, but emphasizes those with 2 or more arguments.
2. This research is funded in part by a 1995-1996 OCLC Library and Information Science Research Grant.

References
Abstract: Implicit associative relationships among Medical Subject Headings (MeSH) relevant to clinical anatomy were investigated to determine the extent to which they could be characterized, organized, and structured, and the degree to which this process might be automated was assessed. The 256 term pairs were clustered into five broad semantic categories that reflected both the linked terms and the nature of the relationship(s) between the terms: Procedures, (other) Anatomic Entities, Functions and their Disruptions, and Chemical Agents. The most frequent category, Procedures, was characterized by a prevalence of morphologically-derivative relationships expressed as neo-classical compounds, with a potential to support automated semantic classification. Results suggest the need for an expressive representation structure rich in both entity types and non-hierarchical associative relationships.

1. Introduction

In a given information system, the exact meaning of a concept is determined by the context in which it occurs; the relationships a concept has with other concepts in the system will define its context, and thus its meaning. To the extent that the relationships among concepts are explicitly expressed, most controlled vocabularies focus on hierarchical relationships. Far less attention is paid to non-hierarchical, or associative relationships. Where they are identified, they are rarely further characterized beyond being grouped with other similarly non-specified relationships under such descriptive designations as "related terms," "see also," "see related," or "other." The larger vocabularies may contain hundreds or thousands of these relationships. Although the existence of such implicit non-hierarchical associative relationships has long been formally recognized by explicit reference, this is of little practical use when the precise meaning or significance of the relationship between the concepts or terms in the context of the vocabulary is unknown because it is only implied. While the nature of the relationship may be or become apparent to subject experts or experienced users of the vocabulary, novices to either the content area or the workings of the particular information system are at a loss to discern meaning. The importance of a more thorough understanding of the nature and use of inter-concept relationships, and some current research activity in this field has been addressed in some detail by Molholt (1996) and Green (1996). Analyzing and characterizing the nature of associative relationships within the specific contexts or applications in which they occur will supplement previous work to enumerate and classify them from a more global perspective (e.g., Neelameghan & Maitra 1978; Soergel 1967), providing additional detail and further dimension to our understanding of these important relationships.

Because the conceptual content of a controlled vocabulary covers an expressly delimited subject domain, the relationships among concepts may also be expected to exhibit a certain domain dependency. For example, a medical model would contain such general relationships as <part_of>, <causes>, and <precedes> that would be common to almost any other
subject domain; it will also include relationships such as <treats> and <diagnoses> that might not be included (or necessary) in representations of other types of knowledge (McCray & Nelson 1995). Furthermore, the relative proportion or frequency of particular relationships will vary among subject domains, presenting a different relationship profile for each. It seems likely that there might exist within controlled vocabularies conceptual groupings of similar, related inter-concept relationship types, particularly among the implicit non-hierarchical associative relationships. Significant manifest clustering would warrant formal specification and characterization. Classifying and expressing these in a consistent, systematic, and meaningful way would allow us to take advantage of the semantic knowledge inherent in the vocabularies. The present research is developing and assessing the usefulness of a methodology to characterize implicit associative relationships that are explicitly identified in controlled medical vocabularies, and investigating the extent to which this methodology can be automated. This is part of a long-term effort to compile a comprehensive inventory of meaningful and consistent associative relationships that operate in the medical domain, toward development of conceptual models and tools for integrating distributed knowledge base systems in medicine. Current advances in technology support development of knowledge bases capable of expressing and thus exploiting this rich source of semantic information.

2. Present Research

A subset of implicit associative relationships among Medical Subject Headings (MeSH) was investigated to determine the extent to which they could be characterized, classified, organized, or structured in some systematic fashion, and the degree to which this process might be automated (i.e., whether the relationships form regular, predictable clusters that can be identified and manipulated by machine). Non-hierarchical associative relationships are indicated by "see related" references in the print MeSH, and are thus explicitly recognized, but the nature of the relationship between terms is not specified, and thus implicit.

Reviewed and revised in annual updates by the National Library of Medicine for subject access to the biomedical literature, MeSH is among the largest and best-known controlled medical vocabularies. MeSH is included in the Unified Medical Language System (UMLS) Meta-Thesaurus, and there serves as the main authority for Preferred Term selection from among synonymous terms and concepts. When the data set used in the present study was obtained, the corresponding edition of the UMLS (1994) contained in its Meta-Thesaurus close to half a million interconnected terms for almost 200,000 biomedical concepts, compiled from 26 source vocabularies. The Semantic Network provides a unifying structure of high-level categories (or "semantic types"; 133 in 1994) to which Meta-Thesaurus concepts are mapped, that are linked by semantic relationships (50 in 1994). The present research uses this high-level structure for broad semantic categorization, and considers potential relationships from the perspective of the contemporaneous version of the UMLS; changes in subsequent editions that are relevant to the test set will be also be addressed. MeSH terms are indicated by the use of italics and UMLS semantic types by uppercase; relationships are set off using angle brackets.

This analysis has concentrated on anatomical terminology to restrict the parameters of investigation to a relatively narrow subject domain and functional enterprise within medicine. Clinical (or "gross") anatomy in medicine is a discrete, well-circumscribed subdomain that crosses boundaries of both content and function, with a conceptual basis that is both well-established and relatively static. The findings of this research are being applied in the development of a conceptual model and knowledge base rich in inter-concept relationships to support the first-year anatomy course component of the integrated electronic medical
3. Approach

The test set of focal terms was derived from the set of all descendants of head terms for the "system" trees from the 1994 MeSH A-Anatomy axis, excluding the Nervous System (i.e., A2-A7), and including Body Regions (A1) and Sense Organs (A9). The 536 terms thus identified account for most of the MeSH concepts relevant to the domain of gross anatomy. Entries for each of these terms in the annotated alphabetic list in the print version of MeSH were examined to find all terms linked by "see related" cross references, yielding a total set of 256 such related term pairs for 140 unique terms. Not surprisingly, a large proportion of the focal terms (68.4%) correspond to the UMLS semantic type BODY PART, ORGAN, OR ORGAN COMPONENT; the remainder correspond to BODY LOCATION OR REGION (10.7%), TISSUE (6.0%), CELL (4.7%), BODY SYSTEM (4.0%), BODY SPACE OR JUNCTION (4.0%), and to HAZARDOUS OR POISONOUS SUBSTANCE (0.7%), PHARMACOLOGIC SUBSTANCE (0.7%), and CELL COMPONENT (0.7%). (Numbers total more than 100% due to multiple semantic types for some concepts. For example, *Bone and Bones*, *Muscles*, and *Pericardium* all correspond to both semantic types BODY PART, ORGAN, AND ORGAN COMPONENT and TISSUE.)

The term pairs could be grouped into five broad clusters or semantic categories based on how the second, or related term of the pair was related back to the original, or focal term: Procedures (n=142), (other) Anatomic Entities (n=52), Functions (n=31), Disruptions to Function (n=14), and Chemical Agents (n=16). One additional pair exhibited a relationship unique in this analysis. Each broad category will be considered in turn to discuss pertinent examples of predominant relevant relationship types or principles of organization identified for that group.

4.1 Cluster 1: "Procedures" Terms

In the largest group, which accounted for more than half the pairs in the entire test set (n=142, 55.5%), the related terms referred to a therapeutic, preventive, or diagnostic procedure performed on the focal anatomic entity. Within this group, in eight subgroups accounting for the bulk of the procedure terms (n=123, 86.6%), a consistent pattern was noted: the related terms were neo-classical compounds consisting of one or more anatomical referents (root morphemes) joined with a combining form (terminal morpheme) that designated a specific procedure performed on the focal anatomic entity. Compounds contained Greek, Latin, and English root forms. Thus, some of the root morphemes were lexical derivatives of the focal anatomic entity (*Bronchi*: bronchoscopy and *Trachea*: tracheotomy), while others were conceptual or synonymous derivatives (*Nose*: rhinoplasty and *Kidney*: nephrectomy). The eight terminal morphemes, with their common meanings and frequencies, were -ectomy (excision, removal; n=29), -graphy (writing, description; n=28), -stomy (mouth, opening; n=22), -scopy (viewing with an instrument; n=19), -plasty (forming, shaping; n=12), -otomy (incision, cutting; n=10), -metry (measuring; n=2), and -desis (fusion, binding; n=1).

In fourteen of the nineteen remaining (non-neo-classical) term pairs, the related term likewise names a particular procedure performed on the focal anatomic entity. In some cases the referent is not named, but rather implied, probably because the procedure is specific only, or chiefly, to that anatomic entity (*Gonads*: castration and *Skull*: trephining). The final four
pairs refer generally to surgery on the focal entity.

4.2. Cluster 1: "Procedures" Relationships

The contemporaneous UMLS (1994) allowed no relationships between semantic types denoting anatomical and procedural entities. Presumably this reflects the more general conceptual viewpoint that procedures act by affecting the functioning of anatomic structures. Concept classes can only be related within the UMLS Semantic Net using binary relationships; expressing complex or composite concepts, which join multiple unitary or atomic concepts, may only be accomplished through a series of binary relationships. This obscures any direct relationship that may exist between procedure and anatomy. Using this system to express the above relationship that exists between a therapeutic procedure and an organ, one must chain together the four binary relationships: THERAPEUTIC OR PREVENTIVE PROCEDURE <treats> PATHOLOGIC FUNCTION; PATHOLOGIC FUNCTION <is_a> BIOLOGIC FUNCTION; BIOLOGIC FUNCTION <has_location> FULLY FORMED ANATOMICAL STRUCTURE; FULLY FORMED ANATOMICAL STRUCTURE <inverse_is_a> BODY PART, ORGAN, OR ORGAN COMPONENT. Although this expresses an important conceptual element of the relationship, it neglects an equally valid aspect in a very real sense: namely that many procedures are direct actions on some particular anatomic entity, affecting it both physically as well as functionally. In fact, it may be more realistic to view a procedure as having some physical effect on an anatomic entity, which in turn leads to a change in function. This also highlights the necessity of making careful distinctions between high-level models and low-level reality, and in the application of general principles to specific instantiations.

The current UMLS (1996) does allow a direct relationship between these two semantic types, but only one denoting a spatial location, for example, compressing the above sequence of four binary relationships into a single one: THERAPEUTIC OR PREVENTIVE PROCEDURE <has_location> BODY PART, ORGAN, OR ORGAN COMPONENT; this still limits the expression to only one aspect of what can be a very complex multifaceted relationship or set of relationships. In the UMLS, procedures are semantically distinguished by the general motivation for undertaking a procedure, i.e., for therapeutic, preventive, or diagnostic purposes; the terminal morpheme clusters observed here argue for more distinct groupings according to the means by which that objective may be accomplished. This suggests not only a more specific level of motivation for processes in the conceptual schema for the medical domain, but also a set of distinctive relationships that reflect the precise actions of procedures on anatomical entities. For example, within the set of THERAPEUTIC AND PREVENTIVE PROCEDURES, we may first subclassify the terms by whether they serve to <join> (-desis) or separate anatomic parts, and then among the latter, whether they <excise> (-ectomy), <incise> (-otomy), or <create an opening in> (-ostomy) some part. Likewise, there exist subcategories among the anatomical entity types that may be linked in a meaningful way to only certain subtypes of procedures; for example, an -ostomy would only be performed on a BODY PART, ORGAN, OR ORGAN COMPONENT that is hollow or tubular.

The derived morphological clusters have inherent meaning associated with the combining form. These suggest specific classificatory divisions along which to form and align the new and unique relationships relevant to this particular subdomain. These term pairs can be considered to have the relationship between anatomy and procedure embedded in the morphological features of the related term. The prevalence of neo-classical compounds among
these associative relationships thus suggests not only lexico-semantic clustering, but also the potential for automated semantic classification of a substantial portion of such relationships: a simple pattern-matching program could be used to automatically classify the neo-classical compound clusters described based on string matchings of regular and consistent morphological features of the terms. Limitations of precision may arise because neo-classical compounds tend to be underdetermined by their constituent parts (McCray et al 1988), but this method should be robust enough to provide at least preliminary semantic classification of many relationships, and form the basis for a more sophisticated representation. Such an approach should be particularly fruitful in the present context, given the highly nominalized sublanguage of medicine.

5. Cluster 2: "(Other) Anatomic Entities" Terms

Of the 52 term pairs providing links within or among anatomic entity types, more than half (55.8%) reflected one or more physical or spatial relationships between BODY PART types or between BODY PARTs and regional or spatial concepts, which were well-accounted-for by "legal" UMLS relationship types. The rest were widely distributed among the remaining semantic types, with too few per category for definitive analysis; however, the relationships observed did seem to be consistent both with "common sense" and the UMLS.

This cluster provides evidence that common morphological features can also be used to suggest in the semantic context a broader relationship that may be more relevant than the specific one embedded. For example, a small cluster of three term pairs shares the common word "bones" (Head:facial bones, Metatarsus:metatarsal bones, and Tarsus:tarsal bones). The relationship that characterizes each of these term pairs is that the second term refers to the bones within the body site named in the first term. Five other term pairs may be added to that cluster because they are characterized by the identical relationship within the same semantic context (Head:skull, Heel:calcaneus, Heel:talus, Neck:cervical vertebrae, and Thigh:femur). However, upon consideration, this relationship may be somewhat more specific than is optimally useful, because it essentially reflects precoordination of specific atomic concepts into a composite term; for example, would not a similar relationship exist between specific muscles or vascular elements within particular body sites? Looking for other occurrences of term pairs in similar semantic contexts yields six more (Arm:forelimb, Heel:achilles tendon, Mandible:chin, and Jaw:-, Mandible:-, and Mandibular Condyle:- with temporomandibular joint). Examining the relationships in these pairs along with the others reveals other potentially more useful physical or spatial relationships (e.g., <has_part> or <contains> or <location_of>) that characterize the notion of specific body parts having particular associations with specific body sites. (The last four example pairs above reflect the inverse of the relationship between participants, with interesting semantic implications of the relationships between positive and negative spatial entities.)

6. Cluster 3: "Function" Terms

A total of 31 term pairs linked focal anatomic entities to corresponding physiological functional concepts such as organ or tissue function and organism function. For 25 of these, the UMLS relationship <location_of> is allowed (by virtue of inheritance), which is useful when referring to the site where functional process occurs (Pylorus:gastric emptying). In addition to accounting for this spatial relationship, it would also be useful to capture or express a specific functional attribute, such as differentiating the relative roles played by organ
components (*Kidney Glomerulus: renal blood flow, effective*) from an over-arching role played by all the components combined in and of the whole entity itself (*Kidney: diuresis*). (Oddly, no legal relationship exists between a BODY SYSTEM, which is defined as a functional concept, and any functional semantic type.)

7. Cluster 4: "Disruptions to Function" Terms
Fourteen term pairs contained relationships relevant to disruptions to function, such as diseases, injuries, and symptoms. Although there were too few in this cluster to develop meaningful subgroups, it seem likely that subclassification by general pathological processes such as neoplasia or inflammation would yield useful axes along which to define specific relationships with anatomic entity types, in a manner similar to that for Procedures. Further, it is likely that using neo-classical semantic classification techniques (e.g., -itis, -plasia) would prove fruitful in this group as well.

8. Cluster 5: "Chemical Agents" Terms
Half of the sixteen related terms in this cluster were pharmacologic agents, composed of a referential root designating the anatomic entity that is the object of the agent (sometimes implied) joined with the pharmacologic action of the agent on the object (*Bronchi: bronchodilator agents* or *Pupil: miotics*). The remainder portrayed relationships in which the focal anatomic entity produces (*Endocrine Glands: hormones*), or in the case of anatomic substances, is produced by or composed of the related entity (*Cartilage: chondroitin*).

9. Discussion
This research has demonstrated the existence of clusters of conceptually-related relationships among MeSH associative relationships linking anatomical concepts with the terminology of other conceptual categories. Some of these may be expressed by using "legal" relationships among the corresponding semantic types in the UMLS Semantic Net; however, for others, this does not satisfactorily express the relevant aspects of the relationship in the surrounding context. Results suggest both the need and a means to fill the gap existing between high-level conceptual models where both the categories and the relationships between them are so broadly expressed that important distinctions may be obscured, and the low-level term instantiation where both concepts and relationships are too narrow and precisely focused to be generalizable beyond the specific instantiation. Evidence has also been presented indicating the potential utility of automated semantic classification using morphological analysis of neo-classical compound terminology.

Further research will assess the expression of larger sets of a greater variety of these conceptual relationships. Research already underway is examining the "other" relationships in the UMLS for terms assigned to these and additional semantic types relevant to anatomy, drawing terms (focal and related) from all 40+ vocabularies currently represented in the UMLS. This will be valuable both for evaluating existing systems and structural models, and especially so for designing new ones that adequately reflect the semantic structure of a knowledge domain through a richly expressive system of categories and relationships. Beyond the need to know what relevant relationships are operative in a given representation or knowledge model, it is also critical to determine how they "work." That is, whether they have some consistent operating logic, such as regards transitivity, for example, would be important.
in implementing vocabularies in computerized information systems, because of the increased storage and retrieval efficiency that can be gained by applying subsumption and transitive "inheritance" principles.

Notes

1. This work was supported in part by NLM Training Grant LM07079. Thanks are extended to Rebecca Green, Pat Molholt, and James J. Cimino for their insightful comments regarding the research reported in this paper.

References


Hypertext and Indexing Languages: Common Challenges and Perspectives

Abstract: Conceptual considerations on possible synergies between hypertext and indexing languages’ research. The existence of a common conceptual and theoretical background encourages and facilitates cooperation. Researchers and developers in both fields face also common challenges in an environment of rapid technological and sociological change. The possibilities of crossed fertilization in both directions are carefully explored. It is concluded that he conjunction of the almost biological flexibility of the hypertext and the logical and conceptual consistency of indexing languages is a promise of better and more effective information management systems in the future.

1. Introduction

As it is well known, hypertext is an alternative to alphabetical and sequential access to information. Hypertext is also able to integrate dynamically quite different kinds of data—text, numbers, graphics, films and sounds—in what is called hypermedia. Finally, hypermedia seems to be the near future for all global information systems, due to the ever growing size and popularity of the WWW. But hypertext design may improve if the knowledge from classification and indexing research is applied. Inversely, old problems in documentary control may be solved with the aid of hypertext technologies.

In fact, both documentary languages and hypertext share a common problem —optimizing the representation and organization of universal and/or specific domain knowledge—and the same theoretical background, based on cognitive sciences and social epistemology. In both cases, their epistemological models rest on the constructs of concept and term, and the concept of relationship, that are called in hypertext systems nodes and links. Both of them rely on the same principle of non sequential access to information. Both of them are in pursuit of a universal organization of knowledge, like the universal classifications or Vannevar Bush’s Memex. Finally, both approaches have in common a flavor of utopia.

So, it is not strange that a basis for cooperation may be found in both directions. First, documentary languages—as vocabulary and conceptual control tools, and repositories of relational information—may provide the cognitive maps that can help us to fight two well-known problems of hypertext systems: cognitive overload and disorientation. Use of thesauri will lead to easier and wider retrieval and to improved automatic generation of hypertext relations. Classifications, on the other hand, will provide the universal frame for true conceptual retrieval, not simply an alphabetical one. This is of the greatest importance in a biologically growing environment like the Web, whose growth is not lead by central planning.

Inversely, hypertext opens new possibilities to the different documentary languages. First, hypertext can improve the interfaces on which the automation of indexing and classification systems relies. Second, the two-layers’ theory for hypertext systems can provide a frame for considering the connection of the rational and pre-fixed lexical structures of the documentary languages and the idiosyncrasy and personal cognitive maps of the users. Third, hypertext can be
used to link associative and hierarchical languages, and those with texts and references, creating truly global and integrated information systems. Also, hypertext interfaces can facilitate the tools for designing and maintaining thesauri and classifications.

Finally, indexing theory and hypertext offer not only a potential for cooperation, but also share common challenges. Both of them require increasingly dynamic systems for logical and graphical representation, able to cope with the multidimensionality of the worlds of concepts. Both of them have as a common goal the development of platforms for cooperative and concurrent work. Both of them need to integrate automatic knowledge and linguistic processing. And finally, both hypertext and indexing theory are to gain a deeper understanding of the complementarity of the hierarchical and associative approaches to the organization of information systems.

2. A Common Problem and Theoretical Backgrounds

a) Though, as we have actually said, documentary languages and hypertext systems seem at a first glance two completely different things, both of them share a common underlying problem: optimizing knowledge representation, organization and transfer as a response to the information overload that our societies suffer.

In this sense, they are nothing more than two knowledge technologies, that find a common ground in the theoretical disciplines that study their very same problem, this is, cognitive sciences, Epistemology, etc.

b) The response to this common problem has been in both cases the development—more precisely, its importing from other sciences devoted to basic research—of a common theoretical background on the way that human beings acquire, represent and store information in the form of knowledge (from Cognitive Psychology) and how the society manages, co-ordinates and determines this knowledge (from Social Epistemology).

In fact, information scientists have given attention over all to one field of Cognitive Psychology: the study of the writing and reading processes. But information scientists have also made an important contribution to Social Epistemology: they have created an image of the Society of Knowledge as a macro-text, linked like a web by the conceptual relations existing among terms and the very genealogical relations existing among the documents themselves (The society of text, 1989).

c) In both cases, their core epistemological models are based on the constructs of concept and relation, which are called in the hypertext terminology nodes and links. As in thesauri, hypertext relations may be single or bi-directional, and also hierarchical or associative.

These relations can be labeled in quite more complex ways, as occurs in some quite complex documentary languages, like PRECIS.

Generally speaking, hypertext is an effort to transcend the limits of sequential communication and alphabetically organized information systems. Documentary languages achieve also this goal by means of their a priori structure.

d) Certainly, documentary languages and hypertext share a common core principle: non-sequentially in information access.

That is what Bush (1945) called “associative indexing” and what he proposed as the main feature of his “Memex”, that is, that each item may be linked to any other, and the other way, too. Both systems aim at providing the universe of knowledge, which is mainly in a textual and sequential form, with a meta-textual organization, able to integrate every text in a super-structure that would surround, surpass and transcend it.

e) The above-mentioned reasons explain also that both disciplines use a common language.
As Churcher suggested in 1989, hypertext specialists have adopted the very language that it is used in Psychology, Artificial Intelligence, Pedagogy or Information Science—the so-called Cognitive Sciences—to describe knowledge and information structures.

f) Finally and differently from other information technologies, both systems share a strong utopian component.

Certainly, hypertext systems, on one hand, and universal classifications and inter-connecting indexing languages—as the BSO, for example—on the other, aim at building tools for global knowledge organization. Not other was the driving idea of Dewey, Ranganathan, Bliss, Bush or Nelson.

Anyway, indexing languages and hypertext do not only share a common theoretical foundation, that leads them to similar views of the world of knowledge, but also, as we will see now, many possibilities of synergy and common challenges.

This conceptual and empirical proximity is not strange if we consider some common historical backgrounds. For example, thesauri and hypertext systems have a precedent in encyclopedias, dictionaries and complex and multi-referenced texts.

3. Contributions of Documentary Linguistics to Hypertext

Hypertext creates well-known problems either to the administrator of the system, either to the final user. These problems are the cognitive overload and disorientation (Conklin, 1987).

Documentary languages are, precisely, the tools that can help us to overcome this handicaps. Indexing languages are tools for terminological control and they are also deposits of lexical relational information. Because of these, they conform true knowledge maps and models for conceptual organization that may improve navigability in hypertext environments.

The possible contributions of indexing languages for better hypertext systems can be summarized in these four points:

a) Terminological (or vocabulary) control. According to Frank Halasz (1988), hypertext systems should be able not only to provide with a navigational access to information, but also with suitable search and retrieval tools. Indexing languages, specially thesauri, are the tool for controlling search terms. In general, hypertext systems may get profit from every form of vocabulary control, from the pure morphological to the lexical ones. Thus, for example, Hyperlexis (Otal et al., 1992), which is a prototype of an interactive hypertext dictionary for lexical learning, has three kinds of retrieval tools: keyword search, a database of lexical relations for each word, and an assisted hypertext with a parser that leads from the clicked word—whichever its morphological form may be—towards its entry.

b) Automated link generation. From the point of view of vocabulary control, documentary languages can constitute an excellent knowledge database for automatic node identification and link building. Considering the costs of manual link creation, using databases of technical vocabularies with relational information is the only alternative to build massive hypertext corpora with minimal conditions of profitability (Westland, 1991).

c) Knowledge organization (structuring of information). As soon as 1991, Urr prevented us against the limitations of hypertext systems. They lack a coherent and—as Fugmann (1994) would call it—predictable underlying conceptual system, which is necessary to achieve one of its more important aims: giving quick and effective access to information with independence of the intellectual background of the user.

In his paper, Urr considered an analogy between hypertext and OPACs, which are less effective when they lack carefully developed classification schemes. Nowadays, an almost general agreement exists that a well-build hypertext system must have a structure able to lead users
without needing any kind of formal or informal teaching about its use. In pursuit of this aim, several authors have proposed developing hypertext models with two or more levels, each of them teaching and giving access to the next and more difficult one (Agosti et al., 1992; Salminen and Watters, 1992). Documentary classifications could serve as conceptual superstructures, being common to the different local systems, and configuring a common cognitive interface to the different hypermedia products. In fact, multimedia products have been somehow criticized for their excessive idiosyncrasy and the low predictability of their relations. In this sense, this approach could complement other hierarchical approaches based on statistical models, as the one proposed by Botafogo et al. (1992).

d) Finally, as world-wide virtual hypertext systems develop in the Internet, completely new challenges in the field of documentary control are appearing. A brilliant example of this fact—nets of hypertext documents distributed throughout different computers along the globe—is the World Wide Web. As these nets grow, they will need progressively stronger knowledge organization tools and, of course, a tough work of terminological normalization. This is a new field for Knowledge Organization experts, and a very promising one, in fact. The possibility of studying logs of HTTP connections, for example, opens completely new fields in the empirical validation of knowledge organization structures.

Though all these are obvious considerations for any information scientists, very little work has been made in the above mentioned directions. As Gibbs (1992, 40) points out: “Hypertext is a technology which has caught the imagination of information scientists and authors alike. However, as with the first phases of expert systems, it has probably been oversold as a solution to all information production and retrieval problems. It has been shown to be effective in assisting users who have ill-defined information needs, or wish to discover information, but is generally less successful for goal-oriented searches. This is a direct result of the fact that hypertext research has concentrated on the preservation and creation of the explicit intellectual and structural links between text units which facilitate browsing. Despite observations concerning the potential of document surrogates for creating hypertext links (Reimer and Hahn, 1988), the indexing of text units has been generally ignored or conveniently overlooked by hypertext developers.”

4. Contributions of Hypertext to Classical Knowledge Organization

Reciprocally, the concept of hypertext is offering new possibilities of controlling the vocabulary relations that traditional documentary languages provide. In current practice, non-expert users find documentary languages—both thesauri and classifications, and, of course, more complex systems—difficult to use. Novel users are hardly able to take advantage of their conceptual rigor and their possibilities for better information retrieval. We think that hypertext environments can constitute the perfect platforms for the development of tools for both the professionals—modules for automatic management of thesauri and classifications, and the users—friendly retrieval interfaces.

Thus, hypertext systems might contribute to knowledge organization tools with the following advantages:

a) Make easier accessing to and using pre-coordinated languages. Hypertext systems can be the base for the friendly user interfaces to the documentary languages, especially the pre-coordinated ones, because hypertext interfaces usually combine keyword searching, alphabetical browsing and powerful relational navigation. Thus, for example, Bjorklund (1990) has studied the advantages and problems of using a hypertext system to make easier the use of the Swedish SAB for both users and professionals.

It is necessary, anyway, to have in mind that hypertext is more than a mere friendly user
interface technology. It is an alternative and complementary model in the access to information, apart from that of sequential, alphabetical or conceptual browsing. It is necessary to make clear the distinction between live hypertext, capable of storing (authoring) new relations, and fossilized hypertext, the set of relations that are provided by the creator of the system. It would be a pity to reduce hypertext to a mere 'wysiwyg' interface.

b) Connect the indexing languages with the cognitive maps of the users. As we have said before, a traditional problem with documentary languages is the difficulty to link their pre-defined structures—featured by their excessive crystallization and dependence from the ideology and intellectual backgrounds of their creator—with the idiosyncrasy and personal conceptual maps of the user, that differ from one person to the other, depending on its previous intellectual background, professional group, etc.

In this sense, hypertext has something new to offer to Knowledge Organization: the concept of path, which is a bridge between socially accepted knowledge and the individual processes of knowledge acquisition, which is always constructed in a personal and idiosyncratic manner (Novak and Gowin, 1984). In this sense, hypertext, as a living documentary system, suggest the storing of knowledge and learning paths from different points of view (disciplinary slanted, project slanted and so on). This would open a new dimension in the empirical research on indexing tools: it would be possible to compare the different paths established by users among them, and all of them with the indexing language. Stotts and Furuta (1991) have conceptualized the dynamic relation that exists in hypertext between the objective conceptual maps proposed by the creator of the system and subjective ones as the liking of two planes: the fixed one developed by the designer of the system—which is called by these authors fixed underlying information structure—and the flexible one, generated on the march (dynamically) by the user as it interacts with the system. The relation between documentary languages and the cognitive maps and linguistic uses of the users could be conceptualized in a similar manner. Thus, the documentary language could be adapted to the preferences of an information center or a group of users and even to a special user, but preserving an underlying vocabulary control and a common conceptual and lexical standard.

c) Dynamic interconnection of documentary languages among them, and of these with primary documents and references. Hypertext can be very useful to connect dynamically documents and their representations in the documentary systems, and thereafter with the documentary languages that give access to them.

As Alberico and Micco (1990, 180) stated: "Hypermedia provides the ideal tool for mounting and displaying bibliographic tools such as encyclopedias, Dewey and LCSN subject headings, and reference tools that can serve as guides to other information sources". Of course, hypertext interfaces for OPACs already exist, for example HYPERCAT of the Linköping University in Sweden (Hjerppe, 1986).

We have to keep in mind, however, that hypertext interfaces have not done as well in situations where there is a large, unfamiliar, heterogeneously structured network, such as traditional library databases, in which the user easily gets lost in the hyper-space (Alberico and Micco, 1990, 182). Use of hypertext in these environments requires ensuring the user with "intelligent" help, as it happens in RABBIT (Tou et al., 1982).

Finally, hypertext systems could probably be used for joining dynamically classifications and thesauri among them, creating little by little some sort of interconnection languages.

d) Design, building and management of thesauri and classifications. Hypertext environments can be used for developing and managing thesauri and classifications: its terms, relations and graphical and alphabetical presentations (Richard, 1990).
5. Common challenges

As we told in the introduction, indexing languages' experts and hypertext designers share some common challenges of great importance, too. Here are some of the most important ones:

a) Development of dynamical systems for logical and graphical representation of the relations into and among documents. From a merely documentary point of view, one of the best strategies to improve information retrieval with keywords and descriptors has been to include them within a context. This context can be a priori, like lexical relations in thesauri; a posteriori, pre-coordination of indexing terms; or a mixed solution. Creating context throughout classifications and conceptual maps is also the best tool against cognitive overload and disorientation of hypertext users.

From a more theoretical point of view, documentary languages and hypertext systems face the same challenges in the logical and graphical representation of very complex conceptual systems. Especially, if we consider that this representation must be dynamic, because it refers to a world that is changing continuously. The map of sciences and the relations between disciplines are changing all the time, and so subject fields and their terminologies.

The world of graphical representation of concept maps is a very interesting field where we have a lot to learn from mathematicians (Wille, 1984). One of the most central problems is to conciliate the complexity of the relations among concepts and easy-to-use and navigable graphical presentations. We have to do also more research in which are the most suitable graphical presentations from the point of view of the users. It seems that presenting the more immediate relations of the node that is being consulted is more useful for the users than the presentation of the whole classification schedules (Girill at al., 1991).

b) Compatibility and hospitality for concurrent and cooperative work. A typical critic that both hypertext and associative documentary languages have deserved is that they have a very idiosyncrasy character. Each hypertext system is different from the rest of them. And that occurs also among, for example, thesauri. This results in problems when retrieving and interchanging information and while doing cooperative work. This is a very important question with indexing languages, because their magnitude needs of the cooperation among great groups of experts, and a fitted balance between the centralization of control functions and the respect of the principle of subsidiarity. These needs could be better served by the possibilities offered by cooperative electronic edition and distributed databases under a common interface.

c) Incorporation of automatic knowledge processing. The incorporation of inference engines to these structures of objects that are, after all, both indexing languages and hypertext systems would be able to transform them in dynamic retrieval systems. They would be able to assist the user in the search process and the professional in the indexing and classification processes, specially when determining complex relations.

In this sense, expert systems, on one hand, and indexing tools and hypertext, on the other, can complement each other: Expert systems shells provide the technology able to support inference processes, while indexing and hypertext systems store conceptual and pragmatical knowledge.

d) Deeper research in the complementary between hierarchical and associative approaches to knowledge organization. A typical aspect of the evolution of knowledge organization research has been fluctuation between associative and hierarchical models. But nowadays there is a broader sense of their complementary. In fact, the first conceptualizations of hypertext where based in an extreme form of the associationist theory of human thinking and memory. This was the case of Bush's Memex.

Recently, most of the authors highlight the relevance of a strong hierarchical organization of hypertext systems. Inversely, documentary languages have evolved from extremely rigid
tree-structures—like the first universal classifications—toward the assumption of associativity and multi-dimensionality of the world of knowledge in thesauri and facet classifications. Certainly, from a broader point of view, it can be postulated that classification and association are complementary cognitive strategies in the task of representing and processing into knowledge (Garcia and Esteban, 1993).

In this line of complementary between associative and hierarchical approaches to knowledge organization, it is not strange that Jan Wyllie (1990) defended recently that the future of hypermedia relies on the integration of free-text retrieval technology and of, in a broad sense, the know-how and theoretical basis of the discipline of documentary classification.

6. Conclusion

The conjunction of the almost biological flexibility of the hypertext project and the logical and conceptual consistency of indexing tools is a promise of better and more effective information management systems in the future. This is why working for a closer cooperation among researchers in both fields is so needed. While this occurs, research is needed in two directions. On one hand, we need to improve the theoretical background of both information technologies and connect their common references and referents with basic research carried on in the field of Cognitive Sciences. On the other, we have to implement practical models of cooperation between hypertext and knowledge organization tools.

Notes

1. Aims of such a system are quite meaningful: allowing the browsing of records and the navigation among them, building better structured catalogs than the traditional ones; offering alternative systems to show and illustrate the record structures and their information and relations; developing tools for establishing links and navigate among them; creating a dynamic system in itself, not only by continuously adding new records; ensuring each user the possibility of building and storing its own navigational routes throughout the library, so that the system itself can be enriched by the utilization that users do of it; allowing the storage of user models, both stereotyped and individualized, enriching the traditional information that catalogues offer, through relations among fields, records and files; and finally, offering information not only about individual documents, but also about the collections themselves.

References


Visual Dewey: DDC in a Hypertextual Browser for the Library User

Abstract: A new design for the Dewey Decimal Classification (DDC) allows for display of the schedules and indexes in a hypertextual browser which was originally designed for the more conventional thesaurus with BT-NT, RT and Use-UF relations. Features of the DDC which approximate those relations and the presentation of the rich vocabulary found in the hierarchical structure, captions, index entries and notes are represented for the library user's perusal and choice before searching a library's holdings. By dragging and dropping parts from such a display into a search window, the search for bibliographic items can begin and be revised painlessly. Information such as class numbers or subject headings for retrieved items can be dragged and dropped into the search window or the Visual Dewey display window as it suits the user to review vocabulary or revise the search results. This seamless to-ing and fro-ing allows for a truly interactive and spontaneous search environment, with maximum assistance at point of need. "File folders" exist for saving any Visual Dewey information, any retrieved items, etc.

1. Introduction and Background

More than a decade has gone by since the first results came in about user difficulties with online public access catalogs (OPACs) (Ferguson et al., 1982; Hildreth, 1983; and Markey, 1983). Had we taken these findings seriously and begun the suggested improvements then, maybe we would not find the situation still as bad as ever. In our opinion, the technologies at our disposal at the time were not sufficiently advanced to provide the functionalities required. What is frustrating now is to find that system designers today, with better technologies, have learned little or nothing from those early OPAC user studies, from the analytical papers on preparing classification and thesauri for use online (Chan, 1986; Cochrane and Markey, 1985; Richmond, 1974; Svenonius, 1983; Wajenberg, 1983; Hill, 1984), from the early attempts to mount thesauri and classification systems into retrieval/search systems (Freeman and Atherton, 1968; Olson, 1974; Rodriguez, 1978; Markey and De Meyer, 1986; Bjorklund, 1990; Allen, 1994; Liu and Svenonius, 1991). Is the half-life (or the institutional memory) of the library and information science literature less than a decade? How sad. Neither the continuous stream of international conferences on classification research nor the conferences on Library Classification at Allerton in 1957 and 1994 (Cochrane, 1996), nor the numerous Subject Access conferences sponsored by the Council on Library Resources in the 1970s, have delivered the message that library users need more help than they are receiving from the current designs of OPACs and other retrieval systems and that presenting classification information may help.

A brief summary of these early research findings is found in Table 1. We have added a recent finding by Palmer (1996).
Participants in some of the focused interviews conducted as part of the online public access catalog evaluation project talked about the idea of a “knowledge tree” which could help a user broaden or narrow a search (Cochrane, 1985, 381). All of the devices listed in Table 1 could help form such a knowledge tree.

Carole Palmer (1996) described the information gathering habits of members of an interdisciplinary research team as “seeking, gathering, and probing.” These “boundary crossers,” as she called them, often tried to “foreground the periphery” and then focus on the core of their discipline orientation. As they needed to accumulate new knowledge at the boundaries of their work with others in the team, they had to scale multiple cognitive structures and go from a novice level of knowledge acquisition to an expert level in very short periods of time. Often their knowledge of the vocabulary of related fields was minimum. They needed multiple meaningful domain displays and some way to handle their rapid accumulation of new knowledge.

Do we provide proper access to information stores for such seeking, gathering, and probing? How would we improve our search systems so that user difficulties could be avoided? Maybe a partial answer is provided by better vocabulary displays and easier interaction between displays of related terms and the information retrieved from catalogs and indexes. That at least is what we have attempted to do, first with a hypertextual thesaurus browser (complete with keyword and KWIC windows), then this combined with an automatic "concept space" display and a classification scheme.

### Table 1: Needs and Devices for Seeking, Gathering, and Probing

<table>
<thead>
<tr>
<th>NEED After or before Keyword search:</th>
<th>HELPFUL DEVICES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Narrow the search after too broad a term yields too many hits</td>
<td>Thesaurus, classification</td>
</tr>
<tr>
<td>2. Broaden the search after too narrow a term yields too few (or zero) hits</td>
<td>Thesaurus, classification</td>
</tr>
<tr>
<td>3. How to organize choice of subject terms (Markey, 1983)</td>
<td>HOLD file, classification</td>
</tr>
<tr>
<td>4. How to find &quot;correct&quot; terms (Hildreth, 1983)</td>
<td>KWIC, thesaurus, classification</td>
</tr>
<tr>
<td>5. How to view related terms (Cochrane, 1985; Daily, 1972)</td>
<td>KWIC, thesaurus, co-occurrence lists, indexing records in relevant retrieved items, classification outline</td>
</tr>
<tr>
<td>6. Link text words to controlled vocabulary (Hildreth, 1983)</td>
<td>(See Johnson, 1995; Schatz et al., 1996)</td>
</tr>
<tr>
<td>7. How to easily cross disciplinary boundaries (Palmer, 1996)</td>
<td>Polyhierarchical displays, switching vocabularies and HOLD files during searching, drag and drop</td>
</tr>
</tbody>
</table>

2. Need for Better Vocabulary Displays

Figure 1 shows a display from the first online classification system used as a retrieval language. Developed in the mid-1960s and called AUDACIOUS (Freeman and Atherton, 1968), the best we could do with the existing technology was a permuted keyword display followed by a class number in the schedule display. The searcher would then have to retype
the class number in command mode for retrieval of documents. This served as an example for the DDC Project (Markey and Demeyer, 1986).

**Online Search**

> Radioactivity

- RADIOACTIVITY
- RADIOACTIVITY IN THE ATMOSPHERE 539.16
- RADIOACTIVITY IN THE EARTH 550.378
- RADIOACTIVITY IN OCEANS AND SEAS 551.48:539.16
- RADIOACTIVITY IN RIVERS 551.48:539.16
- RADIOACTIVITY STANDARDS 539.16.081

> 551.510.7

**Explanation**

The user inputs Radioactivity. The system responds with lines from the UDC schedule where the term Radioactivity appears.

The user is interested in Radioactivity in the Atmosphere and inputs its number. The system response is a list of lines from the UDC schedules in the neighborhood of that UDC number, 551.510.7, thereby showing some of the hierarchy.

Confer (cf.) notes suggest alternate placement of this concept.


**Fig. 1: UDC as an online browsing tool for subject access.**

Other efforts included Charles Hildreth’s experimental online catalog with a menu bar that allowed for “Bookshelf Browsing” in the Library of Congress Class Outline (Hildreth 1993).

As library classification schedules became available in electronic form ways were devised for this information to be presented to users. Electronic Dewey for edition 20 of the DDC (Electronic Dewey, 1993), on CD-ROM, was designed for classifiers as was the current classification schedules on CD-ROM from the Library of Congress.

By 1996 producers of these electronic versions were well into designing their products using graphical user interfaces (GUIs) and hypertext links. Because of the complexity of the data they were trying to present, readability and ease of use may be serious problems (see figure 2). Classification displays were designed for the printed page, for easy manipulation of pages, back and forth between index, auxiliary tables, summaries, manuals or footnotes and schedules. This same type of “flipping” is not as easy in the electronic environment, as the rich interconnectedness of hypertext can cause the “art museum phenomenon” and multiple overlapping windows can cause “death by windows.”
The problems of handling the myriad notes, index entries, and nested summaries of hierarchical ranges was discussed in 1985 (Cochrane and Markey). In this paper a list of the minimal data elements for a classification schedule, its auxiliary tables, index, and outline took up five pages (Cochrane and Markey, 1985, 102-106). How to display these data was not answered by the MARC format for Classification schedules and it appears that special editing will be necessary before displays for end users can be devised.


R. A. Wall (1980) contrasted the advantages and disadvantages of hierarchical classifications and thesauri. If either is used alone to display searching vocabulary it will come up short, just as keyword searches or controlled vocabulary searches are insufficient alone. A thesaurus may have more specific terms than a classification schedule and its index, but the hierarchies in a thesaurus are often too short and broad terms are infrequently posted. On the other hand, classification notation provides a map of search routes from broad to narrower numbers so that a full sweep of an area can be made easily,

Fig. 2: Alcoholism displayed in Dewey for Windows (OCLC/Forest Press)
but the cryptic messages about related areas to search are confusing to all but expert classifiers. (Mann, 1993)

Eric Johnson has developed a hypertextual thesaurus browser which allows for an interactive term suggestion display, using subject thesauri and co-occurrence lists, rotated word-in-context and keyword displays (Johnson, 1995; Johnson and Cochrane, 1995; Schatz et al., 1996; see also figures 3-4). Such a design and the power of new technologies may now provide what is needed for the classification display for a library user. With a drag and drop mechanism from any field within the retrieved document (figure 4) back to the thesaurus or classification “window” in Figure 3 most of the needed devices listed in Table 1 become real and the controlled vocabularies become as easy to use as keyword searching.

4. Ready DDC for a Visual, Interactive and Hypertextual Display

In his 1967 paper Victor Rosenberg concluded that “the preference for a given method [of information retrieval] reflects the estimated ease of use of the method rather than the amount of information expected.” That appears to have held true with keyword searching being preferred over any other means yet devised. Instead of lamenting the fact that users are not retrieving the best relevant items this way, we think the time has come to make other searching mechanisms and displays as easy to use. Because the thesaurus records were “more ready” for hypertextual browsers (Johnson and Cochrane, 1995,
Fig. 4: Hypertextual display of search, retrieved document, and hold file. The thin arrow indicates a mouse click on a short record, which then displays the full bibliographic record; the thick arrow indicates a drag-and-drop operation that adds the subject term B-ISDN to the hold file.

figures 1 and 2, pp. 79, 80) we started there. We began with the INSPEC Thesaurus and went on to the Legislative Index Vocabulary (LIV) of the Library of Congress/Congressional Research Service (see figure 5).

In 1996 we decided to explore the possibility of mounting a library classification scheme in this hypertextual environment. We preferred this to the top down approach (e.g., first DDC summary, second DDC summary) which has been used in so many Webpages and browsers on the Internet. OCLC/Forest Press have encouraged us by giving us the DDC database in electronic form for research purposes.

To illustrate how our work is progressing, we will use the three areas of the DDC schedule represented by three subjects: Alcoholism, Malaria, and Solar System. These illustrate both the promises and the difficulties of readying DDC for such a visual display.

Alcoholism: The Relative Index of DDC shows several class numbers for Alcoholism: 362.292, 616.861, 344.04461, 613.81, 291.17832292. Qualifiers show the context: social welfare, medicine, law, personal health, and social theology, respectively (Search #2 in Figure 2, above, shows seven hierarchies, but we will show only the five referred to in the Relative Index in our example.). Starting with 613.81, the easiest to explain in terms of transformation, has the following “record,” with data elements for entry number (en), entry heading (eh), and index entry headings (ieh):
What this record does not show is the hierarchy which places this topic in the 613.8 Substance abuse hierarchy (similar to Figure 5). This could be done in a manner similar to following the BT-NT relationships in a conventional thesaurus, but can not be done easily until the class numbers have been linked up and down the hierarchy as shown in figure 2. Such visual displays and summary tables in the printed DDC are a big help for showing context and hierarchy.

In LIV, which is used to index policy literature for Congress, there is only one broader term context, namely, Social science/Crime and criminals/ Crimes without victims (Figure 5). In the DDC, alcoholism appears in five hierarchies which the hypertextual browser must show or the other contexts will be missed and can not help the searcher who may really have other contexts in mind.
Fig. 6: Proposed display of DDC search for alcoholism. Note similarity to figure 5. The KWIC display is in a floating window (shown at the left, but can be repositioned). The phrases in the KWIC window come from a scan of notes in the DDC schedules. The hierarchy on the right-hand side of the figure comes from the display at 613.81, in the Personal health hierarchy. Other hierarchies are accessible via the list in the box at upper right. Clicking on the name of another hierarchy (e.g. Law) would display that hierarchy (344.04461). Scope notes and related areas are shown at upper left.

Figure 6 illustrates how the DDC information could be displayed. Analysis of all pertinent DDC records had not been completed at the time of writing, but we remain confident that the richness of the records can be used and made interesting to searchers.

For Malaria, the DDC record again appears simple, but the nfa note complicates the display in the hypertextual browser:

en 616.9362
eh *Malaria
nfa *Add as instructed@bunder 616.1-616.9
ich Malaria@2medicine

Nfa is a note to “add as instructed” from another part of the classification. This centralizes the instruction to the classifier to handle all diseases alike, using standard subdivisions with certain modifications. It also forces the classifier to read other notes at “the top of the hierarchy” about how to handle therapy, rehabilitation, public measures for
prevention, social factors contributing to the spread of a disease, etc. References are also
given to parts of the manual which could help the classifier make decisions about two
possible numbers which could be used. Incorporating all this information is one of the
most difficult parts of converting the DDC for use in a hypertextual browser. In this paper
we do not attempt to illustrate how the browser would handle such a record, but suffice it
to say the “related areas” in the display would be the way into such information.

The term Solar System in DDC leads to a complete hierarchy and an interesting way
of linking subjects via auxiliary tables. The core records for the Solar System in DDC are:

```
en  523.2
eh  Solar system
nse For@bspecific parts@c,see@d#523.3-523.7
ieh Solar system
\\
cen 523.3-523.7
eh  Specific parts of solar system
nce Class@bcomprehensive works@c@cin@d#523.2
\\
ten 2.99
eh  Extraterrestrial worlds
ndf Worlds other than Earth
nce Class@bspace@c@cin@d#2.19
nsn see Manual at@b2.19 vs. 2.99
ieh Extraterrestrial worlds@ya51
ieh Solar system@y151
\\
tcn 2.991-2.994
eh  Solar system
nce Class@bcomprehensive works@c@cin@#2.99
```

The richness of the DDC numbers for describing complex subjects comes from
building numbers composed of schedule and table numbers. For example volcanic activity
on Mars would be 551.21099923, using the class number for volcanoes with the number
from Table 2, representing Mars, the planet. We want to alert the user to such phenomena
as a “related area” and hope someday that we could parse DDC numbers to uncover all
the incidents of such number-building.

Cochrane and Markey (1985) enumerated many of the problems which occur
because of many notes and the nesting of DDC hierarchies. DDC editorial work will have
to be done to make all of these hierarchic ranges explicit and to convert classifier's notes
for use with the catalog user. We think we are going in the right direction when we leave
the DDC numbers in the background (unless explicitly requested), focusing instead on
displays of the rich hierarchical and related areas, using the index entries as lead-in
vocabulary to the right display(s) for that term, showing polyhierarchies, when they exist,
in an unobtrusive manner (see figure 6). We intend to continue our research taking
advantage of feedback from Forest Press, the DDC Editorial Office, and participants at this conference.

5. Movement between Retrieval and Vocabulary Selection Mechanisms

Links back and forth between the bibliographic records and these hypertextual displays should be very straightforward since the class number is separated from the Cutter/author-book number in the MARC record field for this data. Class or call number searching has been a feature in many online catalogs since the late 1960s (for example, the SCORPIO "browse call" command allows you to search LCC class numbers in the bibliographic records at the Library of Congress). We would intend for the class number portion of the bibliographic record to be dragged and dropped onto a classification window for displays similar to that in Figure 6. Then the library searcher would be free to seek, gather, and probe.

Smith et al. (1989, 252) remarked that "a number of issues arise when an intermediary system is given thesaurus-type knowledge. Issues arise concerning organizing perspectives, means of access to hierarchy nodes (top-down versus direct activation), focusing of attention (pruning of hierarchies), and learning. Such computerized intermediaries, whether they should be considered 'intelligent' thesaurus systems or simply useful thesaurus systems, clarify some of the issues that bona fide knowledge-based systems must deal with. Thesauri contain certain types of knowledge that must be dealt with in designing a knowledge-based system." We couldn't agree more. The only change we would make in Dr. Smith's quote would be to add the word classification wherever the word thesaurus appears.

References
Cochrane, Pauline Atherton. (1965). Ranganathan's Classification Ideas: An Analytico-Synthetic Discussion. Library Resources & Technical Services, 9, 463-473. (First written as a talk at CRSG June 1960 meeting; reprinted in Cochrane [1985], 300-311.)


Classification to the Rescue—Handling the Problems of Too Many and Too Few Retrievals

Abstract. The first studies of online catalog use demonstrated that the problems of too many and too few retrievals plagued the earliest online catalog users. Despite fifteen years of system development, implementation, and evaluation, these problems still adversely affect the subject searches of today's online catalog users. In fact, the large-retrievals problem has grown more acute due to the growth of online catalog databases. This paper explores the use of library classifications for consolidating and summarizing high-posted subject searches and for handling subject searches that result in no or too few retrievals. Findings are presented in the form of generalizations about retrievals and library classifications, needed improvements to classification terminology, and suggestions for improved functionality to facilitate the display of retrieved titles in online catalogs.

1. Introduction

The first studies of online catalog use demonstrated that the problems of too many and too few retrievals plagued the earliest online catalog users (Tolle, 1983; Larson, 1983). The former problem has grown more acute due to the growth of online catalog databases through the addition of new library acquisitions, retrospective conversions, and proliferation of databases.

A recent study demonstrated that a little more than half (56.4%) of user queries for subjects exactly matched assigned subject headings in the Library of Congress subject headings system (Drabenstott, 1994, 45). Overall, about one in every three of such matches resulted in retrievals (i.e., subject headings or bibliographic records) exceeding 100. High retrievals were especially a nuisance for users whose queries exactly matched geographic subject headings—almost two of every three such queries resulted in over 100 retrievals. By themselves, such findings are not that enlightening. When we consider them in view of findings about low levels of user persistence displaying as few as 35 or as many as 100 or 200 titles (Wiberley and Daugherty, 1988; Wiberley, Daugherty, and Danowski, 1990, 1993), we realize that online catalogs are penalizing the most successful of online catalog users—those who enter matches of the catalog's controlled vocabulary—by overwhelming them with large numbers of retrievals. Generally, online catalog responses in the form of long lists of subject headings or titles provoke searchers into terminating their ongoing search prematurely, entering an entirely new search, or walking away from the online catalog terminal.

This paper explores the use of library classifications for consolidating and summarizing high-posted subject searches in online catalogs. It also examines the use of classifications to assist users whose subject searches produce too few retrievals or too few useful retrievals.

2. Literature Review

Several approaches to alleviating the problem of large retrievals have been offered. Some researchers (Larson, 1989; Prabha, 1990; Lynch, 1990; Buckland, 1992) have experimented with system functionality that allows users to summarize or reduce their retrievals by criteria that do
not necessarily touch upon subject. Other researchers (Massicotte, 1988; Drabenstott and Vizine-Goetz, 1994) have recommended the assignment of broad category codes to enable systems to summarize large numbers of subdivided subject headings by broad subjects. Still other researchers (Markey, 1984; Cochrane, 1986; Larson, 1989; Hildreth, 1993; Walker and De Vere, 1990) have recommended and/or tested the usefulness of classification as a subject retrieval device to aid searchers whose searches are plagued by too many or too few retrievals.

An early attempt to link subject heading displays and classification numbers was produced in the form of a book catalog called the *Subject Guide to the Collections of the Public Library of Columbus and Franklin County* (PLCFC, 1978). Assigned subject headings were listed in alphabetical order followed by an ordered list of classification numbers and the number of books in the library's system assigned the listed subject heading. PLCFC's *Subject Guide* gave library patrons an alternative to scrolling through the library's microfilm catalog. It encouraged patrons to consult this single-volume book catalog, find a subject heading of interest, and browse a particularly fruitful shelf area where several books on their topics of interest would be found according to the statistics in the guide. Unfortunately, explanatory captions from the classification schedules did not accompany listed class numbers so patrons did not know until they arrived at the shelves whether their topics of interest were treated from their desired perspective or point of view.

The Dewey Online Catalog (DOC), an experimental online catalog that featured the Dewey Decimal Classification (DDC) as a searcher's tool for subject access, browsing, and display, featured the subject outline search that consolidated retrievals in ways similar to PLCFC's *Subject Guide*. DOC also included captions in its displays of particularly fruitful areas of the library classification where patrons could find materials on their topics of interest (Markey and Demeyer, 1986). Searches in the experimental CHESHIRE system demonstrated how classification can be used to manage high-posted searches (Larson, 1989). Although OCLC's Electronic Dewey Decimal Classification (EDDC) is targeted at library catalogers, it has an interesting feature—not much different from PLCFC's Subject Guide and DOC's subject outline search—that allows librarians to search for subject headings and retrieve class numbers that DDC catalogers at the Library of Congress have assigned to large numbers of bibliographic records. If available to library patrons, such information could help them find materials on their topics of interest. Catalogers could also use such information to assign DDC numbers to library materials.

Classification can be used as a device to find additional retrievals in low-posted searches. Research using experimental online catalogs developed separately by Hildreth (1993) and Walker and De Vere (1990) demonstrated the usefulness of classification for relevance feedback. This technique has captured the imagination of designers of commercial online catalogs who have added capabilities for users to switch from the display of a bibliographic record that was the result of a subject search based on subject heading or keyword retrievals to a display of titles in the same classification neighborhood as the original title.

A preliminary USMARC format for classification authority data has been available for a half decade providing the keepers of library classification data with a vehicle for expressing such data in machine-readable form and enabling them and library organizations to share such data amongst themselves. Other than adding the capability to switch between subject searching results and bookshelf browsing, designers of commercial systems hardly paid any attention to library classifications as a tool for subject access, browsing, and display.

This paper adds more fuel to the fire, so to speak, in terms of providing a rationale for the incorporation of library classifications in online catalogs. Their job would be focused on summarizing large retrievals and on helping users find additional retrievals in low-posted searches.
3. Methodology

To study the problem of too many retrievals, the researchers chose several moderately high-posted subjects entered by users of academic library catalogs. This paper features searches for "acid rain" and "costa rica;" the full-length report features three additional searches (Drabenstott, 1994). We chose queries that were exact matches of controlled vocabulary terms because such terms, especially geographic terms, usually result in high-posted searches. We searched these subjects in the online catalogs of Duke University and the University of Michigan—two catalogs with sizable databases—using these catalogs' subject heading searches that listed retrieved titles for subdivided and unsubdivided forms of the matched subject heading. Retrievals in the Duke and Michigan catalogs were classified in the Dewey Decimal (DDC) and Library of Congress Classifications (LCC), respectively. We consolidated retrievals in classification number order and summarized large numbers of retrievals using captions from classification summaries and outlines. We examined two levels of consolidation using classification.

To study the problem of too few retrievals, we chose queries that yielded few or zero retrievals in keyword searches of academic library catalogs. We examined the elements in the few records retrieved, especially in useful records, to determine how systems could use such elements to find additional useful records.

4. Using DDC to Summarize Too Many Retrievals

Acid Rain

The query "acid rain" retrieved 136 titles assigned the unsubdivided subject heading "Acid rain" and subdivided forms of this heading. Duke retrievals were distributed into 37 different classification numbers. There were 17 unique classification numbers beginning with the same three-digit number. The three most common numbers and percentages of retrievals per number were 363.7386 (25.7%), 363.7392 (14.7%), and 363.7394 (8.8%). Overall, these three classification numbers accounted for a little under 50% of retrieved titles. Table 1 enlists first-level DDC captions to consolidate the 136 retrieved titles.

<table>
<thead>
<tr>
<th>DDC no.</th>
<th>No./titles</th>
<th>Classification captions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
<td>Generalities</td>
</tr>
<tr>
<td>3</td>
<td>96</td>
<td>Social sciences</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>Natural sciences and mathematics</td>
</tr>
<tr>
<td>6</td>
<td>21</td>
<td>Technology (Applied sciences)</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>The arts</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>Geography and history</td>
</tr>
<tr>
<td>0-9</td>
<td>136</td>
<td>Total</td>
</tr>
</tbody>
</table>

Table 1: First-level DDC Captions for "acid rain"

The majority of retrievals were found in the social sciences section of the Dewey Decimal Classification. About 15% of retrievals were found under the "Technology (Applied sciences)" caption. For users interested in this topic from the perspective of technology, such a breakdown might have been useful for users to pursue. If users selected the "Social sciences" caption where the majority of retrievals occurred, systems would summarize retrievals using second-level captions (Table 2).
Table 2: Second-level DDC Captions for “acid rain” in the Social Sciences

The second-level display might not be as helpful as the first-level display because the vast majority (87.5%) of retrievals were centered in one classification area. Also, the caption “Social services; association,” was not too descriptive.

Costa Rica

The query “costa rica” retrieved 477 titles assigned the unsubdivided subject heading “Costa Rica” and subdivided forms of this heading. Duke retrievals were scattered into 142 different classification numbers. There were 58 unique classification numbers beginning with the same three-digit number. The most common number was 972.86; however, it accounted for only 16.4% of titles. Overall, the top five ranked classification numbers in terms of numbers of retrieved titles accounted for only 37.3% of retrieved titles. Table 3 enlists first-level DDC captions to consolidate the 477 retrieved titles.

Table 3: First-level DDC Captions for “costa rica”

Large numbers of retrievals were split between the “Social sciences” (300s) and “Geography and history” (900s) sections of the Dewey Decimal Classification. If users selected the former, retrievals would be summarized using second-level captions (Table 4). A manageable number of retrievals were given under all but one second-level caption so that users might be tempted to display brief title information under one or more listed captions. If users selected the geography and history (900s) section of DDC, almost all retrievals would occur under a single caption.
5. Using LCC to Summarize Too Many Retrievals

Acid Rain

The query “acid rain” retrieved 141 titles assigned the subdivided subject heading “Acid rain” and subdivided forms of this heading. MIRLYN retrievals were distributed into 30 different Library of Congress Classification numbers. There were 25 unique classification numbers beginning with the same number disregarding alphanumeric characters beyond the decimal point. The most common number was TD196.A25 which retrieved 51.8% of the titles bearing “Acid rain” subject headings. In our analysis of summarization using the DDC, no single number accounted for more than 42.5% of retrievals (“greek sculpture”) (Drabenstott, 1994). The class number “QH545.A17” accounted for about 14% of retrieved titles and came in a distant second in magnitude to the most frequently-occurring classification number. Table 5 enlists first-level LCC captions to consolidate the 141 retrieved titles.

<table>
<thead>
<tr>
<th>LCC no.</th>
<th>No./ titles</th>
<th>Classification captions</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>6</td>
<td>Social sciences</td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>Political science</td>
</tr>
<tr>
<td>K</td>
<td>3</td>
<td>Law</td>
</tr>
<tr>
<td>Q</td>
<td>32</td>
<td>Science</td>
</tr>
<tr>
<td>S</td>
<td>5</td>
<td>Agriculture</td>
</tr>
<tr>
<td>T</td>
<td>88</td>
<td>Technology</td>
</tr>
<tr>
<td>Z</td>
<td>6</td>
<td>Bibliography and Library Science</td>
</tr>
<tr>
<td>A–Z</td>
<td>141</td>
<td>Total</td>
</tr>
</tbody>
</table>

Table 5: First-level LCC Captions for “acid rain”

The majority of retrievals were found in the “Technology” section of the Library of Congress Classification. This was different from the result for DDC retrievals in which most retrievals were concentrated in the Social sciences class (Table 1). Manageable numbers of retrievals occurred in all classes but the Technology class. If users selected low-posted classes, systems would display brief-titles lists. If they selected the Technology class, systems could summarize retrievals using second-level captions; unfortunately, such captions do little to
summarize retrievals because all but one title occurs under one of the two second-level captions. At this point, users would be better off using a summarization technique other than classification, e.g., date, library branch. A third-level display would have been more helpful than a second-level display for summarizing retrievals (Table 6).

<table>
<thead>
<tr>
<th>LCC no.</th>
<th>No./titles</th>
<th>Classification captions</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA418</td>
<td>1</td>
<td>General works [Elastic properties and tests]</td>
</tr>
<tr>
<td>TD195</td>
<td>10</td>
<td>Special industries, facilities, activities, etc.</td>
</tr>
<tr>
<td>TD196</td>
<td>73</td>
<td>Other environmental pollutants, A–Z</td>
</tr>
<tr>
<td>TD427</td>
<td>1</td>
<td>Special pollutants and organisms, A–Z</td>
</tr>
<tr>
<td>TD883</td>
<td>1</td>
<td>General works [Air pollution and its control]</td>
</tr>
<tr>
<td>TD885</td>
<td>2</td>
<td>General works [Gases. Flue gases]</td>
</tr>
<tr>
<td>TA–TX</td>
<td>88</td>
<td>Total</td>
</tr>
</tbody>
</table>

Table 6: Third-level LCC Captions for “acid rain” in Technology

The third-level display distributed titles into six classes. The majority (83.0%) of titles occurred in a single class (TD196). Remaining classes featured ten or fewer titles. The same, awkward wording was used for three of the six LCC captions, viz. “General works.” Placed in brackets and accompanying these three phrases in Table 11 are captions at the next highest level of the classification which give “General works” captions context.

Costa Rica

The query “costa rica” retrieved 344 titles assigned the unsubdivided subject heading “Costa Rica” and subdivided forms of this heading. MIRLYN retrievals were scattered into 109 different classification numbers. The most common number was “HC143.” It accounted for only 11.3% of titles. Overall, the top three ranked classification numbers in terms of numbers of retrieved titles accounted for only 30.8% of retrieved titles.

Fifteen first-level LCC captions were needed to consolidate the 344 retrieved titles. About half (177 titles) occurred in class F (“History of North and South America”) and a little less than a quarter (74 titles) occurred in class H (“Social sciences”). Remaining titles were distributed into thirteen classes in which the number of titles in classes ranged from 1 to 23. Large numbers of retrievals were split between the “History of North and South America” (F) and “Social sciences” (H) sections of LCC. The split between these two frequently-occurring classes (F and H) was the same as the split for retrievals in the DDC (Table 3). Numbers of retrievals in other classes were manageable; if users selected them, systems would show brief-titles lists. If users selected the captions representing the two high-posted classes, lists with two different distributions result. Since class F has no classes beginning with two-letter combinations, classification number ranges could be used to represent second-level classes. When this approach is enlisted, over 90% of retrievals were classed in the “Costa Rica” range (F1541–F1557).

A total of thirteen second-level classes summarized retrievals in the “H” (“Social sciences”) class. One such class (“HC”) summarized 56.8% of retrievals in second-level classes. The remaining retrievals were distributed into a dozen second-level classes ranging from one to seven retrievals. If low-posted classes (i.e., classes with one or two retrievals) were combined into a single “Other social sciences” class, the number of second-level classes could be reduced to nine classes. Two second-level classification captions (HC and HD) were worded exactly alike, viz. “Economic history and conditions.” Retrievals could be combined into a single caption; however,
some searches might have different results in which the separation between retrievals from the two classes might be helpful. To reduce end-user confusion, new name(s) should be given to one or both captions. The caption term “Woman” under HQ could be offensive to some end users. Second-level results in the Social sciences demonstrate the need to review LCC schedule terminology to improve understanding and reduce awkward or potentially offensive terminology.

6. Findings—Too Many Retrievals

Overall, we searched five moderately high-posted subjects to explore how the Dewey Decimal and Library of Congress Classifications would summarize large retrievals. The subjects were “acid rain,” “costa rica,” “greek sculpture,” “pornography,” and “racism.” (This paper featured two of these five subjects.)

Numbers of titles retrieved for these five subjects in Duke's online catalog ranged from a low of 111 titles (“pornography”) to a high of 477 titles (“costa rica”); numbers of titles retrieved for these five subjects in Michigan's online catalog ranged from a low of 141 titles (“acid rain”) to a high of 497 titles (“racism”). The average number of retrievals per unique, retrieved DDC classification number ranged from a low of 2.1 titles (“racism”) to a high of 5.2 titles (“greek sculpture”) and per LCC classification number ranged from a low of 2.1 titles (“racism”) to a high of 4.8 titles (“greek sculpture”). The lowest percentages of titles retrieved for the most common DDC number and LCC number for a particular subject were 11.8% for the synthesized number “305.800973” under “racism” and 11.3% for “HC143” and “HT1521” under “costa rica” and “racism,” respectively. The highest percentages of titles retrieved for the most common DDC number and LCC number for a particular subject were 42.5% for “733.3” under “greek sculpture” and 51.8% for “TD196.A25” under “acid rain.” Although our analysis of subject heading and classification number retrievals was far from comprehensive, the analysis resulted in several generalizations.

Generalizations about retrievals were:

- Retrievals were distributed into many unique classification numbers across many broad disciplines.
- Retrievals were likely to result in one common classification number; however, the percentage of retrievals for such numbers could vary considerably, i.e., between ten and fifty percent of the titles assigned the subject heading.
- Retrievals were not likely to result in more than two common classification numbers that, together, retrieved more than half of the titles assigned the subject heading.

When we applied first-level DDC captions to summarize retrievals, a total of only six captions were connected with an unmanageable number of retrievals. One subject featured two first-level captions with an unmanageable number of retrievals (“Costa Rica”) and the remaining four subjects featured a single first-level caption with an unmanageable number of retrievals. When we applied first-level LCC captions to summarize retrievals, a total of only ten captions were connected with an unmanageable number of retrievals. Two subjects featured two first-level captions with an unmanageable number of retrievals (“costa rica” and “greek sculpture”) and one subject featured three first-level captions with an unmanageable number of retrievals (“racism”).

When users select captions with unmanageable numbers of retrievals, systems could use various techniques to facilitate user browsing and targeting of relevant titles. For captions connected with a hundred or fewer titles, they could display titles in reverse chronological order. For captions connected with more than a hundred titles, they could give users options for displaying titles by date, format, language, library branch, other non-subject criteria, and combinations of these criteria. When the original query that provoked the classification summary
is an exact match of a subject heading, systems could summarize retrievals using subdivided forms of the matched subject heading. If large numbers of subdivisions are retrieved, systems could use the available three or four types of coded subdivisions, i.e., topical, period, geographic, and, possibly, form, to summarize them.

We would have liked to have shown such a display of subject headings but the catalogs we searched did not feature subject heading searches that could be limited to specific classification numbers and sorted by subdivided forms of the matched heading. Future explorations of the use of classification to summarize retrievals should experiment with such displays. They should also focus on very high-posted subjects—subjects that retrieve several thousands of retrievals—because, in addition to using classification for summarization, such results are likely to also require summarization of subject headings.

For the most part, first- and second-level classes did a satisfactory job summarizing large numbers of retrievals. Numbers of retrievals were not evenly distributed amongst listed classes. A handful of classes summarized the majority of retrievals. Summarization by classification numbers could be especially helpful to users who find a first- or second-level class that treats their topics of interest from a less-than-common perspective.

Generalizations about classification terminology:
- Generally, DDC terminology would be understandable to end users.
- LCC terminology was sometimes not suitable for conveying the subjects of classification numbers which were assigned to many titles bearing the same subject headings. Considerable editorial work would be necessary to improve the wording of captions at all levels of the Library of Congress Classification.

On occasion, it is likely that all the typical methods of summarization—classification, subject headings, year of publication, language—would be exhausted and numbers of retrievals would still be unmanageable. To facilitate browsing in such searches, catalogs should feature display techniques that allow users to browse retrievals as quickly as possible. For example, the capabilities of "balloon help" from the Apple Macintosh's graphical user interface and of "show URL's" from Mosaic's graphical user interface could be extended to online catalog interfaces. When end users dragged a mouse over particular locations on the screen, detailed explanations of the underlying text would be given on the screen. Imagine dragging a mouse over a brief-titles list bearing truncated author names, call numbers, and titles. When users dragged the mouse over the titles, the system would display the full titles in a balloon overlay. In this way, end users could drag the mouse down a column of a dozen truncated titles, call numbers, author names, etc., and read the information that interests them in balloon overlays in a matter of seconds.

7. Using Classification to Increase Retrievals

We used the same approach as experimental online catalogs developed by Hildreth (1993) and Walker and De Vere (1990) to find additional retrievals in low-posted searches. That is, we used one relevant title to find additional ones. Our results were similar to these previous researchers, that is, this approach was not always successful because it was based on a classification number in a single title. This approach was more successful when several relevant titles were assigned the same classification number. The problem was that there were so few titles in low-posted searches on which to launch shelflist-browsing searches.

Table 7 summarizes three searches in which browsing titles in the same DDC or LCC classification area as retrieved, useful ones led to the retrieval of additional useful titles. The left-hand column gives the original query followed by terms used in the keyword-in-record search; the plus sign (+) indicates where explicit truncation was applied.
<table>
<thead>
<tr>
<th>Query/Search statement</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>arts fund raising/ arts fund raising</td>
<td>3 of 5 titles retrieved in a keyword-in-record search are potentially useful: <em>ArtsMoney: raising it, saving it, and earning it</em>, <em>Successful fundraising for arts and cultural organizations</em>, and <em>Cash in!: funding and promoting the arts</em>. All 3 titles bear the class number “700.681.” Searches of this class number result in two more titles, viz. <em>National guide to funding in arts and culture and Non-profit enterprise in the arts</em>.</td>
</tr>
<tr>
<td>gays in the military/ gay+ and military</td>
<td>5 of 23 titles retrieved in a keyword-in-record search are potentially useful: <em>Conduct unbecoming: lesbians and gays in the U.S. military</em>, <em>Vietnam to the Persian Gulf</em>, <em>Gays—in or out?: the U.S. military &amp; homosexuals: a source book</em>, <em>Barrack buddies and soldier lovers: dialogues with gay young men in the U.S. military</em>, <em>Gays and the military: Joseph Steffan versus the United States</em>, and <em>My country, my right to serve: experiences of gay men and women in the military</em>. All five are assigned the number 355.008664 or 355.0086642. Searches of these class numbers result in only one more title, viz. <em>Torn allegiances: the story of a gay cadet</em>.</td>
</tr>
<tr>
<td>women and dieting/ women and dieting</td>
<td>The only title retrieved in a keyword-in-record search is potentially useful: <em>Beyond dieting: psychoeducational interventions for chronically obese women</em>. Searches of its class number RC552.025 result in four promising titles. Examples are <em>Fat oppression and psychotherapy: a feminist perspective</em>, <em>Such a pretty face: being fat in America</em>, and <em>Fat is a feminist issue: the anti-diet guide to permanent weight loss</em>.</td>
</tr>
</tbody>
</table>

Table 7: Using Classification Numbers to Increase Retrievals

8. Findings—Too Few Retrievals

The three searches summarized in Table 7 described seemingly successful searches in which class numbers assigned to titles retrieved in keyword-in-record searches led to the retrieval of additional, potentially useful titles. This table failed to show the following problems with this strategy for finding additional useful titles:

- The many subject searches we conducted in which class numbers assigned to retrieved titles did not lead to additional useful titles.
- The many subject searches we conducted which led to the retrieval of a handful of useful titles that were all assigned different class numbers.
- The many class number searches we conducted which led to so many titles with the same class number that browsing title lists to find potentially useful ones was too tedious and time-consuming.
- The many class number searches we conducted in which titles in the same classification area were more general than the topics users sought.

For the most part, searching classification numbers extracted from titles retrieved in low-posted keyword-in-record searches was a hit-or-miss proposition. This strategy could result in
the retrieval of additional, potentially useful titles. It could fail to yield useful retrievals or result in retrievals that were more general than the topics users had in mind. It could also result in requiring patience and perseverance of end users who must browse many retrievals to find promising ones. Hildreth (1992) and Walker and De Vere (1990, 67) experimented with browsing using classified lists of titles and reported comparable results. When such browsing yields many titles, systems could enlist non-subject criteria to enable users to summarize classified lists of titles.

The place of classification-based browsing to overcome the problem of "too few retrievals" would probably be as a search of "last resort." This approach is probably less fruitful than showing users the results of alphabetical searches, that is, list of assigned subject headings in the alphabetical neighborhood of their query.

9. Conclusions

Library classifications hold considerable promise for summarizing the results of high-posted searches in terms of their subject matter. Systems could use broad ranges of classification numbers to consolidate retrievals and captions from the classification schedules to summarize the subjects of consolidated retrievals. Depending on the number of retrievals, captions used to represent broad ranges of classification numbers could be taken from the almost two dozen alphabetical characters (A–Z) of the Library of Congress Classification (LCC), the ten numerical characters (0–9) of the Dewey Decimal Classification (DDC), or from smaller ranges within these larger ranges.

When users selected a listed caption connected with a manageable number of retrievals, systems would respond with a brief-titles display. When they selected high-posted captions, systems could prompt them to display brief title entries by date, format, language, other non-subject criteria, or combinations of criteria. In the case of high-posted captions that resulted from an exact match of a subject heading, systems could sort displays using subdivided forms of the matched heading; if more than thirty-five subdivisions occurred, systems could use the available three or four types of coded subdivisions, i.e., topical, period, geographic, and, possibly, form, to summarize large number of subdivisions.

LCC has a longer way to go than DDC with respect to becoming a summarization tool. LCC's terminology was sometimes not suitable for conveying the subjects of classification numbers which were assigned to many titles bearing the same subject headings. Considerable editorial work would be necessary to improve the wording of captions at all levels of the Library of Congress Classification. Since one classification caption is often connected to most retrievals in less than first-level displays, systems should determine whether the vast majority of retrievals are summarized by a single LCC number before displaying less than first-level displays to users. Other techniques such as language, year of publication, or format should figure prominently into LCC summarization displays because less than first-level summarizations were not always useful.

The problem of too many retrievals also called for improved methods of browsing titles. On occasion, the use of classification to increase retrievals in searches plagued by too few retrievals would have to enlist improved methods of browsing titles. We recommended capabilities similar to the Apple Macintosh's "balloon help" and Mosaic's "show URL's" in which dragging a mouse over truncated information would produce balloon overlays that showed the desired information in its entirety.

Our exploration of the use of classification to enhance the results of searches that produced too few retrievals concluded that searching classification numbers extracted from titles retrieved in low-posted keyword-in-record searches was a hit-or-miss proposition. This strategy could result in the retrieval of additional, potentially useful titles. It could fail to yield useful retrievals
or result in retrievals that were more general than the topics users had in mind. It could also result in requiring patience and perseverance of end users who must browse many retrievals to find promising ones. A next step would be to demonstrate classification-based summarization in an experimental or operational online catalog and test the effectiveness of this approach in quasi-controlled retrieval tests with online catalog users.

Acknowledgments

The OCLC Online Computer Library Center, Inc., provided support for this study through its Library and Information Science Research Grant Program. Celeste M. Burman and Marjorie S. Weller assisted in system development and data analysis tasks.

References


PLCFC. (1978). Subject guide to the collection of the Public Library of Columbus and Franklin County. Columbus, Ohio: The Library.


Research Paper 72.


Eduard R. Sukiasyan
Russian State Library

Change as a Problem of Classification System Development

Extended Abstract

1. Stability and Changeability: Principal Qualitative Characteristics of Classification Systems (CS)

Stability of a classification system is determined by its ability to function for a long time without substantial changes in its structure and content. Stability of CS favours the preservation of its external form, and permits the organization of the process of study of the CS on the level of general principles and stable basic characteristics.

Changeability of a CS is determined by its ability to reflect new elements of content within the framework of a present structure, and in cases, when it is necessary, reflect structural innovations, which make it possible for the CS to retain, in the main, its functional properties, which are necessary and sufficient for normal use. Changeability of a CS ensures the possibility of using the CS in conditions of our changing world and information society.

Stability and changeability are contradictory qualities of a CS, yet they are not mutually exclusive.

2. Limits of CS Stability and Changeability

There exists objective limits of CS stability and changeability. Quantitative and qualitative parameters of changes in structure and content, introduced in CS, can lead to the generation of CS, which will belong to other typological groups. Thus, the process of transformation of an enumerative CS into a combinative one, through the introduction into the structure of auxiliary tables leads to the appearance of a new CS (a thing that happened to the Dewey DC in Europe in the end of the 19th century.)

There exists an opinion that all kinds of changes in the existing CS speak of their imperfection, weakness, ageing, absence of a scientific base, etc. However, we cannot agree with this—in our opinion the development of CS is a normal "physiological" process.

A CS that does not change and develop is doomed. Yet how else can we develop it if not through the introduction of innovations of a structural and substantial character into the existing CS?

3. Internal and External Reasons of Change Introduction

The development of CS is a manageable and, in many things, a planned and comprehended process of incessant character. At its basis lie internal and external factors.

Internal factors are the result of accumulation of intrasystemic contradictions, provoked by the introduction into CS of separate additions, which are frequently not connected with one another. With the passing of time, their quantity begins to approach a critical level and then this or that structural or substantial change of a generalizing character becomes evident. It is in this way that absolutely new classification structures, which absorb previously existing concepts and groups, appear. Moreover, in a number of cases, classes, which formerly had no additions, are
subjected to a radical and substantial transformation, simply because society sees the necessity of this or that development "in general."

External factors are the result of appearance of Change-phenomena outside the boundaries of CS.

We understand that the term "Change-phenomena" covers any new social, political, cultural, technical, etc. phenomenon, scientific discovery, regularity, a subject of animate or inanimate nature and so on, the appearance of which calls for the introduction of changes and additions into the CS.

A very principal and important is the differentiation of Change-phenomena by at least two types. To the first type we shall attribute those Change-phenomena the introduction of which into the limits of CS calls for immediate and urgent actions. An example of such a change is the appearance on the map of the world of a new state. Such changes, firstly, do not call for serious preliminary scientific research and, secondly, they are supported by standard algorithms precisely enough (one knows beforehand what should be done, including even the notation solutions.)

To the second type of Change-phenomena we attribute such phenomena, which do not demand immediate operative actions, and which can be accumulated, analysed and evaluated without hurry so as not to violate the CS structure.

4. Processing of Change-Phenomena

Processing of Change-phenomena is conducted in accordance with a standard technological plan, a detailed description of which would take up a lot of time. We shall limit ourselves to a brief list of the principal operations, the succession of which is designated by digits in brackets.

Search (1) and discovery (2) of Change-phenomena should be the special duty of information workers, who can scan a large number of publications. Only published information can be the source of Change-phenomena. It is desirable to introduce such a system of registration in which the indication of the source is envisaged. The identified Change-phenomena are analysed (3) with the help of dictionaries and manuals, and further in accordance with CS tables and alphabetic-subject indexes to them. If, as a result of this primary analysis, it is established that the given concept is really lacking in the CS, then it is put into (4) the data bank and registered (5).

Further, Change-phenomena are processed by experts, who not only assess the Change-phenomena from the stand of its content (6), but also give their opinion on the expediency (7) of reflecting the Change-phenomena in CS. In the case of positive decisions, approximate economic effectiveness (8) is estimated and one or more drafts of additions and/or corrections are prepared.

This draft is then widely discussed (9). When there are a number of drafts, it is recommended that they be compared in a practical experiment (10). It is only after all these procedures have been realised that a final decision is taken (11). This decision is published and got to the users (12).

Introduction of it into practice (13) should without fail be accompanied by a selective examination of results (14), with the estimation of realistic effectiveness (15). Indexes of estimated or expected effectiveness are compared with indexes of actual effectiveness and then appropriate conclusions are arrived at (16).

5. Prognostication of Change-Phenomena

It is generally known that there exist common regularities of development of science. We can assume that there are also exist regularities of appearance of Change-phenomena. We are familiar with some of them (for example, we do know that the proceedings of a major international
scientific congress will without doubt contain materials of this kind.) Thus it follows that we can and should prognosticate the appearance of a "flow" of Change-phenomena, and this means that we can organize and plan our work correctly.

6. Conclusion

In my opinion, we have good possibilities of cooperation and coordination among universal CS. A few stages of Change-phenomena processing are common for all. We may and must use information coming from international organizations (for example, from ISO). This conclusion allows us to speak of possible coordination and cooperation, and of the expediency of more active information policy.
New Wine in Old Bottles: Problems of Maintaining Classification Schemes

Abstract: Editors of long-standing classifications have to adapt their schemes to totally different circumstances from those they were originally designed to suit. The need to retrieve information accurately in an online environment and a world dominated by the Internet is vital and contrasts with the basic linear approach for which these classifications were intended. The latter need still has to be satisfied, so ways of achieving both goals must be explored. The need for greater synthesis, clearly defined facets, with distinctive notation and a closely adhered to citation order is essential. An expressive notation is attractive in an online environment. Modern educational approaches make traditional structures meaningless and the current economic climate and expense of developing new publication formats is reducing revision budgets. Co-operation between editors and the use of one scheme, e.g., DDC, as a switching language between other schemes, e.g., UDC, thesauri and subject headings lists are one way forward. Existing co-operative projects and future plans between the editors of DDC, UDC and BC2 are discussed and the complementing of classification by a thesaurus is recommended.

1. Introduction

It has often been stated that there is the need for a new classification scheme every thirty years. This pattern actually was roughly achieved through the early part of the present century, but today we are here, in part, to celebrate the 120th birthday of the classification scheme that has a valid claim to be regarded as the standard method for subject retrieval, comparable to the ISBD for author title cataloguing.

But, we are also faced with a totally new situation from that which existed when that classification, and the others that were either derived from it or produced as an alternative to it, were created. We now find that the structures that we have traditionally used for books no longer hold good for a number of reasons which we shall look at in a moment, and the usual policy is to fall back on the familiar—words, normally in English, and even more frequently unsystematically organized.

Attempts have been made in the post-war period to devise new general classification schemes. In the 1960s the CRG in the UK spent several years on such a project (McIlwaine, 1993) which eventually saw the light of day in a different form—a verbal indexing system, PRECIS (Austin, 1984), which now itself has fallen into disuse. In the early 70s Unesco as part of its UNITISIST programme, financed the construction of the Broad System of Ordering (ISO, 1978) as a switching language in an attempt to link various databases that were ordered in different ways. This has been revised, but has never been effectively put into use. The profession prefers to stick to the tried and tested and this means that there is a great onus on the traditional classifications to refurbish themselves to meet these new demands while at the same time causing as little disruption as possible to their existing frameworks.

Those of us who are concerned with editing classification schemes therefore have to try to find ways in which we can cope with the present and the future situation. The history of the Universal Decimal Classification provides an excellent example both of the problems and of the solutions that
have been attempted, with varying success. It is, itself, an attempt to make the original Dewey more flexible and accommodating to complex concepts and for "documentation" rather than the arrangement of library shelves. Dewey's "divide like" approach was carried to its logical conclusion by Otlet and LaFontaine with the introduction of the linking colon. They added a whole range of auxiliary tables—six in addition to the colon facility in the first edition of the classification of 1907 (I.I.B., 1907)—and copied various other devices from Dewey, such as the practice of parallel subdivision which doubtless seemed brilliant at the time, but is now causing a lot of trouble. (See Fig. 1.)

2. Requirements of Present Day Situation

Today, we are faced with a situation where there is far too much information and far too little communication between those responsible for devising automated systems and putting up information on the Internet and those whose professional skill has traditionally been the organization of knowledge. Librarians and information workers make extensive use of online sources of information, but they are not making their expertise sufficiently indispensable to those who are creating the vast range of diverse data available to all and sundry. So, we need act quickly before it is too late. Editors and revisers of classification schemes have a difficult task because we now have to attempt to satisfy two very different types of need—those of our traditional users for (essentially though not exclusively) the shelf arrangement of libraries, i.e., a linear approach to knowledge, and the organization of bibliographical tools, and the needs of the Internet surfer where the consistent ordering of concepts has far less significance. It seems that the traditional schemes have much to commend them for this changing role and that there is little enthusiasm for new systematic approaches.

The use of library classifications for organizing material on the Internet has not been ignored. The British Standards Institution, who holds the copyright for all English language versions of the UDC has an advisory committee that meets regularly and consists of people with an interest in UDC, representatives of professional organizations such as the Library Association and Aslib and recently it has added three other members. These are the people responsible for the BUBL bulletin board with which you may be familiar, for a Social Science network centred on the University of Bristol and a member of a research team in the Computer Science Department at the University of Leeds. All of these are using the UDC as an approach to organizing information online. (See Fig. 2.) There is no reason why more than one system should not be used, for example BUBL is experimenting with using LCSH in addition to the UDC-based subject trees as a means of retrieval.

3. Changes Needed in Traditional Schemes to Meet Today's Needs

As far as the existing general schemes are concerned, there are two problems. One is the need for conformity. The second is the need for revision in certain parts of all of them. Ranganathan identified the need for conformity when he posited his facet formula and identified the familiar five fundamental categories of PMEST. These we have developed into what we now call the "standard citation order". Dewey embodies this in the instructions given in "add" notes and in the general instructions to users on the use of standard subdivisions. It also gives advice on choice between facets in its "Tables of preference." Since this order is also recommended in the international standards for thesaurus construction, we should probably do more to advocate its use universally.

The UDC is as much at fault as any in this regard, since it has always been "all things to all men" and this is why we are now attempting to produce fully faceted revisions with a recommended citation order, together with examples of combination in the tables, and advice to editors to provide
synthesized numbers in the index where they represent compound concepts or terms that are likely to be sought. (See Fig. 3.)

Many regard this, quite rightly, as a loss of autonomy for the individual and a dilution of one of the scheme's most attractive features—its adaptability to a very wide range of circumstances. To those who wish to retain an independent approach, there is no need to follow the recommended order if another seems preferable, but it is at least made clear that in doing so they are not conforming to the recommended practice. Even though a structure is not apparent to the searcher in an online situation, it is essential if the indexing language used is to work effectively. Hence the recent attempts to make the structure of LCSH conform better to a systematic approach (Subject Subdivisions Conference, 1992) and the basis of MESH on a series of tree-structures, to give but two instances.

Notation is a major problem. It is a topic that has seen various changes in fashion. Dewey, and UDC in his wake, thought that a notation that expressed the structure of the classification was essential. Ideal though this may have been in theory, it has broken down constantly and is subject to frequent disruption by new developments in knowledge. In the 1950s Vickery wrote an influential series of articles in the *Journal of Documentation* (Vickery, 1952-1959) on notation which resulted in it becoming fashionable to consider notation simply as an ordering device and to look for no further enhancement of its function. The schemes produced by the CRG at this time provide plenty of examples—the *British Catalogue of Music* (British National Bibliography, 1960) scheme, for instance, with its long strings of capital letters—and this view is still held by Mills and can be seen in practice in the current Bliss revision.

Nowadays, however, the wheel is turning and we are reverting to the view that there is some value in an expressive notation, because in an online environment it enables one to search up and down a hierarchy, provided that the notation is reflecting a true hierarchy. There is the additional attraction of being able to rely upon a consistent, clearly identified, marking of a topic so that one can have confidence that one is locating all the available information on that topic. This is much easier to achieve. The problem with relying on notation to express hierarchies is that, because we are using old frameworks, there are many, many places where what is notated as if it were a hierarchy is in fact no such thing. We are also hamstrung by having a small notational base. The UDC's way round the brevity of Dewey's decimals was to introduce a centesimal notation for long arrays. This is quite wasteful of notation and it has not, in fact, been very widely implemented. Further complications are to be found in the practice of parallel subdivision that shortens notations but conceals the relationship of concepts. This is why we are trying to edit out many such examples in UDC, so as to facilitate retrieval of a concept universally on the same piece of notation. The revision of class 9 demonstrates this very clearly. (See Fig. 4.)

As editors, we need to take account of what we want of our classifications and we have to remember that we have a duty to our traditional users as well as wanting to make our schemes attractive to the modern information world. We do, of course, want to use them as the tool for retrieving recent information immediately, but we also need to retain their traditional use in many contexts. Just because libraries have computerized catalogues they do not necessarily throw out all their nineteenth century works that may be needed by history, literature or even social science students. Therefore the works that were classified 120 years ago by Dewey's scheme still, in some cases, have to be organized today. We should not overlook that totally. Additionally, we have to remember the innate conservatism of our users, and as an editor of a classification scheme I mean the classifiers as the users. They are more resistant to change than anybody else. We publish an annual *Extensions and corrections to the UDC* in which we always ask for comment and contributions for future issues. Invariably, the only comments that I ever receive relate to numbers that have been changed. We do not change numbers in order to irritate our users, we change them in order to produce a sounder arrangement and one that we feel, after due consultation with specialists, will be of greater benefit to
more people. We will never produce a perfect classification, even for one moment in time for one particular place, so we have to be pragmatic and forget some of our idealism, and this is difficult, and unacceptable to some people. The best we can do is to aim for a workable order that is predictable. It is precisely this that a structured notation which permits the searching on one symbol for one concept, or the use of a predictable arrangement for compound concepts, such as that embodied in the standard citation order, aims to achieve.

We are also faced with the problem of a more "woolly" approach to knowledge. The clear division of life into compartments, instilled from one's early days at School, when one learnt Mathematics or Latin or History, and one had a common understanding and expectation of what these disciplines comprised has gone. Young children now have an "integrated day" in which they are encouraged to develop their independent approaches to schoolwork. Universities offer courses in Women's Studies, Environmental Studies, or degree programmes that are made up of individual "pathways" selected by the student, across a very wide range of disciplines. The result of this is the personal loss of a structural base on which to begin the quest for knowledge; the eventual outcome may become the replacement of traditional structures by others that so far have eluded the general classification schemes for obvious reasons. So a systematic approach via well known signposts begins to appear less appropriate.

4. Possible Solutions

We try to cater for this more unpredictable approach by building synthetic devices into our classification schemes. Here, from the outset the UDC further developed the trend already inherent in Dewey's scheme. The increased use of synthesis does make for much greater flexibility and eases the problems of accommodating new combinations. The problem is that in our classifications it has not been developed consistently, and now it cannot be without upsetting the entire structure, and thereby pleasing no-one. So, for example, in the UDC we have a Common auxiliary of materials. This was not in the original classification—it was thought of as a good idea much much later. Consequently, there are many places in the tables of the scheme where the same concept is catalogued with a different notation. Stone, for example, as a material rather than an entity, or an era (as in Stone Age) has the notation .032.5, but as a building material it is 691.2, as a deposit in geology it is 553.5, structures using stone in engineering are 624.012.1, in art as a substance painted upon it is 75.023.12, while in graphic arts it is 76.023.2 and as a stone block for printing it is 681.652.2. We are at present attempting to revise the Special auxiliary subdivisions of class 7. In order to eliminate the use of different numbers in 7 for individual materials, therefore, the following instruction will appear:

_for building materials, at 72, denote kinds of material by .691...

Example(s) of combination:
72.023.691.2 Architectural use of natural stone
_for materials in other contexts, use Table 1k-03 - Common auxiliaries of materials, e.g.,
7.023.1-032.4 Precious metals. Including: gold, silver
7.023.1-032.5 Stone. Including: marble

This means that in future stone as a material in any of the fine arts, e.g., as a surface for painting, as a material for sculpture, as a tool in the graphic arts, and so on will always have the notation from the common auxiliary of materials, and the same is true for all other materials. That is just one example, selected at random and it is far from atypical. Almost all the classes where stone appears as a material have other materials, similarly listed in the common auxiliary and similarly enumerated with a range of different notations.
Superficially, it might seem a reasonably straightforward task to "clean up" these anomalies, but actually it is not, because they are not isolated instances but examples of a problem for a whole range of materials, or whatever, throughout a number of classes. The end result is a total reorganization across many classes. This can only be done slowly and after plenty of preliminary warning.

Therefore, since it seems that, for the foreseeable future, a "make do and mend" approach is favoured by the community that we try to serve, we have to adopt a pragmatic approach to revision. How can this best be done? One possibility is to give serious consideration to one classification scheme as the standard, in much the same way as the ISBD is regarded as the standard for descriptive cataloguing. The obvious candidate for such a role, as I said at the beginning, would be the Dewey Decimal Classification, at least for the English-speaking world, and probably for many other parts of the globe as well. Such usage would permit an agreed minimum for subject description that could then be extended as required through use of another or other systems, for example the UDC. These two schemes make a good starting point for such an experiment because of their common origin. In order to accomplish this, it would be necessary to examine the two schemes and to co-ordinate the broad framework more closely than is at present the case. Clearly, it would work a great deal better with some classes than with others. However, even in a fairly specific and troublesome section of the scheme, such as 621, the broad outline does not differ too impossibly, as Fig. 5 demonstrates, though as one progresses to five digits under 621.3 the variations become greater. There might well be possibilities for mapping a UDC structure on to the Dewey base, in the same way as Ranganathan envisaged the basic classification and depth schedules for the Colon classification.

At present, there are places where DDC is more detailed and more up to date than UDC, though to some extent UDC's superficially out of date appearance in certain classes masks the facility for creating numbers for emerging concepts which can frequently be done through synthesis. To develop an approach whereby the flexibility of UDC could be built on to the structure of Dewey would require close editorial co-operation and agreement in the development or amendment of the scheme, at least to an agreed minimum level. It would also require where the increased flexibility and synthesis of UDC was employed, the need for a clear citation order that was rigidly adhered to along the lines outlined earlier.

The editors of Dewey and the UDC have not closed their minds to such co-operation and to the need for developing the two classifications along more closely integrated lines. One place where we thought that this might be done is in the Area Table and we have great hopes that before too long we may be able to publish a joint Area Table, agreed at least to the level of nation state. With this in mind, when UDC prepares its Area Table for annual publication, we consult with the Decimal Classification Office and keep them informed of what we are proposing to do. Two years ago we revised the Area subdivisions for Australia and for South Africa. The UDC Area Table for Australia was distressingly brief, consisting of only about eight subdivisions. That for Dewey was considerably larger, and there is also in existence a special expansion for Australia that was developed for local use. In expanding our table for Australia we consulted both with Washington and with the Australian member of the Editorial Policy Committee, Giles Martin and we received a great deal of helpful advice. In general outline, apart from one or two instances which seem to be deeply rooted in the historical organization of the continent, such as a main number for Central Australia we kept broadly to the same structure of DDC. (See Fig. 6.)

We based our arrangement on more recent census information that the Dewey table, which has led to a few towns being collected under a different "Including note" from that in DDC, but by and large the overall structure and notational symbols are very similar and there would not be too great a problem in co-ordinating the two tables at a future date, if that were felt to be worth while. This is, with all due respect to any Australians here, only a very small part of the whole, but it does demonstrate that there is no major problem to closer integration between the child and its parent scheme.
We would like to extend this across to the whole Area Table. There seems great good sense in the two classifications sharing a table, and indeed in promoting such a table as a standard for wider use, since there are many instances where a current listing of notations for nation states could be useful. The principal stumbling block at the present time lies in Eastern Europe, where unfortunately the two classifications diverge quite significantly. UDC is used extensively in this part of the world and, of course, to alter an area number does mean upsetting the entire collection in many libraries who use the classification as a means of shelf arrangement. Class 9, which is the most affected by any such change, will occupy a large proportion of a general library's collection. It is also a class that has recently been revised in UDC. We therefore feel that we do have to pay some attention to this problem and are seeking ways of resolving it.

At the present time the publishers of general schemes have a tendency to invest more money into experimenting with new formats in which to publish and in developing tools to assist with the application of the classification in an online environment. This is all very necessary, but the product that is being subjected to these developments must be right also, and this means a constant programme of revision. Both the development of new formats and the production of revised classes are costly. Revision must be at two levels, the overall tidying up that has to go on in order to maintain any scheme, and the “bottom up” revision of the type that DDC has just undertaken for the life sciences and that Professor Williamson will be referring to in relation to Medicine in the UDC. I would like to think for a moment about the latter.

As you may know, the UDC has entered into a mutual agreement with the Bliss Bibliographic Classification whereby, after payment of a royalty and on condition the origin is acknowledged in any published revisions, the structure of the Bliss Bibliographic Classification may be used as the basis for the revision of a class in the UDC. This application is fairly advanced for Medicine, and the Mathematics class is now in the early stages of being similarly employed. In a sense, this is simply legitimizing a state of affairs that has been unacknowledged for years, since all revisers of classifications do and must look at what has already been done and will inevitably either borrow ideas or learn from the mistakes of others. It does seem to me, however, that at a time when funding is short and such little money as there is frequently needed to develop the format of publication, there is scope for much more co-operation between the editors of classification schemes and their editorial boards. It does not make sense for two or three separate sets of people to spend a great deal of time in developing new classes independently. It would be sensible to pool resources, both financial and technical, and build one new, mutually agreed, revised class as a common research project. There would be great benefit in having a core structure that was common across several indexing languages. It could then be translated into a variety of notational symbols (or words) to convey the meaning. If an agreed basic structure could be developed, there would be no reason why further detail could not similarly be grafted onto the basic classification, in the same way as I demonstrated earlier with UDC and DDC. We could also provide the added facility of a thesaurus as UDC is doing in its current revisions, so as to accommodate the users who prefer the purely verbal approach. In time, given the range of languages in which editions of DDC and UDC exist, this could be extended to a multi-lingual facility.

This would mean paying more attention to the index than is frequently the case with the various editions of the UDC. Unlike UDC, the DDC editions are always centrally produced by those responsible for revising the scheme. The UDC Master Reference File (the machine-readable master version at the Hague) does not exist in hard-copy and has no index. Therefore, indexes are entirely at the discretion of a publisher in the individual language, either one of the six members of the Consortium of publishers whose property it jointly is, or some other organization publishing under licence. The Master Reference File at present exists only in English but the hope is that a German version will be created in the very near future. This will provide the potential for the production of a bilingual index, with the added possibility of French also, since a French Medium Edition (CDU, 1990-1995) was
published in 1990-95, complete with index. The conversion of these three indexes into a multi-lingual thesaurus with the link to the classification, and a tie in with the Dewey Decimal Classification and its many language editions, could provide a very valuable retrieval tool and one that would commend itself to the current climate. Searching via natural language is clearly much more user friendly and any measure that can combine this facility with the increased recall that a systematic structure gives is clearly to be commended, whether in an individual library's OPAC or in the deeper morass of the Internet.

5. Conclusion
The present European tendency as exemplified by Germany or the Netherlands, of using a basic classification for the shelf arrangement of libraries and undertaking detailed subject searching by means of an independent thesaurus is less helpful, as it is language dependent whereas a root system based on a classification that could be expanded has far greater universal appeal. Cutter attempted this a hundred years ago. Perhaps we should think about reinventing that wheel too.

References
Fig. 1. Use of "parallel subdivision" to express the same concept in a different context, in DDC and UDC
BUBL WWW Subject Tree - Arranged by Universal Decimal Classification

0 - Reference Works
001 - Research
003 - Writing, Semiotics ...
33 - Economics
330.342 - Development Studies
331 - Labour, Work, Employment, Jobs, Vacancies
339.5 - Conservation of Natural Resources
34 - Law and Legislation
343 - Crime, Criminal Law and Criminology
347 - Copyright, Intellectual Property
35 - Public Administration, Government
355 - Military Art and Science, War, National Defence, Armed Forces
36 - Social Relief, Welfare, Insurance
362.65 - Disability Issues
369.4 - Community and Youth Studies

Under each number are listed appropriate sources:

BUBL Subject Tree: UDC

362.65 - Disability Issues
Disability
  BUBL Gopher Resources on Disability
...
Computers and Disability
  Documents on computers and disability issues
Deaf Education
...
Disabilities Access
  Includes news from the magazine Disability News, and many other resources

Fig. 2: Example of UDC Subject Tree on BUBL
Prehistory. Prehistoric remains, artefacts, antiquities.
Interpretation and synthesis of the material relics of ancient humans,
their culture forms and civilizations

Example(s) of combination:
903.21 Archaeological, prehistoric, protohistoric periods and ages
903-03 Materials of remains and artefacts
For enumeration of materials see Table 1k-03
Example(s) of combination
903-032.42 Gold
903-033.64 Earthenware
903-034.4 Copper

Special auxiliary subdivisions
Three different auxiliary tables may be used at this point, each with
its own distinctive notation. The citation order (reverse of the filing
order) is: (1) type of culture and level of civilization, (2) materials and
techniques, (3) shape and form of remains. These concepts are
introduced by -4 to express the shape and form of remains, .0 to
express the materials from which objects were made and ' (apostrophe) to express the type of culture and level of civilization. All
these may be used together, if required, or each may be used
independently, according to need.

Example(s) of combination:
903.2112.05 Metal tools of hunting peoples
903.2215.02 Wheel-made pottery of nomadic peoples
903.516 Burial remains of advanced farming cultures
903.25-032.42 Gold ornaments
903.2614.08-035.3 Carved wooden objects of worship of pastoral
peoples

Shape and form of remains
903.4 = 62-4
Example(s) of combination:
903-422.11 Round objects
903-462 Tubes
Prefer special subdivisions at .01/.08 where available, e.g. 903.01, not
903-494.2 for flaked objects

903.01 Stone and bone objects according to shaping method. Including: Flaked,
chipped, polished objects
903.02 Pottery according to method of manufacture. Including: Moulded,
ing-for-mmed, wheel-made pottery
903.03 Glass according to method of manufacture

Principal divisions
- 903.2 Artefacts
- 903.21 Tools, Implements
- 903.22 Weapons
- 903.23 Vessels, Jars, Urns, Bowls
- 903.24 Clothing

Fig. 3: Example of instructions on citation order in UDC revision
Examples of combination

<table>
<thead>
<tr>
<th>Class No</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>94(100)</td>
<td>World history (chronological summation of facts)</td>
</tr>
<tr>
<td>94(100)&quot;.../05&quot;</td>
<td>Ancient history in general. History of ancient peoples</td>
</tr>
<tr>
<td>94(100)&quot;05/...&quot;</td>
<td>Mediaeval and modern history in general</td>
</tr>
<tr>
<td>94(100)&quot;1914/1919&quot;</td>
<td>First World War, 1914-19</td>
</tr>
<tr>
<td>If preferred, place under History of Europe at 94(4)&quot;1914/1919&quot;</td>
<td></td>
</tr>
<tr>
<td>94(100)&quot;1939/1945&quot;</td>
<td>Second World War, 1939-45</td>
</tr>
<tr>
<td>If preferred, place under History of Europe at 94(4)&quot;1939/1945&quot;</td>
<td></td>
</tr>
<tr>
<td>94(3)</td>
<td>History of the ancient world</td>
</tr>
<tr>
<td>Alternatively class areas of the ancient world with the history of their modern counterparts</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class No</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>94(416)</td>
<td>History of Northern Ireland</td>
</tr>
<tr>
<td>94(417)</td>
<td>History of the Irish Free State and Republic of Ireland</td>
</tr>
<tr>
<td>94(420)</td>
<td>History of England and Great Britain</td>
</tr>
<tr>
<td>Class here the history of the United Kingdom as a whole. For the history of the countries of the UK individually, see 94(411), 94(415) and 94(429) [as well as 94(420)]. See also 94(41) and the note there.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class No</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>94(420)&quot;.../1066&quot;</td>
<td>Prehistory, Roman and Anglo-Saxon period</td>
</tr>
<tr>
<td>94(420)&quot;1066/1154&quot;</td>
<td>Norman kings, 1066-1154</td>
</tr>
<tr>
<td>94(420)&quot;1154/1399&quot;</td>
<td>House of Plantagenet, 1154-1399</td>
</tr>
<tr>
<td>94(420)&quot;1399/1485&quot;</td>
<td>Houses of Lancaster and York, 1399-1485. Wars of the Roses, 1465-1485</td>
</tr>
<tr>
<td>94(420)&quot;1485/1603&quot;</td>
<td>House of Tudor, 1485-1603. Tudor and Elizabethan periods</td>
</tr>
<tr>
<td>94(420)&quot;1714/1837&quot;</td>
<td>House of Hanover, 1714-1837. Georgian and Regency periods</td>
</tr>
<tr>
<td>94(420)&quot;1837/...&quot;</td>
<td>Victoria and House of Windsor, 1837-</td>
</tr>
<tr>
<td>+ 94(420)&quot;1837/1901&quot;</td>
<td>Victoria, 1837-1901</td>
</tr>
<tr>
<td>+ 94(420)&quot;1901/1910&quot;</td>
<td>Edward VII, 1901-1910</td>
</tr>
<tr>
<td>+ 94(420)&quot;1910/1936&quot;</td>
<td>George V, 1910-1936</td>
</tr>
<tr>
<td>+ 94(420)&quot;1936/&quot;</td>
<td>Edward VIII, 1936</td>
</tr>
<tr>
<td>+ 94(420)&quot;1936/1952&quot;</td>
<td>George VI, 1936-1952</td>
</tr>
<tr>
<td>+ 94(420)&quot;1952/...&quot;</td>
<td>Elizabeth II, 1952</td>
</tr>
</tbody>
</table>

Fig. 4: Example from revision of class 93/94 - History in UDC
<table>
<thead>
<tr>
<th>DC</th>
<th>UDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>621.1 Steam engineering</td>
<td>621.1 Heat engines in general, generation, distribution and use of steam. Steam engines. Boilers</td>
</tr>
<tr>
<td>621.2 Sound and related vibrations</td>
<td>621.2 Hydraulic energy, Water power. Hydraulic machinery</td>
</tr>
<tr>
<td>621.3 Electric, electronic, magnetic, communications, computer engineering; lighting</td>
<td>621.3 Electrical engineering</td>
</tr>
<tr>
<td>621.4 Heat engineering and prime movers</td>
<td>621.4 Heat engines (except steam engines)</td>
</tr>
<tr>
<td>621.5 Pneumatic, vacuum, low-temperature technologies</td>
<td>621.5 Pneumatic energy, machinery and tools. Refrigeration</td>
</tr>
<tr>
<td>621.6 Fans, blowers, pumps</td>
<td>621.6 Fluids handling, storage and distribution plant and techniques</td>
</tr>
<tr>
<td>621.7 Mechanical technology in general: processes, tools, machines, equipment</td>
<td></td>
</tr>
<tr>
<td>621.8 Machine engineering</td>
<td>621.8 Machine elements, Motive power engineering, Materials handling, Fixings, Lubrication</td>
</tr>
<tr>
<td>621.9 Tools and fabricating equipment</td>
<td>621.9 Working or machining with chip formation, Abrasive working, Hammers and pressers; including Cutting, grinding, Thread-forming</td>
</tr>
<tr>
<td>621.3.0 sp auxiliaries...</td>
<td></td>
</tr>
<tr>
<td>621.31 Generation, modification, storage, transmission of electric power</td>
<td>621.31 Generation, supply and control of electricity. Electric machines and apparatus. Electrical measurement. Applied magnetism. Applied electrostatics</td>
</tr>
<tr>
<td>621.32 Lighting</td>
<td>621.32 Electric light sources</td>
</tr>
<tr>
<td>621.33 Electric power transmission for railroads</td>
<td>621.33 Electric traction, locomotion, propulsion</td>
</tr>
</tbody>
</table>
621.34 Magnetic engineering

621.36 Applied optics and paraphotic engineering

621.37 Testing and measurement of electrical quantities

621.38 Electronics and communications engineering

621.39 Computers

621.35 Electrochemical technology

621.36 Thermoelectricity. Electric heating

621.37 Technique of electric waves, electromagnetic waves, oscillations, pulses

621.38 Electronic devices. Electron tubes. photocells. Particle accelerators. X-ray tubes


Fig 5: Comparison of 621 in DDC and DDC
Australia. Commonwealth of Australia

Western Australia

Perth
Including: Armadale, Belmont, Canning, Cockburn, Fremantle, Gosnells, Kwinana, Melville, Nedville, Rockingham, Stirling, Subiaco, Wanneroo

South West, Great Southern, Midlands and Central
Including: Bunbury, Bridgetown, Busselton, Collie, Mandurah, Manjimup

For Perth Metropolitan District, see (941.1)

Lower Great Southern
Including: Albany, Broomhill, Denmark, Gnowangerup, Katanning, King, Kojonup, Pallinup

Upper Great Southern
Including: Boddington, Coobaling, Narrogin, Wickepin, Williams

Midlands
Including: Merredin, Northam, York

Central
Including: Carnarvon, Geraldton, Wiliuna

Pilbara
Including: Dampier, Karratha, Mount Magnat, Newman, Pannawonica, Paraburdoo, Port Hedland, Roebourne, Tom Price, Wickham; Barrow Island, Bernier and Dorre Islands, Dick Hartog Island

Kimberley
Including: Broome, Derby, Halls Creek, Kununurra, Wyndham

South Eastern
Including: Boulder, Coolgardie, Esperance, Kalgoorlie, Kambalda, Norseman

South Australia and the Northern Territory

South Australia

Adelaide
Including: Elizabeth, Gawler, Noarlunga, Salisbury

Yorke and Lower North
Including: Clare, Kadina, Maitland, Moonta, Wallaroo

Eyre
Including: Ceduna, Cleve, Port Lincoln, Tumby Bay

Northern
Including: Quorn, Leigh Creek, Peterborough, Port Augusta, Port Pirie, Whyalla

Northern Territory

Southern district
Including: Alice Springs

Class here Central Australia (a term still in general use but of specific administrative meaning only between 1926 and 1931; today the term "Central Australia" usually refers to Alice Springs and an area of about 700 km radius around it - generally a major tourist destination)

Northern district
Including: Arnhem Land, Jabiru, Katherine, Litchfield, Tennant Creek; Groote Eylandt, Melville Island, Bathurst Island

Darwin
Including: Palmerston

Queensland

Fig. 6: Section from Area Table for Australia

(a) In UDC
Australia

Class here *Great Dividing Range

SUMMARY

-941 Western Australia
-942 Central Australia
-943 Queensland
-944 New South Wales
-945 Victoria
-946 Tasmania
-947 Australian Capital Territory
-948 Outlying Islands

Western Australia

-941 1 Perth metropolitan district
   Including Fremantle

-941 2 Southwestern district
   Including Albany, Bunbury, Collie, Geraldton, Katanning, Manjimup,
   Merredin, Narrogin, Northam; Kalbarri, Nelson and Hays, Noramlup,
   Stirling Range National Parks; Darling, Stirling Ranges; Blackwood,
   Greenough, Swan Rivers
   For Perth metropolitan district, see —9411

-941 3 Northwestern district
   Including Carnarvon, Port Hedland; Barrow Island, Bernier and Dorre
   Islands, Cape Range National Parks; North West Cape; Dampier
   Archipelago; Dirk Hartog Island; Hamersley Range; Ashburton,
   Gascoyne; *Murchison Rivers, Lake Austin

-941 4 Kimberley district
   Including Broome, Derby, Wyndham; Bonaparte Archipelago; King
   Leopold Ranges; Fitzroy, Ord Rivers

-941 5 North central district
   Including Lake Nabarlek
   Class here Gibson, Great Sandy, *Great Victoria Deserts
   Class Great Victoria Desert in South Australia in —94238

-941 6 South central district
   Including Coolgardie, Kalgoorlie, Wiluna; Lakes Barlee, Carey,
   Carnegie

-941 7 Southern district
   Including Esperance, Norseman; Cape Le Grand, Esperance National
   Parks; Archipelago of the Recherche

Central Australia

-942 3 South Australia

-942 31 Adelaide metropolitan district

-942 32 Central district
   Including Angaston, Clare; Gawler, Mannum, Murray Bridge, Port
   Noarlunga, Port Pirie, Salisbury, Strathalbyn, Victor Harbour;
   Chawcoys Line Reserve National Park; Mount Lofty Ranges
   For Adelaide metropolitan district, see —94231

-942 33 Eastern district
   Including Barmera, Berri, Loxton, Renmark, Tailem Bend; Billilatt,
   Mount Rescue Conservation Parks; Lake Alexandrina
   Class here The Coorong

*Class parts of this physiographic region or feature as instructed under —4-9

Fig. 6: Section from Area Table for Australia

(b) in DDC
Thesaurus-aided Searching in Search and Retrieval Protocols

Abstract: Open system protocols for search and retrieval have not provided explicit ways in which to implement thesaurus-aided searching. A number of different approaches within the existing protocols, as well as a proposed service, are evaluated. A general approach to implementing thesaurus-aided searching, particularly during consultation of a thesaurus, requires an entirely new service, whose main features are described.

1. Introduction

Thesaurus-aided searching is the explicit use of a thesaurus by a searcher to enhance information retrieval in an automated system. Thesaurus-aided searching has been used in many information retrieval (IR) systems to enhance retrieval, usually for subject access, but sometimes for other hierarchically-structured authority files such as those for personal or corporate names. However problems have arisen when trying to generalize this function for information retrieval standards such as the ANSI Z39.50 Search and Retrieval Protocol (ANSI/NISO, 1995) or the associated international Search and Retrieval Protocol ISO 10162 and 10163 (ISO, 1992). The ANSI Z39.50 and the ISO 10162 and 10163 standards (hereafter referred to simply as the SR protocols) describe in detail the interaction between client and server systems (referred to in the standards as origin and target systems) in providing distributed search services. Though designed to be "comprehensive information search and retrieval protocol[s]" (St. Pierre, 1996), these standards contain as yet no explicit thesaurus-aided search capability, largely because of the difficulty in agreeing on the way in which thesauri can be used to enhance the search process. This situation is increasingly restrictive in the open systems world of internetworking, and has discouraged developers of SR origin or target software from considering implementation of full thesaurus-aided searching.

This paper briefly describes previous efforts to implement thesaurus-aided searching using current SR protocols, and a recently proposed service based upon the model of the thesaurus as a database separate from the principal searchable database. In order to provide a generalized, simple and comprehensive approach to thesaurus-aided searching, however, another new service is also required. The main features of this new service are described, along with several scenarios describing how it could be used to support thesaurus-aided searching.

Broadly speaking, there are two distinct phases of user-thesaurus interaction during thesaurus-aided searching: a consultation phase, in which the searcher (i.e., the end user) explores the thesaurus, collecting information which can be used in formulating a search strategy; and a query phase, in which the searcher submits a search strategy with explicit instructions to enhance that search using the structure of a thesaurus. (The enhancement of a search without the knowledge of the user is excluded from this discussion, since it depends only the target software, and has a minimal effect on a search and retrieval protocol.) Each of these two phases is examined separately.
2. Thesaurus Consultation

Thesaurus consultation is the phase of thesaurus-aided searching in which the user consults the thesaurus in order to gather information that will help to develop the strategy actually submitted during the query phase. Currently there are a number of possible approaches to implementing thesaurus consultation within the SR protocols.

The ability to learn about possible search keys in a database is provided in the SR protocols by the Scan Service of the Browse Facility, an addition to Version 3 of the protocols. This service was modeled on a feature supported by many IR systems which provides the user with a list of possible search keys, usually extracted from the inverted or index file which provides access to the records stored within the database. In formulating a Scan request, the origin system supplies a number of data elements, including the field with which the index is associated, the search key to be used as the entry point into the index, the maximum number of index keys to be returned, and the relative position within the list at which the search key is or would be found. In response to this request, the target system returns an ordered set of search keys, embedded in a very simple data structure. The ordering of the search keys, and the ability of the origin to provide a relative position of the search key, allow the origin to request lists of keys preceding or following the current list, in effect providing the means for the searcher to "page up" or "page down" through the entire list of possible search keys in the index.

This simple Scan service was clearly intended for sorted key lists that have no particular structure, such as title keywords, and in this role it serves very well. However two different techniques have been used as a means of passing structured thesaurus information using this service. The first technique is to make use of an undefined data element called Otherlnformation found in the simple data structure used to return keys in the target's Scan response. This data element was designed to allow targets to include, with each key, certain specific pieces of information that apply to that key, and can be used to graft another data structure onto the returned index key. Using this data structure, target systems can pass back to origins a complete thesaurus record structure including scope notes, history notes and other language versions of descriptors as well as a full range of thesaurus relationships including Broader, Narrower, and Related Term relations.

There are two objections to this approach. First, both target and origin must agree on the structure that will be passed back to the origin with the search key: while the developer that provides both origin and target software to customers can arbitrarily determine such a structure, this works against the notion of interoperability which is at the heart of the SR protocols. Second, any definition of this Otherlnformation data element requires a fixed, pre-determined structure, and a structure sufficiently general to accommodate a wide range of thesauri could very quickly become large, complex and difficult to modify for unanticipated future requirements.

A second approach to expanding the current Scan service to support thesaurus consultation has been to use an SR construct called attribute sets to request the target to return information about thesaurus associations. Attribute sets are a Z39.50 construct that allow the specification of a number of characteristics to be associated with a search term: for example, a Use attribute such as "Date of publication" indicates that the origin hopes to find a specific search key used in a date of publication field, and a Relation attribute, such as "less than", indicates that the origin wishes to search for values less than the corresponding search term. A basic attribute set called Bib-1 defines common characteristics bibliographic data and has been responsible for much of the success of the standard in ensuring interoperability between different Z39.50 origins and targets. Another attribute set called the Scientific and Technical Attribute Set (STAS-1) was developed to include data elements appropriate to scientific and technical information; this attribute set includes in its Relationship attributes the basic thesaurus relationships, as well as values not
normally considered relationships, e.g., Scope Notes, and combinations of various relationships, e.g., All (STAS Maintenance Agency, 1995). Though designed principally for searching, by attaching attributes for these kinds of relationships to a Scan Request the origin could in effect request that the target return certain associated values for each search key, i.e., keys and text strings that had a specific kind of relationship to the main index key.

The use of attribute sets in this context seems somewhat non-intuitive, especially for identifying thesaurus data elements that do not apply directly to searching at all, such as definitions, scope notes and history notes. The size and complexity of attribute sets has already been noted as a problem in SR protocol development (Lynch, 1996): including in each attribute set data elements necessary for the display of information from any associated controlled vocabulary will only increase their size and complexity. Finally, like the use of OtherInformation, the use of attribute sets only allows a single sequence of controlled terms, whereas some thesauri could profitably be consulted in both an alphabetic and systematic sequence.

While both of these approaches allow enhancement of the Scan Service to include some information about a thesaurus, neither of them is alone sufficient for full thesaurus consultation. In particular neither provides a model for viewing only certain selected thesaurus terms. A common requirement during thesaurus consultation involves a request to view only a particular subset of thesaurus terms, such as all descriptors that contain a particular keyword; the resulting terms, and associated data elements are returned to the user but not necessarily presented in any particular order. This behavior is difficult to model using the Scan service which is implicitly based on a sequential display of all search keys from an index file. As a result of some of these concerns, a proposal has recently been advanced within the Z39.50 Implementors' Group to develop a new service within the Browse Facility, separate from the Scan service.

This proposed new service, called Structured Vocabulary Browse or simply SVB (Kunze, 1995), recognizes that the consultation of a structured vocabulary such as a thesaurus is more complex than a simple "scan" of index keys. The underlying model for this new service is not that of a list of index keys, but rather a database search, where the database being searched is not a bibliographic database, but a thesaurus or other structured vocabulary: the origin submits a specific keyword, and receives back a number of records that match the request. This proposed new service has several advantages. First, it allows for only a subset of thesaurus terms being returned in response to a consultation, by allowing developers to make use of some of the same software and functions already developed to support bibliographic searching in which results are usually a subset of records provided without an explicit sequence. Second, the proposal suggests returning information in a flexible structure that depends upon the thesaurus in use by the target, in the same or similar record syntaxes already cited within the SR protocols for the return of bibliographic information. This use of these existing record syntaxes (such as MARC authority formats, or flexible, general syntaxes such as the General Record Syntax GRS-1), can accommodate a variety of different types of thesaurus structures and hence promotes interoperability.

However, this new service is deliberately limited in that it does not attempt to replace Scan in terms of providing access to a full thesaurus in an explicit sorted order; nor does it address the limitations of both of the previously mentioned approaches towards using Scan for thesaurus information. It also introduces an evident duplication, since in one service (Scan) thesaurus information is presented in a fixed, pre-determined, private format, or in the form of an attribute set, whereas in another service (SVB), the same information associated with each term is presented in a record syntax. Origin software wishing to support thesaurus-aided searching across a variety of target systems and databases must now be able to process essentially the same data in three different ways: once as extended information to the Scan response (using perhaps several different database-specific structures), once in the form of an attribute set, and once in the format
of a database record in one of the recognized SR record syntaxes, such as GRS-1 or a MARC authority format.

The Structured Vocabulary Browse service represents a move towards a more general and flexible support for thesaurus consultation by modeling the thesaurus as a database, and returning records in a record syntax. These features need to be extended to the sequential viewing of thesaurus contents. The problems evident in extending the Scan service to thesaurus information suggests the need for a new SR service that (like the Structured Vocabulary Browse service) is implicitly based on the notion that consulting a thesaurus is like consulting a database. This notion should be familiar to many IR system developers because in many systems the controlled vocabulary is stored physically as a database, in a file separate from the index file containing keys used for searching the main bibliographic database. A complete thesaurus consultation process would make use of the existing SR Explain service to discover information about the thesaurus, the proposed Structured Vocabulary Browse service to identify specific unsorted subsets of thesaurus terms, and a new proposed service, which we will call Structured Vocabulary Scan, to provide access to sorted segments of thesaurus terms. With the use of these three services, a complete and generalized interface for thesaurus-aided consultation encompassing all navigation of thesauri by the end user of the thesaurus could be implemented in the SR protocols.

First, an origin can discover the name of a structured vocabulary database by requesting information about the field in the main bibliographic database using the SR Explain service. The Explain information on a given field could easily be extended to include the name of the database that contains the structured vocabulary associated with that field. The origin can then submit a second Explain request against the structured vocabulary database (i.e., the name of the thesaurus database), to obtain information about the structure of the particular thesaurus, such as the number and nature of different thesaurus relationships (BT, NT, RT), of different language versions of a term (English, French, Spanish), and of different purely informational fields (scope notes, history notes). The only additional piece of information required by this thesaurus use of the Explain service is an indicator of which fields represent relationships (e.g., BT, NT, RT) which can be used either in searching the bibliographic database or in navigating the thesaurus, and which fields represent purely information data elements, such as scope notes.

Second, an origin can request a search of the thesaurus database, using (for example) a keyword taken from a descriptor. This request would use the proposed Structured Vocabulary Browse service described above, in which the origin would request a search of the thesaurus database using some search criteria (e.g., "all terms with Canada in them"); in response, the target would usually reply with a number of thesaurus records, returned in a record syntax. The origin could then display those records to the searcher in a format which would allow the searcher to decide which term to use in another search (thus ending the thesaurus consultation) or would allow the searcher to decide what record or records provided an interesting point for further consultation.

Third, an origin can use the new proposed service, the Structured Vocabulary Scan service, to browse the thesaurus. In this new Service, the origin would send a Structured Vocabulary Scan request to the target, and the target would return a Structured Vocabulary Scan response. The request would include a number of specific data elements including the origin's preferences for the sequence of the thesaurus records returned, a starting term value, and a preferred record syntax. The response would include many of the same elements, indicating what the target was able to supply, usually with a sorted set of thesaurus records returned in response to that request. A brief description of the key data elements in the request and response follows.
Database name is the name of the structured vocabulary (thesaurus) database from which the records would be drawn. The origin would determine the database name through the Explain service.

Term is the thesaurus term which provides the entry point into browsing the thesaurus. The origin supplies this value; no response is required from the target.

Sequence is the order in which thesaurus records are returned. The origin requests a preferred sequence, and the target responds with the sequence which has been supplied. Typical sequence values would be alphabetical (records in alphabetical order by term) or systematic (records in order according to an explicit or implicit classification scheme). If the origin requested a specific sequence and the target could not supply the requested sequence, the target would specify in the response the sequence that was being supplied.

Number of entries is the number of thesaurus records which the origin expects to receive back in the response to the request (in a request) or the number of records actually returned by the target (in a response).

Position in response is the position of the record corresponding to the requested term in the sequence of returned records. The origin indicates a preferred position, and the target indicates the actual position in the response.

Levels up and levels down indicate a number of hierarchical levels above or below to include in the response. The origin indicates the preferred number of levels to display; the target indicates the number of levels actually present in the records returned. This value allows the supply of hierarchical record displays, in which BT and NT terms are provided for a certain number of levels.

Element set name provides a way for the origin to indicate the desired fullness of records in the response. Element set names such as B for Brief record format or F for full format should be supported by all targets. If the origin intended to display only a limited amount of information about each term, such as the term itself, as would be found in a systematic listing in which no explicit BT, NT or RT relations or notes fields are supplied, a Brief record format would be preferred. If the origin intended to display more extensive information about the term, e.g., complete relationship information such as is found in many alphabetical displays, then a Full record format could be requested.

Record syntax is the syntax in which origin would prefer to have the records returned, or the record syntax in which the target actually returns the records. Record syntaxes would include Simple Unstructured Text records (SUTRS), General Record Syntax (GRS-1) records or MARC authority format records. A response in GRS-1 or a MARC authority format, together with the information about data elements in the thesaurus obtained through the Explain service, would allow the origin that made use of a Graphical User Interface (GUI) to format records, so that a user could click on a BT relation, automatically submit a request for another browse based on this particular term, and then see the selected term display. This would allow searchers to navigate simply and easily through multiple thesaurus displays searching for terms to add to their search strategies.

Finally, response records would the set of records returned by the target as part of the response.

Like the current Scan service, the new SVS Service allows the consultation of an ordered list of items. Unlike the current Scan service, the origin can request one of several different sequences, the thesaurus structure is determined by the target system based on the characteristics of the particular thesaurus used, and the results are returned in a record syntax both can express a wide range of thesaurus structures, but in such a way as to allow origins to interpret record contents and present them to the searcher in a useful way. Two or three simple scenarios,
describing the interaction between origin and target, will indicate how this new service could be
used to support consultation of a thesaurus.

Example 1: The user wants to consult an alphabetic thesaurus such as the OECD Macrothesaurus, displaying records starting with the term "DEVELOPMENT AID". The client submits a request indicating the target's name for the Macrothesaurus database, term "DEVELOPMENT AID", sequence alphabetic, number of entries 5, position in response 2, levels up and levels down 1, F as the element set, and record syntax USMARC authority format. The target returns five records, starting with the record for "DEFENCE POLICY", (the term that immediately precedes "DEVELOPMENT AID" in the thesaurus) and continuing for the four records that follow "DEFENCE POLICY" in alphabetic order in the thesaurus. The records are provided in GRS-1 syntax. The origin is able to extract BT and NT fields from this record syntax, identify them as selectable, and to format them on the screen so that the searcher can then click on the BT term "INTERNATIONAL COOPERATION"; the origin then submits a new request as before, but with search term "INTERNATIONAL COOPERATION", and receives back five more records in alphabetic sequence with the record for the descriptor "INTERNATIONAL COOPERATION" in the second position.

Of these five records, the first corresponds to the descriptor "INTERNATIONAL AGREEMENT" and the fifth corresponds to "INTERNATIONAL RELATIONS". The searcher issues a Page Up command, and the origin then requests term "INTERNATIONAL AGREEMENT", sequence alphabetical, number of entries 5, position in response 5. The server responds with a new sequence of 5 records, of which the 5th record corresponds to the descriptor "INTERNATIONAL AGREEMENT". The effect is to "page up" within the alphabetic sequence of the thesaurus.

Example 2: The searcher wants to consult a classified thesaurus, like the Art and Architecture Thesaurus, starting with the term "paint". The origin sends a request with database name for the AAT on the target system, term "paint", sequence systematic, number of entries 20, position in response 3, levels up/levels down 0, element set B (for brief), record syntax GRS-1. The target responds with 10 records, of which the first record corresponds to the descriptor "Maroger medium" (i.e., the term two terms before "paint" in the classified sequence of the thesaurus) passes through "megilp" and "paint", and finishes at "enamel paint", the descriptor 17 positions after "paint". Using information about the relative hierarchical level of each term returned within the record structure, the origin could, for example, format the descriptors one to a line in a 20-line window, with variable indentation to reflect their level within the thesaurus hierarchy.

Example 3: The searcher wants to consult an alphabetic thesaurus like the OECD Macrothesaurus using a hierarchical display. The origin sends a request with the database name known by the target for this thesaurus, the term "INDUSTRY", sequence alphabetical, number of entries 1, position in response 1, levels up 2, levels down 3, element set name F for Full, record syntax GRS-1. The target responds with a single record, pre-formatted by the target, displaying two levels of BT and 3 levels of NT embedded in the record. The origin could then show the searcher the one returned record in this "hierarchical" format of; however because of the simple unstructured format of the returned record, the searcher would likely have to type in another term appearing on the display to navigate to another part of the thesaurus.

3. The Query Phase

While thesaurus consultation requires some major changes in the SR protocols, the second phase of thesaurus-aided searching, the query phase, in which the user actually submits a search
against the source database, does not present the same modeling complexities. Once there is a
mechanism in place for the origin to discover information about the data structure of the thesaurus,
e.g., the number and nature of the thesaurus relationships between one term and another, then the
origin can use these thesaurus relationships to explicitly request an enhancement of the search.
The search request would have to be extended to allow, not only the specification of the descriptor
to be searched, but also one or more expansion requests, where each expansion request consists
of the specification of a thesaurus relationship (e.g., BT, NT, RT) and the number of degrees of
relationship that would apply. For example, to request an expansion of the term "paint" to the
descriptors below "paint" in the thesaurus, the search would consist of the term "paint", and an
expansion specification of NT and level "1". To implement an "explode" operator, the origin
would submit a search request with a term, and an expansion request of NT with the degrees of
relationship either an arbitrary or an unnaturally high value such as 9999. With some modifications
in terms of the means of specifying thesaurus relationships and in the additional specification of
a number of degrees of relationship, the current Search Service could provide all the functionality
required to use a thesaurus in searching a bibliographic database.

4. Summary

Generalized support for thesaurus-aided searching requires a mechanism for discovering
information about a thesaurus, a means of consulting that thesaurus through a sequential display
of records as well as a selected subset of records from that thesaurus; and a mechanism for using
thesaurus relationships and differing degrees of relationship in the search process. Existing or
proposed services in the SR protocols provide, with some relatively minor extensions or
modification, three of those four requirements. However, the sequential display of thesaurus
records demands a new SR service, which would present a number of advantages over the way
in which thesaurus-aided searching can be currently implemented in SR protocols. By providing
thesaurus information to be returned in a full record syntax, it would allow for a richer set of data
elements to be returned, the use of already existing data standards (such as authority record
formats) where appropriate, easier modification of the format of those data elements, and greater
adaptability of the formats to accommodate a broad range of thesaurus and controlled vocabulary
structures. The author is currently compiling a survey of thesaurus-aided searching supported in
currently available information retrieval and library search software. The results of this work will
be used to verify the generality and appropriateness of the proposed approach in meeting the needs
of searchers using standard information retrieval protocols.

References
http://lcweb.loc.gov/z3950/agency (Includes the definition of the Bib-1 attribute set).
UFL.EDU. Sept. 22, 1995. Available from LISTSERV@NERVM.NERDC.UFL.EDU.
Lynch, Clifford. (1996). Defining and maintaining attribute sets for use with the Z39.50 protocol:
a discussion paper. Z3950IW@NERVM.NERDC.UFL.EDU. Feb 2, 1996. Available from
LISTSERV@NERVM.NERDC.UFL.EDU.
STAS Maintenance Agency (1995). STAS-1, scientific and technical attribute set. URL:
http://stas.cnidr.org/STAS.html.
Abstract: Once the library has been through a process of transition from traditional library procedures to automated ones, natural language searching became a necessity for both indexers and searchers. Therefore, aside from the precoordinated classified catalogue we started to build a dictionary of terms in order to make postcoordinated search possible in keeping with the UDC notations assigned to each bibliographic record. After a while we came to the conclusion that the dictionary needed a control of its terms so that synonymous concepts and semantic ambiguities be avoided. The project presented in this paper shows how reality imposed the improvement of the quality of indexing and hence of the searching possibilities. It also shows the reasons why we consider a multilingual thesaurus based on UDC an ideal indexing and searching device. The experiment applied on Class 8 of UDC illustrates the way the UDC tables can be quite successfully used in building a thesaurus due to their qualities and how their limitations can be overcome by a thesaurus. An appendix to the paper contains a sample of the multilingual thesaurus given in both alphabetical and systematic layouts.

1. Introduction

As an institution expected to meet the information needs of a particular category of people (that is to say students, university teaching staff and researchers) the Central University Library of Bucharest - Romania is increasingly confronted with the necessity to find the most suitable solutions to provide search tools able to cope with these needs. Indexing is providing an intermediary between the user and the document on which the information retrieval is largely dependent. It means all the same extracting the main subject matter from each document and representing it in a way that makes retrieval possible. Any piece of information that has not been included in the indexing system will be not identified when searching. Hence the importance of the way the indexing is made upon the quality and effectiveness of the index, and ultimately, upon the reliability of the database as a whole.

2. The Idea of a Word System for Subject Indexing

The preoccupations for an as accurate as possible indexing language was dictated by the growing rate of the library collection reflected in the database. Besides, the number of requests for a certain subject was greater than for works whose authors and titles were known. The existence of an OPAC offered the possibility of using uncontrolled keywords for searching, aside from UDC codes. A word system for information retrieval proved soon to be what our readers expected. We witness nowadays a general tendency to replace subject access with keywords. Because what relevance may have an answer like: "821.112.2375.5" given to an inquirer interested in the German pronunciation? This piece of information is not able to fit into the structure of knowledge in his mind. Words belonging to natural language are more suitable to the user's demands in terms of precision. The same is true when speaking about currentness; a keyword for a new concept can be immediately used in indexing as access point (e.g., 'airbrush painting'). Still then, the vocabulary used for searching is necessary to be known by the indexer and the searcher as well. Therefore, we are already speaking about an agreement between the author of the book as a source of a message
or information and the inquiry matching it so that it can be decoded by the searcher with the least noise possible. We deal then with a vocabulary of keywords which is very flexible on condition that the searcher has the adequate term in mind (e.g., 'entomology' used for the study of insects). But, among other things, free indexing has the disadvantage of little control on the synonyms and homographs, the consistency of terms being also not its very strong point.

So, we came to the conclusion that the best solution was to build a controlled vocabulary that should be used in searching in order to avoid the vagueness and inaccuracies of the natural language which result particularly from homonymy, synonymy and syntactic ambiguities (Fugmann, 1993, 67-107). Our decision was that to each precoordinated UDC notation should be attached one or more descriptors so that postcoordinated search was possible (Riesthuis and Bliedung, 1991, 109). By means of a controlled vocabulary the scattering of synonymous terms denoting the same concept (e.g., 'popular ballad' and 'folk ballad') in an alphabetic list is avoided. The relations, both hierarchical and associative, are meant to make retrieval more effective on condition that the thesaurus is well structured and the descriptors are adequately assigned.

3. Why a Multilingual Thesaurus?

Having a word system to search with was just a stage in the evolution of the search devices our library provides. It is obvious why this system was necessary to be in the mother-tongue of the library users. But, for some years now, it has happened more than once that not only Romanian people come to use our library for their information needs. Then the question arises, how should we satisfy the demands of this category of users? So far, each bibliographic record was assigned one (or more) UDC notation(s) and several corresponding subject headings in Romanian as access points to the subject. Given these circumstances the problem still remains since never-the-less the Universal Decimal Classification "speaks with every tongue" it does not say too much to most of the readers.

Moreover the library has been recently connected to Internet, so our database, being included in an international network can be reached from different places outside our country and having only Romanian subject headings for search is not enough. That is how the idea of a multilingual thesaurus was born in order to meet the present requirements of the library users. We consider English and French as most suitable languages to this purpose, the VUBIS system our library uses having a provision for them in the subject headings management.

4. Why Still UDC?

As mentioned before the main strong points of UDC are that it is international and not language dependent. The international retrieval and exchange of information is made easier by a systematic approach based on symbols than by an alphabetical one based on one language. Information retrieval systems based only on words do not enable the user to see the structure of a subject and search up and down hierarchies or related topics that a classification scheme permits (McIlwaine and Buxton, 1993, 103). Consistency and control of notations are two other strong points of UDC that become more difficult to maintain in systems based on words alone. The hierarchical nature of most of the notations makes it easy for indexers to create verbal indexes to classified catalogues to be used in searching in an OPAC. As a matter of facts this is the feature on which most of the project undertaken here is based. Synonymous terms which are often given in the descriptions of UDC notations in the schedules prove to be
a valuable source of terminology for those interested in building thesauri. None-the-less, cross references are ready to be used for concepts having related meanings though in different domains.

According to McLlwaine and Buxton (1993, 110), the hierarchical structure of UDC gives a solution to multiple searches for broader and narrower terms in online catalogues. The truncation sign placed after a notation gives a comprehensive search for narrower terms belonging to a more general concept without any effort to find out each of these narrower terms separately. If a searcher is interested by the grammar of the Italian language in general he is free to choose between:

811.131.1'36

which will give him all items covering different aspects of the grammar of the Italian language and:

811.131.1'36?

which will give several other numbers such as:

811.131.1'362 Comparative grammar
811.131.1'364 Generative grammar
811.131.1'366 Morphology
811.131.1'367 Syntax
811.131.1'367.2 Sentences generally
811.131.1'367.3 Types of sentences

according to his interests.

The independence of language is an advantage in online systems because irrespective of the country a certain searcher belongs to he will find a UDC notation in the UDC edition in his language and the same number will be found by another one provided that he has a UDC version in his own language. The access to the UDC database being given both searchers have equal opportunities ending in equally good results, irrespective of the language they used in searching.

5. Building the Thesaurus

As Ray Prythearch writes in his Glossary (Harrod's, 1995, 635), a thesaurus may be defined either in terms of its function or of its structure. In terms of function it is a terminological control device used in translating from the natural language of documents into a more constrained system language. In terms of structure, a thesaurus is a controlled and dynamic vocabulary of systematically and generically related terms covering a specific domain of knowledge.

As mentioned before, the first stage of the subject indexing activities in our institution was the natural language or free indexing stage when each record was assigned keywords starting from the text of the UDC notations. This had advantages in more precision in indexing and better recall in retrieval. But disadvantages were soon to occur while the dictionary of terms was growing in dimensions. Most of the inconsistencies were due to various forms of terms assigned to define the same concept given either the singular-plural differences or synonyms or words whose meanings were somewhat overlapping with each other.

E.g., Ballads / Ballad / Popular ballads / Popular ballad
Rhetoric / Art of speaking
Public opinion / Social opinion
Etymology / Origins of the language
This way the information was scattered all over the alphabetical index without any possibility to control the use of terms. So we decided that the control of terms is obviously needed the way to reach it being provided by the use of the UDC tables themselves. That was the most valuable source for both the selection of terms and for their relationships, either hierarchical or associative. There were still some other problems related to the meanings of the words, some of them being very difficult to disseminate. A group of indexers was then established, each of the members being a subject specialist and we tried, and still try, by sharing responsibilities, to overcome the problems we had to face. Thus, we started 'cleaning' the dictionary of keywords having in mind the advantages offered by controlled vocabularies (Lancaster, 1986, 161):

- reducing the semantic ambiguities by distinguishing among various meanings of homographs;
- promoting consistency in the representation of the subject matter through control of synonyms;
- facilitating the comprehensive searches by some superimposed structure that links semantically related terms.

6. Problems and Solutions with Multilingual Thesaurisation

The UNISIST Guidelines for the Establishment and Development of Multilingual Thesauri (1981) recognize three possible approaches to the construction of multilingual thesauri:

1. Begin from scratch and collect terms from all the languages involved, seeking equivalents after the terms are collected;
2. Take an existing thesaurus in one language and translate it into the other languages;
3. Take two or more existing thesauri, in different languages and merge them.

Considering the dictionary of indexing terms our library uses as an existing thesaurus, we may say we are in the second type of approach to the construction of a multilingual thesaurus. Therefore, the main activity to be undertaken is translating, or rather finding equivalents for the descriptors already used. Several kinds of equivalence will be met such as:

- perfect equivalence:
  E.g., Rhetoric = Rhétorique = Retorica (82.085)
- partial equivalence:
  E.g., Various other Genres littéraires Alte genuri
    genres = divers = literare (82-9)
- non-equivalence such as the case of the Romanian word 'Doina' defining a species of lyrical poem specific to the Romanian folklore expressing feelings of love, or revenge, or mourning.
- borrowings or linguistic calques:
  E.g., Computational = Linguistique = Linguistica
    linguistics computationale computationallyal (81'322)

As much as the form of the terms is concerned, the ISO Guidelines (ISO 2788) recommends the use of nouns or syntagms and not adjectives. The most familiar or officially adopted term in a particular language should be chosen as descriptor, the synonyms or equivalents being mentioned in the thesaurus as nondescriptors

E.g., Art of writing = Art de l'écriture = Arta scrisului
    UF Creative writing (82.085)

There are two distinctive situations in which homographs may occur. They may be either
intralingual homographs - when the form of a term has more than one significance within the frame of one language:

E.g., *Sonnet* as:

a) Italian verse form consisting of 14 lines with a formal rhyming scheme;
b) short poetic form.

or

interlingual homographs - when the same form of a term is to be found in more than one language:

E.g., *Rondel* - a French verse form and a short poetic form too has the same spelling in all three languages.

Solutions:

- qualifier that is attached to the term and is not a scope note;
  E.g., *Hermeneutics and Hermeneutics (Religion)*
  *Symbolism (Literature) and Symbolism (Art)*

used for the trend or movement in literature and art at the end of the 19th and beginning of the 20th century but

*Symbolism in literature*
*Symbolism in art*
*Symbolism in religion*

used to denote the use and meanings of symbols in each of these domains

- distinction by using in each situation another form (singular and plural);
  E.g., *Sonnet (801.675.2)*
  *Sonnets (82-193.3)*

each of these being likely to be combined by means of the Boolean operator AND with Prosody and with Poetry respectively during the search.

The descriptions of the UDC notations prove to be an excellent source both for terms and scope notes:

E.g., *Philology (801)*

*SN Used for general questions relating both to linguistics and literature*
*Rhythm (801.65)*

*SN Used for verse line and syllable patterns according to numerical characteristics*
*Rhétorique (82.085)*

*SN Art de parler en publique*

When synonyms are given in the descriptions of UDC notations the preferred term becomes descriptor and the synonym (or quasi-synonym) nondescriptor:

E.g., *Satirical and humorous poems (82-17) becomes:*

*Satirical poems (82-17)*

*UF: Humorous poems*

The combinations recommended by Class 82 - Literature are treated as syntagms in the thesaurus:

E.g., *Aesthetics of literature (82:111.852)*

*F: Esthétique littéraire*
*R: Estetica literara*

*UF: Aesthetics (111.852) + Literature (82)*

In general, the literary genres are no problem when translating the corresponding terms represented in the UDC tables by the special auxiliary subdivisions. The selection of terms has still to be carefully made in order to avoid inconsistencies and overlapping concepts.
The treatment of literatures of individual languages deserves a special attention when speaking about bilingual or multilingual countries like Canada and Switzerland, for instance. We may be tempted to use

- *Canadian literature (English)* 821.111(71)
- *Canadian literature (French)* 821.133.1(71)

The same is true for

- *Swiss literature (French)* 821.133.1(494)
- *Swiss literature (Italian)* 821.131.1(494)
- *Swiss literature (German)* 821.112.2(494)

Yet, in order to be consistent with the UDC notations, the qualifiers should be not language names but country names, this implying a change in the descriptors. Then we have:

- *French literature (Switzerland)*
- *Italian literature (Switzerland)*
- *German literature (Switzerland)*

One more example is

- *Dutch literature* 821.112.5
- *Dutch literature (Belgium)* 821.112.5(493)

UF *Belgian literature (Dutch)*

The same special attention should be paid when it comes to authors whose literary output is created in a different language (or country) than their native language (or country). Thus we will have

- *Romanian literature (France)* 821.135.1(44)

and, in addition to that

*French literature + Romanian authors* for writers like Eugene Ionesco and Emil Cioran.

### 7. Conclusions

It goes without saying that the experiment with Class 8 can be generalized for the other classes of the UDC. Provided that a set of common rules starting from the international standards are established, the duplication of work can be avoided and efficiency achieved to the benefit of our work.

Indexing with words from a multilingual thesaurus will *enhance retrievability of bibliographic items by means of multiple search possibilities*. The systematic thesaurus offering more control on the terms used in subject indexing may also be considered as a *guarantee for consistency in the use of descriptors for documents with similar subjects*. In addition to that it enables the user to make comprehensive searches by giving him the possibility to move up and down the hierarchical scale.

We may speak here about the *compatibility of the information languages* the way Prof. Riesthuis (1995) distinguishes several categories in that. He states that information languages that are enumerated in some parts and faceted in others are potentially more compatible the less precordinated they are. Yet, full compatibility can hardly be reached, the experiment presented in this paper shows that there is a *practical possibility to turn to account the tremendous experience and reliability of UDC with a view to using it in the multilingual exchange of information*.

### References

### ALPHABETICAL THESAURUS—ENGLISH

#### Allegorical Poems
- **F:** Poèmes allégoriques
- **R:** Poeme alegorice
- **UDC:** 82-191
- **BT:** Poetry
- **RT:** Verse fables

#### Art of Writing
- **F:** Art d’écriture
- **R:** Arta scrisului
- **UDC:** 82.081
- **UF:** Creative writing
- **BT:** Literary technique

#### Ballads
- **F:** Ballades
- **R:** Balade
- **UDC:** 82-144
- **BT:** Poetry

#### Children Literature
- **F:** Littérature pour enfants
- **R:** Literatura pentru copii
- **UDC:** 82-93
- **BT:** Literature

#### Classical Strophes
- **F:** Strophes classiques
- **R:** Strofe clasice
- **UDC:** 801.672
- **BT:** Strophes
- **RT:** Saphic strophes

Creative writing
- **USE:** Art of writing

Diachronic linguistics
- **USE:** Historical linguistics

#### Editing
- **F:** Edition
- **R:** Editare
- **UDC:** 82.083
- **BT:** Literary technique

#### English Verse Forms
- **F:** Formes anglais de versification
- **R:** Forme engleze de vers
- **UDC:** 801.673.2
- **BT:** Germanic verse forms

#### Epic Poems
- **F:** Poèmes épiques
- **R:** Poeme epice
- **UDC:** 82-13
- **BT:** Poetry

#### Epigrams
- **F:** Epigrammes
- **R:** Epigrame
- **UDC:** 82-193.2
- **BT:** Poetry

Exegesis
- **USE:** Hermeneutics

#### French Verse Forms
- **F:** Formes françaises de versification
- **R:** Forme franceze de vers
- **UDC:** 801.675.1
- **BT:** Romance verse forms

RT: Quatrain
- **R:** Rondeau
- **R:** Rondel

#### General Linguistics
- **F:** Linguistique générale
- **R:** Lingvistica generala
- **UDC:** 81’1
- **RT:** Linguistic schools

#### Generative Grammar
- **F:** Grammaire générative
- **R:** Gramatica generativa
- **UDC:** 81’364
- **UF:** Transformational grammar
- **RT:** Generative linguistics
- **UF:** Diachronic linguistics
- **BT:** Linguistic schools
GENERATIVE LINGUISTICS
F: Linguistique générative
R: Lingvistica generativa
UDC: 81-116.6
BT: Structuralism
RT: Generative grammar
Generative semantics

GENERATIVE SEMANTICS
F: Semantique générative
R: Semantica generativa
UDC: 81*372
RT: Generative linguistics

GERMAN VERSE FORMS
F: Formes allemandes de versification
R: Forme germane de vers
UDC: 801.673.3
BT: Germanic verse forms
RT: Minnesänger stanza

GERMANIC VERSE FORMS
F: Formes germaniques de versification
R: Forme germanice de vers
UDC: 801.673
BT: Strophes
NT: English verse forms
German verse forms

HERMENEUTICS
F: Herméneutique
R: Hermeneutica
UDC: 801.73
SN: Used for textual criticism and interpretation
UF: Exegesis
BT: Studies of philology
RT: Hermeneutics (Religion)

HERMENEUTICS (RELIGION)
F: Herméneutique (Religion)
R: Hermeneutica (Religio)
UDC: 22.07
RT: Hermeneutics

HISTORICAL LINGUISTICS
F: Linguistique historique
R: Lingvistica istorica
UDC: 81-112

Humorous poems
USE: Satirical poems

ITALIAN VERSE FORMS
F: Formes italiennes de versification
R: Forme italiene de vers
UDC: 801.675.2
BT: Romance verse forms
NT: Madrigal
RT: Sonnet

LINGUISTIC SCHOOLS
F: Écoles linguistiques
R: Scoli lingvistice
UDC: 81-11
UF: Linguistic trends
NT: Historical linguistics
Structuralism
Synchronic linguistics
RT: General linguistics

LITERARY FORMS AND GENRES
F: Formes et genres littéraires
R: Forme si genuri literare
UDC: 82-1/-9
NT: Various other genres

Literary schools
USE: Literary trends

LITERARY TECHNIQUE
F: Technique littéraire
R: Tehnica literara
UDC: 82.0
BT: Literature
NT: Art of writing
Editing
Rhetoric
<table>
<thead>
<tr>
<th>Descriptor Group</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>025.43 Multilingual thesauri</td>
<td>Thesauri in multiple languages focuses on patterns according to numerical characteristics.</td>
</tr>
<tr>
<td>025.43 Multilingual thesauri</td>
<td>Thésauri multilingues</td>
</tr>
<tr>
<td>025.43 Multilingual thesauri</td>
<td>Tezaure multilingve</td>
</tr>
<tr>
<td>025.43 Multilingual thesauri</td>
<td>Thesauri</td>
</tr>
<tr>
<td>025.43 Multilingual thesauri</td>
<td>Multilingual dictionaries</td>
</tr>
<tr>
<td>22.07 Hermeneutics (Religion)</td>
<td>Herméneutique (Religion)</td>
</tr>
<tr>
<td>22.07 Hermeneutics (Religion)</td>
<td>Hermeneutica (Religie)</td>
</tr>
<tr>
<td>22.07 Hermeneutics (Religion)</td>
<td>Hermeneutics</td>
</tr>
<tr>
<td>801.6 Philology</td>
<td>Philologie</td>
</tr>
<tr>
<td>801.6 Philology</td>
<td>Filologie</td>
</tr>
<tr>
<td>801.6 Philology</td>
<td>Used for general questions relating to both linguistics and literature.</td>
</tr>
<tr>
<td>801.6 Philology</td>
<td>Metropolitan rhymes and syllables.</td>
</tr>
<tr>
<td>801.63 Metre</td>
<td>Métrique</td>
</tr>
<tr>
<td>801.63 Metre</td>
<td>Metrica</td>
</tr>
<tr>
<td>801.63 Metre</td>
<td>Prosody</td>
</tr>
<tr>
<td>801.65 Rhythm</td>
<td>Rythme</td>
</tr>
<tr>
<td>801.65 Rhythm</td>
<td>Ritm</td>
</tr>
<tr>
<td>801.65 Rhythm</td>
<td>Used for verse line and syllable patterns according to numerical characteristics.</td>
</tr>
<tr>
<td>801.65 Rhythm</td>
<td>Prosody</td>
</tr>
<tr>
<td>801.66 Rhyme</td>
<td>Rime</td>
</tr>
<tr>
<td>801.66 Rhyme</td>
<td>Rima</td>
</tr>
<tr>
<td>801.66 Rhyme</td>
<td>Prosody</td>
</tr>
<tr>
<td>801.67 Strophes</td>
<td>Strophes</td>
</tr>
<tr>
<td>801.67 Strophes</td>
<td>Strofe</td>
</tr>
<tr>
<td>801.67 Strophes</td>
<td>Stanzas</td>
</tr>
<tr>
<td>801.67 Strophes</td>
<td>Prosody</td>
</tr>
<tr>
<td>801.67 Strophes</td>
<td>Classical strophes.</td>
</tr>
<tr>
<td>801.67 Strophes</td>
<td>Germanic verse forms.</td>
</tr>
<tr>
<td>801.67 Strophes</td>
<td>Nordic verse forms.</td>
</tr>
<tr>
<td>801.67 Strophes</td>
<td>Oriental verse forms.</td>
</tr>
<tr>
<td>801.67 Strophes</td>
<td>Romance verse forms.</td>
</tr>
<tr>
<td>801.672 Classical strophes</td>
<td>Strophes classiques</td>
</tr>
<tr>
<td>801.672 Classical strophes</td>
<td>Strofe clasice</td>
</tr>
<tr>
<td>801.672 Classical strophes</td>
<td>Strophes</td>
</tr>
<tr>
<td>801.672 Classical strophes</td>
<td>Saphic strophes</td>
</tr>
<tr>
<td>801.677 Oriental verse forms</td>
<td>Formes orientales de versification</td>
</tr>
<tr>
<td>801.677 Oriental verse forms</td>
<td>Forme orientale de vers</td>
</tr>
<tr>
<td>801.677 Oriental verse forms</td>
<td>Strophes</td>
</tr>
<tr>
<td>801.7 Studies of philology</td>
<td>Étude de la philologie</td>
</tr>
<tr>
<td>801.7 Studies of philology</td>
<td>Studii de filologie</td>
</tr>
<tr>
<td>801.7 Studies of philology</td>
<td>Hermeneutics</td>
</tr>
<tr>
<td>Code</td>
<td>Description</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
</tr>
<tr>
<td>81.73</td>
<td>Hermeneutics</td>
</tr>
<tr>
<td>81'374</td>
<td>Lexicography</td>
</tr>
<tr>
<td>81'374</td>
<td>Special dictionaries</td>
</tr>
<tr>
<td>81'374.3</td>
<td>Explanatory dictionaries</td>
</tr>
<tr>
<td>81'374.5</td>
<td>Phonetic dictionaries</td>
</tr>
<tr>
<td>81'374.7</td>
<td>Dictionaries (Ordering)</td>
</tr>
<tr>
<td>81'28</td>
<td>Dialectology</td>
</tr>
<tr>
<td>81'364</td>
<td>Generative grammar</td>
</tr>
<tr>
<td>81'372</td>
<td>Generative semantics</td>
</tr>
</tbody>
</table>

**Hermeneutics**

F: Herméneutique  
R: Hermeneutica  
SN: Used for textual criticism and interpretation  
UF: Exegesis  
BT: Studies of philology  
RT: Hermeneutics (Religion)

**Lexicography**

F: Lexicographie  
R: Lexicografie

**Special dictionaries**

F: Dictionnaires spéciales  
R: Dictionare de specialitate  
SN: Used for special dictionaries according to the vocabulary they contain

**Explanatory dictionaries**

F: Dictionnaires explicatifs  
R: Dictionare explicative

**Phonetic dictionaries**

F: Dictionnaires phonétiques  
R: Dictionare fonetice

**Dictionaries (Ordering)**

F: Dictionnaires (Ordre)  
R: Dictionare (Ordonare)

**Dialectology**

F: Dialectologie  
R: Dialectologia  
UF: Areal linguistics  
NT: Dialects

**Generative grammar**

F: Grammaire générative  
R: Gramatica generativa  
UF: Transformational grammar  
RT: Generative linguistics

**Generative semantics**

F: Semantique générative  
R: Semantica generativa  
RT: Generative linguistics
Library Catalogs in the Internet: Switching for Future Subject Access

Abstract: A multitude of library catalogs are now being entered into the Internet. Their differing classification and subject headings systems used for subject access call for a switching system, a black box to facilitate the location of subject fields and their subjects in these systems. The principles on which such a switching system must be built in order to provide the necessary insight, surveyability, reproduceability and ease of concept combinability (e.g. in cases of interdisciplinary subjects) are outlined and compared with the Broad System of Ordering (BSO) which had once been established by the FID in order to serve a switching purpose. The advantages of using the Information Coding Classification (ICC) as a switching system in the Internet are demonstrated, likewise the methodology needed to establish the necessary correlation between library Classification Systems (and if possible also subject heading systems and thesauri) and the ICC. Finally some organizational implications for creating a switching for six universal systems in use are described.

1. Introduction

The national libraries of England, Finland, France, Germany, The Netherlands, Poland and Portugal are already represented in the Internet through the Internet section GABRIEL. It was announced recently, that the catalogues of some 32 national libraries in Europe will soon be connected to Internet (Gömpel, 1995a and b, GABRIEL, 1995). This implies that the Internet users will have a tremendous resource not only for formal access to descriptive data, but also for subject access to the world's literature, provided they know how to find the information they are looking for. If users are not familiar with the classification systems used by these libraries, they can only guess - probably with much frustration - which natural language and which term or combination of terms they might search with hope for success. As has been shown (Fugmann, 1991; Blair, 1996) the often praised retrieval with free text does not constitute a viable alternative because the user's search parameters normally do not occur verbally in the texts.

2. Subject Access by Interconnection?

How could one proceed in helping the users to locate the subjects sought in the literature? One way would be by viewing the classes of one classification system in its outline and its hierarchies, then by showing how each of these classes correlates to the classes of the next library system and thereafter how the first and the second systems relate to the classes of the third system, forth, fifth, sixth system. With this technique, any user will soon be totally confused. And not only will the user get lost, but also those who are trying to correlate the systems in this way. Alone to correlate the six universal systems in use at present, the Dewey Decimal Classification (DDC), the Universal Decimal Classification (UDC), the Library of Congress Classification (LCC), the Bliss Bibliographic Classification (BBC), the Ranganathan Colon Classification (RCC) and the Russian Library-Bibliographical Classification (LBC) one would need 15 correlation tables (Dahlberg, 1995a), as the following diagram shows:
Fig.1: Fifteen one-to-one connections between six universal CS

A better method would be to use a switching system. Then one would only have to correlate the different systems with this single switching system, from which the relevant classes of the other CS can then be located. Such a switching system will then work like a „Black Box“.

![Diagram of connections between six universal systems](image)

Fig. 2: Switching between six universal systems via a general system XXX

This kind of proposal was already made in the UNISIST Report (UNESCO 1971) and it was said that such a system should not be a detailed but rather a broad one. Its classes should cover only broader concepts relating to the broader concepts of other systems. When undertaking to elaborate such a system the Fédération d'Information et de Documentation (FID) called it the “Broad System of Ordering” (BSO) (Coates, 1978; 1979). We will get to this again later.

3. Principles of a Broad Switching System

It is well-known that in order to correlate it to all of the existing and possible future classes, any switching system must be more detailed in its classes than any participating classification system (CS). Hence the problem arises what should such a system look like and what should be its basis?

Several principles for such a system were laid down already some 25 years ago (Dahlberg, 1974). According to these principles, the Information Coding Classification (ICC) was designed and the system constructed (1982). Let us look at some of these principles.

3.1 Main Classes Based on General Objects

All of the existing universal CS mentioned above are based on disciplines (that is, their main classes denote disciplines) such as Physics, Chemistry, Geology, etc. or agglomerations of disciplines, as e.g. Geosciences, Social Sciences, the Arts, etc. However, what has been considered a discipline depends to a great deal on the state of knowledge at the time when the individual system was developed. This fact becomes very obvious if one compares the 10 main classes of the DDC and UDC with the 24 ones of the LCC, the 26 ones of the BBC or the 31 ones of the RCC and the LBC (Dahlberg, 1974, 180-181). Discipline-orientation means that knowledge-accumulation has been looked at as an activity-related matter. It also implies aspect-orientation: each discipline looks at an object under the aspect of its own interest. In two earlier papers (Dahlberg, 1992; 1995b), it was shown that the classification scheme which Francis Bacon devised in 1605 was based on the mental activities of Memory, Reasoning, and Imagination, as already used in the systems of Poliziano (1454-1494) and Juan Huarte (1535-1592). The main classes of Melvil Dewey’s DDC have been related via the system of W.T.Harris (1870) to F.Bacon’s system (Dahlberg; 1995c). One can therefore conclude that
the discipline-oriented library classification systems are based on the mental capabilities of mankind and as these are ever-growing, it follows that they must be a source of never-ending additions and changes with hardly any principles to guide their sequencing. Also, multiplicity of aspects relating to one and the same thing will lead to its being distributed over a number of disciplines in a classification system.

By contrast, a philosophical scheme, as has been shown in (Dahlberg, 1995b), would be based on the ontology and the special methodology of an area or a group or field of knowledge and would therefore provide the necessary objective basis in the double-sense of the word, in the sense of object-relatedness on the one hand and of non-subjectivity on the other. Thus, if a universal CS would be based on the general objects of being, looked at under the different aspects, methodologies and activities which characterize them, its main classes would consistently manifest the object-relationships and would provide an objective basis both for the developers of the system and for its users.

3.2 The Integrated Level Theory to Form the Sequence of Main Classes

If the general objects on which such a universal CS would be based are listed in a way following the Theory of Levels proposed by A.B. Novikoff (1945) and outlined independently of each other by the British philosopher J.K. Feibleman (1954) and the German philosopher Nicolai Hartmann (1964), a sequence of main classes will be obtained, in which each level presupposes the next one, thus constituting an array by which the so-called "Laws of the Levels" can be applied. The British Classification Research Group (CRG) had once tried to follow Feibleman's theory when attempting to construct a new universal system based on form classes (general categories of objects) (Foskett, 1970), however, because they failed to recognize one further necessary principle, an unsurmountable difficulty arose. This additional principle asserts that, besides the levels of being found in nature, one must also consider levels composed of man-made beings i.e. creations, productions of man and society in the form of so-called artefacts, in a material, intellectual and spiritual form. Consequently, but unfortunately, the principle of the integrative level theory was discarded by the CRG as non-applicable for the purpose in mind.

The following Fig. 3 shows the nine areas of general objects as used in the ICC.

1 Form and Structure Area (Forms, Structures)
2 Matter and Energy Area (Matter, Energy)
3 Cosmos and Earth Area (The Cosmos, The Earth)
4 Bio Area (Microorganisms, Plants, Animals)
5 Human Area (Human Beings)
6 Socio Area (Society, Societies)
7 Economics and Production Area (Technical Products and Utilities)
8 Science and Information Area (Scientific Products, Inform./Comm.Products)
9 Culture Area (Cultural/Metaphysical Products)

Fig.3: The Nine General Object Areas

3.3 The Principle of Recurrent Subdivision of Classes

Because our knowledge expresses itself in statements or judgments, the contents of such statements, i.e., their underlying concepts, follow a certain logical, syntactical structure. The Absolute Syntax for which Ranganathan had been searching (Ranganathan 1967) is contained in the primitive structure of each sentence: Subject - Predicate - Object/Complement. In fact, his facet formulae follow this order which he adapted in each discipline according to the valencies which each activity (Energy)-related concept demands. A similar structure to that of Ranganathan's facet order has been applied in identifying the facets of the ICC and constituting the components of subject areas, subject groups and subject fields.
This structure (or "structural outliner", as A.J.N. Judge called it) gave rise to the so-called "systematifier" (Dahlberg, 1995d). It includes the following nine facets:

1. Concepts concerning general and theoretical problems ("theories", "principles")
2. Concepts of the particular object of concern in an area, a subject group or a subject field, also its kinds, parts, with the subfacets of the special properties of the object and its kinds ("objects")
3. Concepts of the particular activities of a field, as related to the objects under 2, also states and processes and applied operations ("activities")
4. Special attributes related to 2 and 3 ("specialities") or
4, 5, 6. Concepts related to specialties of concepts under 2 and 3
7. Concepts influencing or coming from outside fields ("instrument", "technical relationship")
8. Application of the methods in 3 to other fields ("potential", "resource orientation", "application relationship")
9. Distribution of the knowledge contained in 1-8 according to persons, societies through teaching, documents, media, etc. ("actualization", "synthesizing", "environmental relationship")

Fig. 4: The nine facets of the Systematifier

This Systematifier can be applied on every level of abstraction. Thus it was also used for the subdivision of the nine areas of general objects of being: The scheme resulting from applying the Systematifier to these nine areas is depicted in Fig. 5.

We would like to add that the facets shown on the horizontal zero line of Fig. 5 do not necessarily apply in each of the columns shown. In some cases they must be expanded (Area 4) or repeated (Area 2 and 3) on the level of the general objects.

3.4 The Principle of a Decimal Notation

A class number or notation is an element of the conceptual structure of a system and fixes the position of a concept or of a class with its contents on the hierarchical level of a classification system. The most universal notation - applicable in all languages - is a notation based on numbers and the best one to show hierarchical levels of classes is the decimal notation (Dahlberg, 1978). Since we have been dealing in the previous sections with nine Integrative Levels and with nine Facets of the Systematifier, the decimal notation can, without problems be applied in the ICC to fix the system positions of the subject groups and their subject fields - as shown in Fig. 5. This notation has also proved fruitful in expressing the concept combinations in interdisciplinary subject fields. The progressing development of knowledge creates more and more different kinds of cross-disciplinary fields (as shown in Dahlberg, 1994), these can very easily and consistently be expressed according to contents and syntax by notational combinations.

4. What about Switching by FID’s Broad System of Ordering

After UNESCO had published its UNISIST Report in 1971 with the proposal to establish a broad switching system (UNESCO, 1971), the FID established a Working Group to elaborate such a system in 1972 and named the forthcoming system the Standard Reference Code (SRC) and later on the Subject-field Reference Code. Meetings of representatives of the FID/CR Committee and the UDC Central Classification Committee took place in the Hague between 1972 and 1974. Unfortunately no consensus could be reached between the representatives of the two committees, with the result that two proposals for what later was to be called the BSO were submitted to a Meeting of invited Experts in Sept. 1974 at The Hague. These experts decided that a panel of three persons should combine the two controversial proposals into one. It must be said that the scheme of the FID/CR members was a forerunner of the ICC applying the principles outlined above. By contrast, the scheme of the UDC representatives was one without a theoretical background for the general structure, being
<table>
<thead>
<tr>
<th>GENERAL FORM CONCEPTS</th>
<th>01 THEORIES, PRINCIPLES</th>
<th>02 OBJECT, COMPONENT</th>
<th>03 ACTIVITY, PROCESS</th>
<th>04 PROPERTY ATTRIBUTE</th>
<th>05 PERSONS OR CONT'D</th>
<th>06 INSTITUTION OR CONT'D</th>
<th>07 TECHNOLOGY &amp; PRODUCTION</th>
<th>08 APPLICATION &amp; DETERMINATION</th>
<th>09 DISTRIBUTION &amp; SYNTHESIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FORM &amp; STRUCTURE AREA</td>
<td>11 Logic</td>
<td>12 Mathematics</td>
<td>13 Statistics</td>
<td>14 Systemology</td>
<td>15 Organization Science</td>
<td>16 Metrology</td>
<td>17 Cybernetics, Control &amp; Automation</td>
<td>18 Standardization</td>
<td>19 Testing and Monitoring</td>
</tr>
<tr>
<td>COSMO &amp; GEO-AREA</td>
<td>31 Astronomy &amp; Astrophysics</td>
<td>32 Astronautics &amp; Space Research</td>
<td>33 Basic Geosciences</td>
<td>34 Atmospheric Sciences &amp; Technology</td>
<td>35 Hydroospheric &amp; Oceanol.S &amp; Technology</td>
<td>36 Geological Sciences</td>
<td>37 Mining</td>
<td>38 Materials Science &amp; Technology</td>
<td>39 Geography</td>
</tr>
<tr>
<td>BIO-AREA</td>
<td>41 Basic biological Sciences</td>
<td>42 Microbiology and Cultivation</td>
<td>43 Plant Biology and Cultivation</td>
<td>44 Animal Biology and Breeding</td>
<td>45 Veterinary Sciences</td>
<td>46 Agriculture &amp; Horticulture</td>
<td>47 Forestry &amp; Wood Sci. &amp; Technology</td>
<td>48 Food Science and Technology</td>
<td>49 Ecology and Environment</td>
</tr>
<tr>
<td>SOCIO AREA</td>
<td>61 Sociology</td>
<td>62 State and Politics</td>
<td>63 Public Administration</td>
<td>64 Money and Finances</td>
<td>65 Social Aid, Social Politics</td>
<td>66 Law</td>
<td>67 Area Planning, Urbanism</td>
<td>68 Military Science and Technology</td>
<td>69 History Science and History</td>
</tr>
<tr>
<td>ECONOMICS &amp; TECHNOLOGY AREA</td>
<td>71 General and National Economics</td>
<td>72 Business Economics</td>
<td>73 Technology in general</td>
<td>74 Mechanical &amp; Precision Engineering</td>
<td>75 Building</td>
<td>76 Commodity Science &amp; Technology</td>
<td>77 Vehicle Science and Technology</td>
<td>78 Transportation &amp; Technology &amp; Services</td>
<td>79 Utilities and Service Economics</td>
</tr>
<tr>
<td>SCIENCE &amp; INFORMATION AREA</td>
<td>81 Science of Science</td>
<td>82 Information Science</td>
<td>83 Informatics, computer science</td>
<td>84 Information in general</td>
<td>85 Communicat. Science</td>
<td>86 Mass Communication</td>
<td>87 Printing and Publishing</td>
<td>88 Communicatinol Engineering</td>
<td>89 Semiotics</td>
</tr>
<tr>
<td>CULTURE AREA</td>
<td>91 Language and Linguistics</td>
<td>92 Literature and Philology</td>
<td>93 Music and Musicology</td>
<td>94 Fine Arts</td>
<td>95 Performing Arts</td>
<td>96 Culture Sciences, narrower sense</td>
<td>97 Philosophy</td>
<td>98 Religion and Secret Teachings</td>
<td>99 Christian Religion</td>
</tr>
</tbody>
</table>
based again on disciplines for the main classes and using a notation running from 100 to 992. For the subdivision of the main classes it used a centesimal, two-digit notation, separated from the first three numbers by a comma. The work of the three experts was finished in 1978 and presented in a scheme at the Edinburgh FID Conference (Coates et al, 1978). A manual (Coates et al, 1979) followed in 1979. The BSO has also been translated into French and was used in several applications, mostly for demonstration purposes. In this scheme, the principles outlined under 3.1 - 3.4 above were not applied: it used neither the Integrated Level Theory to build a natural array of the main classes nor the Systematifier Principle to combine classes and concepts at the relevant system positions. Some years ago, the FID gave up its ownership of the BSO and the chairman of the 3-man panel, E.Coates continued to elaborate further the scheme in depth, to computerize it and to make it available on diskette. In my view, the greatest handicap for an easy use of the scheme is its centesimal notation; although it seems to offer a lot of possibilities for subdivisions on one level, it does not group the facets together on any one level and it disregards the fact that the human mind usually cannot keep in its memory more than nine digits at one time. Seven digits is already very much, one usually memorizes only 3-6 digits at a time (Judge, 1978/79). By contrast, in the ICC system, the Systematifier uses 3 times 3 digits for its nine facets, then facilitating the mental as well as the practical handling of the system.

5. The Suitability of the ICC for Switching in the Internet

5.1 Origin of Subject fields

The subject fields included in the ICC were derived from very many different sources such as guides, directories, lists of university courses, lists of R&D institutions, special libraries, documentation centers, journals, etc., all of which guaranteed that the names used for these fields represented existing areas of activities. Thus some 13,000 names were collected. The names of these fields were checked against many universal and special lexica, dictionaries, glossaries in order to locate their definitions. By doing this, a lot of synonyms could be identified. Interestingly, this search also turned up names of fields without any actual activity corresponding to them. When no definition of its concept could be found, a field-name was usually discarded.

5.2 Structure of the ICC

The bigger a system may once become, the more its future structure must be taken into account in order that it not - so-to-speak - collapse through further development. As the ICC is a universal classification, comprising all fields of knowledge, it must have the strongest structure possible, that is, its form elements/categories must be like steel pillars supporting a huge construction. Therefore the ICC structure has been based in the first place on general categories of objects/entities and the aspects under which these can be seen were put into the second place. But these aspects are also based on categories, having been arranged recurrently as shown above (3.3).

5.3 Mnemonics of the ICC

The more a system is equipped with mnemonics, the easier it is to be understood and applied. The 3x3 object areas in their natural sequence and the 3x3 facets of each subject group and field make it easy to memorize its special element positions. Once the system is understood in its systematic structure, one can handle its classes with ease, and will find the way from there to the other systems in any switching case.

5.4 Ease of Screen Display and Search

The systematic display of the ICC in diagrammatic presentation - should fill a single screen. With the aid of a telescoping facility, one can click onto one of the field cells and then retrieve the related subject fields and their general categories.
zoom into the next and further subdivisions. Using the terms/captions and notations found, further searches can be carried out in one or all of the other systems.

6. Implementing the Switching Process

How would one go about using the ICC for switching? It is proposed to do this for the six universal systems mentioned above (see Fig.2). From the beginning it must be clear that only subject fields are to be correlated, i.e. fields, the name of which shows that not just an object or an activity is meant but rather a combination of the two, for example, not just ‘information’ but ‘information science’, or, if a subject field is represented by only one term, like ‘metrology’ or ‘forestry’, it is the definition that reveals its field character.

The first step is to correlate the classes of one classification system after the other with the subject groups of the ICC, using the compatibility matrix according to the methodology outlined in the Guidelines for the Establishment of Compatibility between Information Languages... (Dahlberg, 1995a). Doing this, it becomes obvious at which positions in the correlated systems there are gaps or only partial equivalences corresponding to the concept in question. For problems of this sort, a series of symbols taken from the mathematical symbolization languages (such as < , > ) were introduced. It may also happen that, although the concepts in question are equivalent, their names, even in English differ. In this case, the differing names should be added as synonyms to the principal name used in the ICC.

The second step would then be to establish a tabular alphabetical subject index to the terms in the compatibility matrix; in this table, the notations and the different terms of each system should be placed in a standard sequence; say ICC, DDC, UDC, LCC, BBC, RCC, and LBC.

The ICC comprises some 6,500 subject fields. The correlation with the other systems would bring about 6,500 times 6 = 39,000 entries in tabular form (hopefully most of these will not be expressed in all too divergent terms. In fact, the subject index will probably be of smaller size, as in many cases, there will be a verbal overlap between subject fields and, in other cases, no equivalence, although synonyms will also have to be considered.

The two procedures mentioned will yield a) a systematic and b) an alphabetical display of all the concepts under comparison.

The third step will consist in ironing out inconsistencies in the systems under comparison: filling in the gaps and to seeing to it, that they receive the correct symbolization. In order to do this properly, the alphabetic index of the contents of the compatibility matrix will be helpful.

7. Searching via the Alphabet and the Systematic Display

Having raised above the idea of a black box, we meant to suggest that the background correlation of the different classification systems with the ICC should be done in a way hidden for the user. At the surface, the user would have two ways to search:

1) when selecting a subject field name from the alphabetical index of the terms in the compatibility matrix, the user obtains automatically the corresponding class notations in the other systems and from there to the relevant references. Should the user wish to investigate further the systematic environment of his selected class he or she should have the option of asking for the classification system in question and the systematic display relevant to his/her request;

2) when selecting a subject field in the systematic display of the ICC, the user finds the correlated class numbers and captions of the other CS in the compatibility matrix, arranged systematically according to the ICC. Thereafter he/she could proceed further as described under 1).
8. Some Organizational Questions

In order to realize the proposals of this paper and to elaborate the switching mechanism using the ICC, the following organizational problems need to be clarified.

1) Language: All of the systems under comparison must be in English.

2) Background definitions. Concepts must be defined. They can be defined by their hierarchical position in a system or by using external sources. It would be helpful, if all of the concepts under comparison were related at least to a generic and partitive definition.

3) Definitions in English: All of the subject fields of the ICC are related to a definition in German. An English translation of these definitions should be made available.

4) Experts knowledge in subject fields and in classification: Despite an availability of definitions the actual correlation must be done by experts in broad areas of cognition, e.g., the natural sciences, the human and social sciences, engineering sciences, humanities, etc., assisted by classificationists.

5) Procedure. The correlation work must be done on different levels of abstraction, i.e. one should start with the 81 subject groups of the ICC, find their equivalents in the other systems, thereafter go into each of these 81 groups and find the equivalents to the 729 subject fields and so on.

6) Working places: The work must be done at a location where (a) all universal systems are available, (b) good lexicons can be consulted, (c) computers are accessible, (d) a classificationist and a programmer can be consulted.

7) Timing. Depending on the availability of all necessary resources and manpower, the work could be done in a short time of e.g. 3-6 months.

8) Further steps. In the course of the correlation process, the ICC can be progressively elaborated in depth, using the concepts found in the different systems correlated. To do so, one need only to document the concepts found in the course of correlation. In addition, further fields from the latest documentation of research, e.g. in the Vademecum of European Research and University Institutions, could be included after consulting the representatives of these institutions. Finally, one could start correlating subject heading lists and thesauri, field by field.

9) Institutionalization: It is obvious that once such a program will be started, it needs to be supported, continued and represented by a permanent institution responsible for its maintenance. If this program is accepted, as a valuable contribution to library networking and user satisfaction, the ISKO should be asked to create the necessary institution and to find a sponsor for its support.

9. Conclusion

We started by recalling that the classes of a switching system must be more detailed than the classes of any of the participating classification system. In its hitherto published form the ICC is available only to the third level of abstraction. In reality, however, it has in many subject groups been worked out in considerable more depth. Moreover, because all of its concepts refer to subject fields, it is already more detailed in this regard than any of the existing general systems based on disciplines. In addition, as described above, further elaboration can be generated in the course of the correlation work. Thus there is not difficulty in fulfilling this primary requirement of a switching system.

Here a word to the ISKO. If we see ourselves as a society not only in theoretical, but also in practical knowledge organization, then we are called to become active in the way described in this paper. Given the urgency of the task described, I am convinced that it must be possible to find institutions capable of, and prepared to sponsor a project of this type. The need to improve subject access to library systems in the Internet grows more urgent from day to day, we have only to make the decision whether we want to get involved and proceed in doing what needs to be done.
Notes

1 GABRIEL stands for Gateway to and Bridge between European National Libraries.
2 Feibleman tried to formulate the laws indicated by Novikoff for the different levels by elaborating 18 "Laws of the levels", 6 "Rules of explanation" and 5 "Rules of procedure. He showed also that on each level there existed 1) characteristic entities, 2) processes, and 3) properties coming from 1) and 2).
3 The card-file amounting now to some 6,500 field concepts with the stock of definitions is still held in the office of the INDEKS Verlag.
4 We would like to add here that the ICC has been published in German and English with its subject groups and subject fields so far on three levels of abstraction (Dahlberg, 1982). It exists therefore in print in a short form and fully elaborated sometimes down to the sixth hierarchical level in machine-readable form (with definitions in German and equivalents of the terms in English). Thus a first step towards realization of the task would be to publish the entire system with its definitions in German and in English. With this having been accomplished the 3 steps mentioned under 6. could be attacked with respect to one system after the other and one step after the other, whereby Step 2 could be accomplished automatically. It would be quite helpful, if, in addition to the systematic display of the CS in question, their indexes were also available in English. This does not hold for the BBC, the new edition of the RCC and the LBC.

References


Dagobert Soergel  
College of Library and Information Services, University of Maryland

SemWeb: Proposal for an Open, Multifunctional, Multilingual, System for Integrated Access to Knowledge about Concepts and Terminology

Abstract. This paper presents a proposal for the long-range development of an open, multifunctional, multilingual system for integrated access to many kinds of knowledge about concepts and terminology. The system would draw on existing knowledge bases that are accessible through the Internet or on CD-ROM and on a common integrated distributed knowledge base that would grow incrementally over time. Existing knowledge bases would be accessed through a common interface that would search several knowledge bases, collate the data into a common format, and present them to the user. The common integrated distributed knowledge base would provide an environment in which many contributors could carry out classification and terminological projects more efficiently, with the results available in a common format. Over time, data from other knowledge bases could be incorporated into the common knowledge base, either by actual transfer (provided the knowledge base producers are willing) or by reference through a link. Either way, such incorporation requires intellectual work but allows for tighter integration than common interface access to multiple knowledge bases. Each piece of information in the common knowledge base will have all its sources attached, providing an acknowledgment mechanism that gives due credit to all contributors. The whole system would be designed to be usable by many levels of users for improved information exchange.

1. Introduction

"The global information society of the 21st century will rely increasingly on an information infrastructure which must have two essential components: the global telecommunication and electronic networks epitomized by the Internet; and, underpinning it, a conceptual infrastructure reflecting the way knowledge and information are organized." (From the recommendations of the ISKO/Polish Information Society Joint Seminar on the Compatibility of Order Systems, Warsaw, Poland, Sept. 13-15, 1995)

This paper deals with the intellectual infrastructure. It presents the vision and basic architecture of a system for integrated access to data on concepts and terminology. The system would bring together for the user data from a variety of sources that up to now exist largely in separate worlds, including dictionaries, thesauri, and classification schemes. Implementation would require a collective effort by a number of people. People interested in participating in such a project are invited to communicate with the author.

2. Rationale for the Proposed Access System and Knowledge Base

There are now a multiplicity of order systems and classifications, terminological knowledge bases, and linguistic dictionaries, each serving a limited purpose but widely overlapping in their content. The proposed common interface would provide simultaneous access to all concept and terminology knowledge bases on the Internet with limited integration of information "on the fly"; going a step further, the proposed integrated distributed knowledge base would provide a home for all of these knowledge bases who care to join, eliminating
duplication while preserving the integrity of each source and establishing relationships across sources; . As outlined below, the proposed system would serve many functions and thus justify the major investment it requires. It might appear that this proposal is overly ambitious and that serving so many functions at once is impractical. However, the information to be added for each additional function becomes less and less for each function added, and it is precisely the multifunctionality that makes the considerable investment pay off. Since this is an investment into infrastructure, it should properly come from the public rather than the private sector. Thus, the proposed knowledge base would lead to a savings in development effort and a potentiation of usefulness through the assembly of rich information from many sources that complement each other and through establishing relationships among the concepts and terms from different sources.

* Savings in development effort. Much effort is being expended in developing individual knowledge bases of concepts and terminology limited by subject, application, and language. These individual knowledge bases overlap considerably; thus there is much duplication of development effort which would be saved in the environment provided by the proposed knowledge base.

* Potentiation of usefulness. An integrated knowledge base provides rich information on a concept or term, much more so than any specialized system. It provides linkages across scientific and scholarly disciplines. It explicates fine differences of meaning that exist across languages. It makes conceptual structures that are explicit in one language available to users in other languages. Thus an integrated knowledge base has a usefulness that transcends the sum of its parts.

3. Functions of a Knowledge Base on Concepts and Terminology

Knowledge on concepts and terminology, especially knowledge on classificatory structure, can serve one or more of the functions listed in Figure 1. The proposed system could serve all of these functions directly. It could also serve as an environment for projects developing order systems (thesauri/classifications), concordances between order systems, linguistic dictionaries, etc.; in the ideal case, specialized ordering systems, dictionaries, etc., both machine-readable and printed, could be produced simply by extraction from the integrated knowledge base.

4. The Structure of the System

The structure of the system is presented here on the conceptual level, giving a user's view, without making any assumptions about the underlying implementation. The user's view is captured in a template for the arrangement of information about concepts and terms, information obtained from a search of multiple existing knowledge bases and/or from a common integrated knowledge base.

4.1 The SemWeb Template

Figure 2 gives the first draft of an overall outline of the SemWeb template or frame, a list of frame slots (or slot groups) to organize information about concepts and terms. Some slots pertain exclusively or primarily to terms as linguistic entities; others pertain exclusively or primarily to concepts.

The template provides the basis for the common interface: The user starts with a
The basic functions of a knowledge base on concepts and terminology — on which all others rest — are to provide a semantic road map to individual fields and the relationships among fields, to relate concepts to terms, and to provide definitions, thus providing orientation and serving as a reference tool.

Many specific functions build on this foundation:

- To improve communication and learning generally: To assist writers and readers, to support learning through providing conceptual frameworks, to support language learning, and to support the development of instructional materials.

- To assist researchers and practitioners in exploring the conceptual context of a research or practical problem — a research project, policy, plan, or implementation project — and in structuring the problem, thereby providing the conceptual basis for the design of good research and implementation. This includes consistent definition of variables and measures used in research and evaluation to make results more comparable and cumulative.

- To provide classification for action, for example, such as a disease classification to assist with diagnosis, a classification of medical procedures for insurance purposes, or a classification of products for customs agencies.

- To support information retrieval, including knowledge-based support of end-user searching (menu trees, assistance in the analysis of a search topic based on a facet structure, browsing a well-structured hierarchy to identify search concepts, mapping from the user's query terms to descriptors used in a database or to the multiple natural language expressions to be used for free-text searching), hierarchically expanded searching, support of well-structured displays of search results, providing a tool for indexing (vocabulary control, request-oriented or user-centered) indexing), support for the combination of or unified access to multiple databases through providing a common search language to multiple databases, providing a common index language for a number of databases in a field, or mapping indexing descriptors from one system to another.

- To serve as the conceptual basis for knowledge-based systems.

- To serve as a monolingual, bilingual, or multilingual dictionary for human use.

- To serve as the dictionary/knowledge base for automated language processing — including machine translation, data extraction, automatic abstracting and indexing, and natural language understanding generally by providing both the syntactic and the semantic knowledge needed for correct parsing. This includes spell checking and grammar checking.

- To provide a classification/ontology for data element standardization.

Figure 1. Functions of a knowledge base on concepts and terminology
Entry term, concept, or group of terms or concepts (identified through a suitable identifier for the entity, preferably the system-wide identifier)

Other identifiers for the same entity

Broader and narrower frames

Spelling variants (other character strings in the same language)

Pronunciations (with dialect/regional variations and frequency information), in a phonetic alphabet or as digitized sound.

Word root and derivation from the root

Compound terms, phrases, idioms of which the word is a part.

Etymological origin, history (from this etymological cognates in other languages can be inferred and displayed)

Part of speech, inflection rules, and other syntactic information (possible positions in a sentence, rules on combination with other terms to form expressions) (see below for semantically-based combination rules).

Terminological information: Other terms with the same or similar meaning in the same language and in other languages.

Definition

A preferred definition in English, French, etc., and other definitions and scope notes.

Definition in a formal definition language, possibly arranged in a frame hierarchy.

Semantic components, componential or feature analysis. Relevant feature space, necessary and sufficient features. Semantic root and derivation from the root.

For categories: Examples, prototype(s), members with degree of typicalness

For meanings that refer to concrete objects: a picture of the object and/or a picture that shows the designated object as part of a larger whole (as in a visual dictionary).

Usage notes, usage examples and quotations, familiarity and frequency information. For a group of terms that are close in meaning, subtle differences in meaning may be explained through text elaborating on the definition and usage of the terms, with examples.

Category level (basic level, above basic level, below basic level), qualified by population and population subgroup

Detailed conceptual relationships (Broader terms / hypernyms, narrower terms / hyponyms, parts / meronyms, the whole of which the concept is a part / holonyms, concepts with which the concept at hand combines often / compound terms, etc.) and pointers to the concept's place in overall classificatory structures.

Display of the structural relationships among subordinate concepts (a hierarchy, an association map, or a diagram or table showing relationships (closely linked with definition and usage, for examples see the Longman lexicon)

Rules on combination with other concepts to form expressions. For concepts that express relationships, especially verbs: A case frame. Slot filler restrictions in the case frame will define some aspects of usage.

Figure 2. The SemWeb template: Frame slots for information on concepts and terms
template, fills in a term or concept identifier (possibly choosing from a classification displayed as a menu tree), and highlights the slots whose information she wishes to see. The system then accesses all relevant sources it knows about, extracts the information needed, and presents the filled-in template to the user. The template helps the user identify the kind of information wanted and it provides the framework for integrating the information found and organizing it for display. The template also provides the basis for the system’s internal workings: It provides the framework for organizing the system’s knowledge about what information can be obtained from what knowledge bases and how to search each knowledge base. It serves as an input form for contributing data to the common distributed knowledge base. And finally, a frame hierarchy is one useful view of the internal structure of the common knowledge base.

The template focuses on information on individual concepts and terms. To support many of the functions listed above, the system must also show overall conceptual/classificatory structures in various formats (linear listings, two-dimensional maps, etc.) with adequate browsing/navigation capability allowing the user to move from a general overview to detailed classifications. The system should provide access to multiple views, some corresponding to the arrangement in present sources. Multiple views are essential to make an integrated system workable for multiple groups of users, each with different requirements for conceptual arrangement, information given, main language, etc. Some of these views are grand structures of knowledge, such as the great library classification schemes; others are local overviews, such as the tables in the Longman Lexicon that represent the relationships between a number of terms, such as the various specialized terms for horse (filly, mare, stallion, etc) or the usage of various terms for father (dad, daddy, papa, etc.). An overview is referenced in the appropriate slot for the most specific concept that still covers its entire scope. A frame for the universal concept serves as the reference point for universal classifications.

The system must provide for formal definitions that can capture fine nuances of meaning and usage. This is particularly important for establishing the proper correspondence between different languages. Definitions of the various meanings of a word might be arranged in a frame hierarchy as proposed in Chernyatin (1995).

4.2 Entity Types and Entity Identifiers
The SemWeb template gives a broad picture of the system’s conceptual schema. We need to further specify the entity types for which frame instances can be created:

- **Character strings**
- **Terms** (words and phrases, including idioms and slang expressions)
  - Linguistic roots and derived terms in both stem and inflected forms
- **Concepts** (semantic roots and concepts derived through semantic modifiers)
- **Groups/classes of words/terms or concepts** for which some common assertions hold.
  Examples: All fifth declension Latin nouns; all English verbs that agree in their conjugated forms with sing (sing, ring, drink, etc.); all adjectives that could mean either a color or a race (such as white and black) and consequently share a semantic rule: They refer to a color when they qualify a non-human entity and to a race when they qualify a human entity. Frames for groups can represent grammatical knowledge in the same format as lexical knowledge.

A flexible system for identifying words/terms and concepts that uses the identifiers given in existing knowledge bases and is therefore compatible with the coexistence of many
independent knowledge bases is shown in Figures 3 and 4. Note that the same character string may refer to different words in different sources. The word number distinguishes several words in one language represented by the same character string occurring in the same source. Terms or concepts can also be identified by a source ID and a unique term or concept number within a source.

<table>
<thead>
<tr>
<th>EN drill n OED 1 (rivulet)</th>
<th>EN drill n W3 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN drill n OED 2 (tool etc.)</td>
<td>EN drill n W3 5 ... AHD 1 ... RHD 1</td>
</tr>
<tr>
<td>EN drill n OED 3 (monkey)</td>
<td>EN drill n W3 6 ... AHD 4 ... RHD 4</td>
</tr>
<tr>
<td>EN drill n OED 4 (furrow)</td>
<td>EN drill n W3 7 ... AHD 2 ... RHD 2</td>
</tr>
<tr>
<td>EN drill n OED 5 (fabric)</td>
<td>EN drill n W3 9 ... AHD 3 ... RHD 3</td>
</tr>
<tr>
<td>EN drill v OED 1 (draw out)</td>
<td>EN drill v W3 9</td>
</tr>
<tr>
<td>EN drill v OED 2 (trickle)</td>
<td>EN drill v W3 9</td>
</tr>
<tr>
<td>EN drill v OED 3 (bore)</td>
<td>EN drill v W3 8 ... AHD 2 ... RHD 2</td>
</tr>
<tr>
<td>EN drill v OED 4 (sow)</td>
<td>EN drill v W3 4 ... AHD 1 ... RHD 1</td>
</tr>
</tbody>
</table>

Figure 3. Word identifiers: Language; character string; part of speech; source; word no.

(1) EN drill n OED 2.1 (tool)
(2) EN drill n OED 2.4 (military exercise)
(3) EN drill n OED 2.5 (a person who drills others)
(4) EN drill n OED 2.6 (rigorous training)
(5) EN drill n OED 3.1 (Mandrillus leucophaeus)
(6) EN drill n AHD 1.2 (disciplined, repetitious exercise, esp. military) (includes 2 and 4)
(7) EN drill n AHD 1.3 (specific exercise designed to develop a skill) (broader than 4?)
(8) EN drill n AHD 4.1 (Mandrillus leucophaeus)
(9) EN drill n RHD 1.2 (military exercise)
(10) EN drill n RHD 4.1 (Mandrillus leucophaeus)
(11) FR drill n HD 1.1 (Mandrillus leucophaeus)
(12) FR drill n HD 2.1 (military exercise)

Figure 4. Concept identifiers: Terms with sense discriminators of the form .x

The identifiers thus constructed are unambiguous but not unique; a word or a concept has as many identifiers as it has sources. The common knowledge base will establish a correspondence between the different identifiers for the same word and likewise for the different identifiers for the same concept and, to the extent feasible, establish a system-wide identifier, which has the same form with the source ID for the system itself. Group entities require specially constructed identifiers.

5. Sources of Information for the Proposed System

A tremendous amount of information has been amassed and codified in many existing sources. The system will provide access to as many of these as possible. The common knowledge base will—incrementally over time—include as much of this information as is feasible under legal restrictions and limitations of processing.

- **Lexicons and ontologies** from linguistic projects and knowledge-based systems.
- **Monolingual, bilingual, and multilingual dictionaries**, both general and specialized, including guides to usage (e.g., Fowler’s) and guides to concepts (e.g., Kohl, 1992).
• Terminological standards.
• Order systems / subject access vocabularies (thesauri, classification schemes, etc.) used for information retrieval and other purposes.
• Data dictionaries of large information systems.
• Laws and regulations (food regulations contain definitions for many foods, drug laws classify drugs into "schedules" based on their psychoactive effects, etc.)

6. Development of SemWeb

The development of SemWeb requires incremental work on a number of major tasks. Fortunately, development can build on many projects already underway; SemWeb would bring their results together in a unified framework. Two principles make a system of this magnitude possible: multiple contributors and virtual integration, the principles on which the World Wide Web itself and systems like OCLC and software like LINUX are based.

Tasks required for the development of the common interface

• Develop a "super standard" for any kind of information on concepts and terms. A good template/frame structure is central to the success of the system. While there exist standard formats for machine-readable dictionaries, subject authority files, and classification data, there is no one format for all the types of data on concepts and terms as envisioned here. The existing standards must be brought together and augmented to accommodate even very specialized lexical projects.

• Develop a comprehensive master list of machine-readable and printed resources, each described following a standard schema, based on existing partial lists.

• Develop the software for the system: a kernel (gets the user’s request, selects the knowledge bases to be searched, integrates the information found, and displays it to the user) and special modules for searching specific knowledge bases. The master list serves as a knowledge base for this software, including information on user fees and on copyright status (to determine whether information can be copied into the common knowledge base or can be included only by reference).

Tasks for the incremental development of a common integrated knowledge base

• Level 1: The system keeps the concept or term records it assembles in response to a user request, replacing copyrighted information with a reference, and keeps a directory of these records, regardless of where they are stored. For the next request for the same concept or term, only the knowledge bases that have been updated or that contain copyrighted information need to be accessed again.

• Level 2: The user can edit a concept or term record and store the edited copy, thus adding value by establishing correspondences between the numbered entries from several dictionaries (see Section 4.1) and/or creating system-wide identifiers for words and word senses; establishing correspondences with concepts in classification schemes; synthesizing a natural language definition that is better than any of the definitions found in dictionaries; creating a frame hierarchy of formal definitions of word senses. Some of this editing must be done anyway before the information gathered from several sources can be used. The system allows users to share the fruits of their labor.

• Level 3: Development of a well-structured knowledge base that integrates knowledge from many sources. The structure of such a knowledge base must be designed in
accordance with the super standard mentioned above. The integration of information from various sources can be automated to a large extent, using intelligent software that builds on existing work. The information produced through editing would also be used, with the structure of the Web pages facilitating such incorporation. The system could acquire further knowledge from the analysis of text and of term use in searching.

SemWeb is conceived as a federated system with multiple collaborators by subject, application, or language specialization, and with data distributed over multiple sites but appearing to the user as a unified system. Each contributor and each user has a status (which might include description of expertise along several dimensions); some collaborators might be recognized as official contributors, others might just use the knowledge base for their project - the system would provide an environment for the more efficient development of specialized concept and terminology knowledge bases, while at the same time providing efficient storage and wide access to the results. More casual users could also add their own information and suggest additions and corrections, with mechanisms for quality control. Users could restrict retrieval to information entered or reviewed by a contributor meeting certain status requirements.

7. Access to SemWeb

Access to SemWeb would be provided in multiple ways:

- Through the World Wide Web and other online means.
- Through integration into search systems, giving transparent ("behind the scenes") assistance or explicit system suggestions to be modified as needed.
- Through products derived from it (special classifications, dictionaries, etc.).

8. Conclusion

By providing integrated access to a wide variety of lexical and classification knowledge bases and by providing a forum in which users can augment these resources by sharing lexical and classificatory knowledge, SemWeb creates the conceptual infrastructure that is urgently needed to reap the full benefit from global information exchange that is driven by the same information infrastructure that makes SemWeb possible.

Bibliographic References

1. Sample dictionaries and other lexical resources

AHD    The American Heritage dictionary of the English language.
HD     Harrap's new college French and English Dictionary
OED    The Oxford English dictionary.
RHD    The Random House dictionary of the English language.
W3     Webster's third new international dictionary.

McArthur, Tom. Longman lexicon of contemporary English.
2. Other lexical resources


Unified Medical Language System (UMLS).


Good search: http://www.notredame.ac.jp/cgi-bib/wn.cgi

3. Standards for lexical knowledge bases

ISO CD 12620.2: Computational aids in terminology — Data categories.

ISO DIS 12200, Computational aids in terminology — Terminology Interchange Format (TIF) — An SGML application.


4. Other references


Wright, Sue Ellen, and Melby, Alan K., TEI-TERM: A proposed format for the interchange of terminology data using standard generalized markup language. 200-216.


Wright, Sue Ellen. Creating a data element dictionary for computer-aided terminology work. 169-186.

Is a Picture Worth a Thousand Words?
Classification and Graphic Symbol Systems

Abstract: Icons are graphic images with functional roles in human-computer interaction. They may be used as conceptual tools to represent the organization of information or as operators effecting an activity such as printing or moving to another document. An icon may represent its referent either as a sign, a purely arbitrary relationship that must be learned by the user; as a pictograph, a visual image of the object represented; or as an ideogram, where the referent is not a concrete entity but an attribute, a set of attributes, or an abstract concept associated with the referent. The symbolicity of an icon reflects the degree of representativeness that obtains between an icon and its referent(s). We propose to examine symbolic languages composed of sets of icons and to assess their effectiveness as classificatory structures in terms of: 1.) representation of hierarchical structure; 2.) level of symbolicity; 3.) contexts that promote the capability of icons to represent organization; 4.) relationship between an underlying metaphorical framework and iconic representation of the organization; 5.) graphic elements of effective symbolic languages; and 6.) social or cultural factors related to the effectiveness of icons.

1. Introduction

One of the most interesting phenomena engendered by recent technological innovations is the enhancement and popularization of computer-based graphics. As recently as 1992, a popular text on information organization claimed that "Documents are comprised of words" (Rowley, 1992, 166). But advances in the generation, storage and transmission of images in digital format have heightened awareness of the information-bearing role of images and have led to a general recognition of the need for increased access to images in both traditional and digitized formats through the development of visually-oriented cataloging and indexing techniques (Austin, 1994; Shatford Layne, 1994).

While classification research has focused on developing new methods for representing the visual and conceptual content of images (Jörgensen, 1995), enhanced graphic capabilities have influenced how information is organized in a host of computer systems and applications. Once the exclusive province of the Macintosh user, the GUI, or graphic user interface, has become the standard interface for Windows operating systems; and traditionally keyboard-driven applications such as WordPerfect have incorporated the use of icons. The point-and-click navigation provided by hypertext-based systems such as the World Wide Web has further enhanced the utility of graphic images by encouraging their use not merely for initiating system activities, but also for providing access to information. Thus, while research is currently investigating how to represent the content of digitized images, graphic images are being employed as icons both to indicate the functional capabilities of a computer application and to represent the intellectual content and the relational or associative organization of collections of information.

2. Signs, Pictographs and Ideograms

Arnheim (1969) observes that an image functions at one of three representational levels: as a sign, standing for "something" but not reflecting the perceptible characteristics of its referent...
(e.g., the diagonally crossed bars within a circle that serve as the traffic sign indicating a railroad crossing), as a picture, employing a relatively low-level representation of the perceptible qualities of its referent (e.g., the silhouette of a deer to indicate a deer crossing on the highway); or as a symbol, representing its referent at a higher level of abstraction that is frequently, but not always, metaphorical (e.g., the profile of a soldier to represent the concept "war").

A sign constitutes a one-to-one mapping of a graphic to the entity or concept for which it stands, but the basis for the relationship established between the sign and its referent is purely arbitrary. Use of an enclosing circle with a single diagonal bar to represent "no" or "not" is a purely arbitrary convention. In contrast, a picture, or pictograph, is the visual image of an object that refers directly to the object represented. It is, as Gittins points out, "a graphic representation whose decomposition makes interpretation impossible" (Gittins, 1986, 520-521). A pictograph provides a one-to-one or one-to-many mapping between the graphic image and its intended referent(s). For example, in a word processing package, an image of a printer is a pictograph used both to represent and to initiate the system's printing function.

Symbols, however, are more abstract than either signs or pictographs. A symbol is an ideogram (Krull, 1988) in that its referent is not a concrete entity but either an attribute, a set of attributes, or an abstract concept that is characteristic of or associated with the intended referent. There is a one-to-many mapping between symbol and referent in that the referent generally exists not as a specific object but as an interaction between the intended category term as referent and the immediate context. For example, a bent arrow at the end of a long hallway is a directional sign indicating simply that the hallway continues in the direction shown; but a highway sign that represents a sharp curve in the road as a bent arrow is an ideogram in that it is not intended to stand for the curve itself, but to indicate to the approaching driver the need to slow down before entering the curve. Similarly, the picture of a cowboy boot placed over a shelf of books is an ideogram indicating the intellectual content or general subject area of those books (i.e., "Westerns").

A set of signs, pictographs and/or ideograms, when used within a single system or system component, is frequently organized around one general metaphor that serves to constrain the range of possible referents for any given image. For example, the dominant metaphor employed by many word processing packages is that of the "office": graphic images represent objects or activities that are readily understood within the context of traditional office tasks. The set of images developed to handle specific tasks, to represent system functions, or to indicate related components within a single application or informational unit exists, therefore, as a symbolic language.

A symbolic language is a visual language. More specifically, a symbolic language can be defined as an integrated set of graphic images—of signs, pictographs and/or ideograms—usually organized around a central metaphor, that represents a specific domain and is intended to convey information about that domain through visual, non-linguistic means. It is important, therefore, to distinguish between a symbolic language and the symbolicity of a visual language. The symbolicity of a visual language refers to the level of representativeness that obtains between individual images and their intended referents. As such, it may range from the purely arbitrary level of the sign, through the concrete level of the pictograph, to the more complex level of the abstract ideogram. The symbolicity of a visual language may be pure, in that it consists of visual images that function at a single level of representation (e.g., all images are arbitrary signs), or it may be mixed and include images drawn from across the three levels of sign, pictograph and ideogram.
3. Icons

A graphic image, whether sign, pictograph or ideogram, that has a functional role in a computer interface is an icon. GUIs rely on icons of familiar objects to facilitate user interaction with the system. The use of familiar objects encourages inferences about system activities. When selecting a particular icon, then, the user must rely on preexisting knowledge or familiarity with the domain to recognize those attributes which will identify the intended referent in order to invoke and manipulate system processes (Benbasat and Todd, 1993). Because both the semantics and the syntax of this interaction are pre-defined (Gittins, 1986), effective dialogue between the user and the system can be established without the user's explicit knowledge of an underlying command language.

An icon is generally defined as a graphic image that conveys a single, unique meaning and serves as a visual label for an object or process (Benbasat and Todd, 1993; Gittins, 1986; Krull, 1988; Lohse, Biolski, Walker and Rueter, 1994). Thus, for example, Lohse et al. (1991, 433) describe an icon as a form of "shorthand notation for a label [from which] the meaning of each icon can be readily extracted," and Ockerse and van Dijk (1979) define icons as "signs which have features in common with the object, and characterize, picture, or imitate the object" (1979). Thus, emphasis has traditionally been placed on the graphic rather than the conceptual component of the icon—on its service as a visual signpost pointing to a concrete entity, rather than on its potential as an organizational tool within the system.

4. Functionality and Modality of Icons

One current phenomenon not accounted for by the more conventional approach to icons is their utility in providing access to specific materials in hypertext-based systems such as the World Wide Web. Associative links between conceptually related materials can be represented by icons that are activated to effect navigation between documents. Furthermore, as the complexity of many web sites is compounded by the introduction of new documents, icons are employed not simply to provide access to information but, more broadly, both to represent the intellectual content of related materials and to indicate the overall conceptual organization of a collected body of information.

The use of icons either as conceptual tools to represent the organization of information or as navigational tools to access information has not received much attention in the literature. There is, instead, a tendency to focus on the visual characteristics of icons and icon sets rather than on their functional capabilities. Thus, Gittins describes a GUI as "a system populated by 'objects' which the icons represent" (1986, 527); and Krull warns against the use of icons to represent abstract topics unless "the referent (the concept to be represented) is concrete, [in which case] it should be possible to find a pictorial symbol for it" (Krull, 1988, 258).

With the enhanced graphic capabilities of current systems, however, it is imperative that researchers distinguish between graphic images whose inclusion is purely decorative or illustrative in nature and those which serve a functional purpose in effecting an activity, whether the object of that activity involves initiating a system process such as printing, navigating between linked documents, or providing access to the conceptual organization imposed upon a body of related materials. As demonstrated by these examples, the component that distinguishes an icon from a decorative image is the ability of the system user to interact with and manipulate the system through the selection and activation of an icon so as to produce a system response.

The functional distinction between visual graphics and icons is critical for a comprehensive definition of icon, but it is also important to distinguish icons from linguistic strings that provide functionality in a computer-based system. For example, while hypertext systems such as the
World Wide Web use words and phrases as concept-based links between related pieces of information, these alphanumeric strings are simply links and should not be considered icons. Krull (1988) clearly distinguishes between alphanumeric, or digital, encoding and graphic, or analog, encoding. He points out that, while digital encoding uses alphanumeric characters to directly represent an object, concept or activity in its entirety, analog encoding is metaphoric in that it employs a selected few of the referent's physical or conceptual properties to indicate the intended meaning. From Krull's perspective, then, graphic icons are simply analog encodings. But such a distinction between analog and digital encoding does not preclude the association of a graphic icon with a linguistic label. There is, in fact, an extensive body of literature that compares the utility of single modality icons that are purely graphical in their composition with that of mixed modality icons that incorporate an alphanumeric label to support the intended meaning of the graphic image.¹

For the purposes of the present study, then, an icon is defined as a graphic representation of the intended referent—either as sign, pictograph or ideogram—in single (graphic) or mixed (graphic plus text) modality that allows the user to interact with and manipulate the underlying system to produce a desired response.

5. Research Background

Attempts to understand how icons work—and how they work best—have generally focused on the design of icons. One of the most widely cited articles is Easterby's "The perception of symbols for machine displays" which appeared in 1970. This work has served as a jumping-off point for research that assumes a strong correlation between the graphical composition of an icon and its success in conveying information—research that tends to focus on an identification of those graphical components which contribute to the optimum composition of meaningful icons and icon sets.

Easterby identified five components that contribute to the referential meaning of a graphic image. The first three of these components are structural principles and contribute to the establishment of "contextual cues [for] the observer who is attempting to define the meaning of the symbols" (Easterby, 1970, 150). The last two components are figural properties which determine the shape and form of the graphic within the visual field, thereby either facilitating or hindering an appropriate and meaningful interpretation.

- **Pragmatic component**: The context within which a graphic image is to be used or interpreted.
- **Semantic component**: The relationship between the graphic image and the object, concept or activity to which it refers.
- **Syntactic component**: The relationship between the set of graphic images that comprise the symbolic language.
- **Visibility component**: The visual clarity of the graphic image.
- **Discriminability component**: The ease with which one graphic image can be differentiated from other graphic images in the same set.

While each of these components contributes to the conceptual meaning of the graphic image, it is the sum interaction between the five components that constitutes the essential perceptibility of a graphic. Although perceptibility has been subsequently described as "the degree to which a symbol may be quickly and reliably understood in the context of its intended use" (Boling, Johnson, and Kirkley, 1994, 6), Easterby presented the notion of perceptibility as an evaluative measure of symbol performance that would reflect integration between the processes of perception and discrimination and thereby "ensure that the meaning of a symbol is accurately and readily
perceived" (Easterby, 1970, 152).

In line with his efforts to identify "the optimum definition of more meaningful shapes", Easterby proposed that the evaluation of graphic images should attempt to "define the best features for perception" (Easterby, 1970, 152). In so doing, he established a research agenda whose emphasis on identifying figural properties and their interaction has influenced much of the subsequent research. Geiselman, Landee and Christen (1982) developed an index of perceptual discriminability based upon measures of the configural and graphical attributes that were shared by the images within an existing symbolic language. The index could then be used as a performance-based criterion for selecting candidate images to be incorporated within the existing language. In two studies investigating modality in icon construction, Guastello, Traut, and Korieneck (1989) evaluated groups of icons on the basis of meaningfulness and found that mixed modality icons were generally rated by subjects as more meaningful than either verbal or graphic single modality representations, a preference which extended across a range of domains. The authors concluded, however, that, while there is a strong interrelationship between meaningfulness, memorability and the efficiency of system operation, future research should attempt to define dependent measures that are more closely related to the notion of efficiency than are ratings of meaningfulness.

Gittins (1986) attempted to create a classification for graphic images based upon common design features (e.g., type, form and color) and shared display structures (e.g., boundary, location, and figure grounds). While his effort is interesting, it fails to capture essential commonalities that could be used to construct a true classification of icons. More recently, Lohse et al. (1991, 1994) have undertaken a series of empirical experiments intended to provide the foundation for a taxonomy of visual representations based upon measures of visual similarity across a wide range of graphic presentation methods including, among others, maps, graphs, tables, and time charts. Their research identified eleven classes of visual representations, including a class identified as "icons". It is interesting, however, that their findings indicated that subjects viewed icons as unable to convey much information in that each icon represented a one-to-one mapping between the graphic and its referent.

The studies discussed here provide only a very brief overview of research focusing on the domain of icons, but they are highly typical of work in this area. Published research has concentrated on identifying the figural properties of an effective icon or icon set, on the ease of learning and memorability of single and mixed modality icons, and on measures of user performance that reflect meaningfulness and/or memorability. There appears to be very little research addressing the functional ability of icons to represent the conceptual organization of information.

6. Focus of the Proposed Research

The proposed research will attempt to assess the effectiveness of symbolic languages as classificatory structures. To this end, we will focus on the ability of a set of icons to generate a coherent graphic vocabulary that is capable of representing a meaningful organization for a set of documents while supporting effective and efficient access to materials through manipulation of the icons that comprise the language.

To address the problem of the effectiveness of symbolic languages as classificatory structures, a sample of symbolic languages will be evaluated in light of six specific focus areas.

The first focus area addresses the question of hierarchical structure, which we define as three or more levels of representation that evince a superordinate/subordinate structure. The analysis will concentrate on the overall organizational structure of the information. Although an attempt
will be made to determine if there is an observable preference for either monohierarchical or polyhierarchical systems, the primary emphasis will be on the ability of symbolic languages to represent a hierarchical structure.

The second focus area investigates the level of symbolicity at which graphic images are more effective in representing classificatory structure. Each symbolic language will be analyzed to identify the predominant form of graphic representation—sign, pictograph or ideogram. The symbolic languages examined in the study will then be compared to determine if there is any correlation between symbolicity and the ability to represent classificatory structure.

The third focus area addresses the related questions of context and metaphor. Specifically, it will attempt to determine if there are specific contexts within which a symbolic language can more effectively represent organizational structures. The icons that comprise each symbolic language will also be analyzed, both individually and in association, to identify the presence of a general metaphorical framework.

The fourth focus area extends the notion of a metaphorical framework and investigates the possibility that there is a correlation between an effective representation of the organization of information and either the context or a governing metaphor.

The fifth focus area attempts to determine if there are specific graphic elements that are shared across either the more effective or the less effective of the symbolic languages studied. Because the configural elements of icons are not a primary concern of the proposed study, this focus area may receive relatively less attention.

The sixth and final focus area investigates social and/or cultural differences that limit the effectiveness of the icons in a symbolic language.

7. Scope of the Research

The ability of graphic images to indicate the classificatory structure of a collection of documents constitutes a potentially significant problem. When this problem is extended to the organization of information within a highly unstructured system such as the World Wide Web, performance based on context-free measures of meaningfulness or memorability tends to evaluate individual icons in isolation. This approach may fail either to reflect the underlying utility of the graphic representations or to provide an adequate measure of the effectiveness of a particular symbolic language. Because of the unstructured and evolving nature of the Web and other hypertext-based systems, the performance measure employed to evaluate the efficiency of the dialogue between user and system must be able to reflect the user's ability to "accurately and readily" (Easterby, 1970, 152) infer the meaningful intent of an icon. To this end, efficiency of the symbolic language will be measured as the ability of the user to identify and manipulate icons that locate, or retrieve, the appropriate information.

Because symbolic languages that are used to indicate conceptual structure may demonstrate varying levels of symbolicity, we propose to examine the efficiency of user-system interaction for three different kinds of symbolic languages:

1. Many public and school libraries employ graphic classification labels to provide rapid recognition of the intellectual content of an item. These highly stylized graphic labels are available from commercial library supply companies and rely on simplified metaphorical images that indicate very general categories of content. For example, the shape of a turkey placed on a contrasting background may be used to identify materials related to the American Thanksgiving holiday.

2. The Bookhouse virtual library interface uses icons to represent a range of fiction categories (Pejtersen, 1989). Bookhouse images appear to be less stylized than library classification
labels, possibly reflecting the heightened contextuality provided by the central metaphor of a "house of books".

3. Some sites on the World Wide Web use icons both as indicators of conceptual organization and as tools for discrimination, evaluation and navigation. The precise referent of a graphic representation may be less significant at this level than the insight it provides about the classification of a potentially complex body of information or the ability to distinguish one collection or organization from another. Because the need to communicate a general structure relies upon semantic discrimination between icons, the ability of the symbolic language to indicate organization in such a system may be more dependent on the interaction between icons than it is in the two systems described above.

In order to understand how these three types of symbol systems serve to conceptualize, or organize, the relevant information space, each symbolic language will be evaluated to distinguish the context in which it operates, to identify governing metaphors and to assess the level of symbolicity for the icons that comprise that particular language. Using the structural principles and figurai properties that Easterby (1970) has identified as constituting a meaningful graphic, a measure of the perceptibility of each language will also be constructed.

8. Conclusion

The proposed research will investigate the ability of symbolic languages to represent a coherent classificatory structure for a collection of documents. To assess the effectiveness of a symbolic language as an organizational tool, we will use a performance measure that reflects the efficiency of a given language in providing access to documents within the classified set. By investigating the relationship between effective representation of the classificatory structure and the analyses of hierarchy, symbolicity, context, metaphor and social and cultural limitations, it may be possible to identify certain primary characteristics of an effective symbolic language that can be used in the construction of more meaningful and therefore more efficient symbolic classificatory structures.

Notes

1. There are a number of studies comparing single modality mixed modality icons. The following list is short but nonetheless representative of much of the research being conducted in this area:


References


C. Olivia Frost  
School of Information, University of Michigan

The University of Michigan School of Information  
Art Image Browser: Designing and Testing a Model for Image Retrieval

Abstract: A team at the University of Michigan School of Information designed, implemented and is evaluating a World-Wide Web (WWW)-based experimental system which uses classification to facilitate browsing of art images. The research team built a database of approximately 3,000 digitized images from Art History to determine if classification can be used for retrieving images from a digital database in a networked environment. A key premise in this research is that browsing can serve an important role in retrieving image information. The system was evaluated in controlled tests and through a questionnaire available to WWW users.

1. Introduction

The School of Information (SI) Art Image Browser was developed as part of a research project which utilized classification as a means to group image sets into meaningful categories that support browsing. As Kwasnik has observed, browsing relieves the user from the burden of formulating a precise search strategy and capitalizes on the fact that it is easier to recognize what is interesting or useful than it is to specify it in advance. A related advantage is the ability to facilitate exploration without prior knowledge of subject content, and thus enable users to cross over into unfamiliar domains (Kwasnik, 1992). Furthermore, browsing lends itself particularly well to visual images since a pictorial image is able to present itself whole (Arnheim, 1970, 211) and in its own medium of expression.

The browsing approach employed in the SI Art Image Browser provides the user with access to image sets rather than to individual images. Shatford Layne (1994) points out advantages to this type of approach—a user searching for a single image may be able to make a selection by browsing through a group of images. Abstract or subjective concepts are difficult to convey in a textual description. Or, scanning a selection of images may prove more efficient than indexing at a high level of detail. Shatford Layne further argues that “the emphasis when indexing images should be on recall, rather than on precision. Rather than devoting time to extraordinarily detailed or complicated indexing ... it might be better to concentrate on indexing the basic elements of an image and rely on scanning, or browsing, to make the fine distinctions” (1994, 586).

Methods for accessing image databases have relied primarily on either indexing of individual items or automatic image recognition (Cawkell, 1992; Leung, 1990). Both approaches have their limitations. While systems based on feature-extraction and object-recognition can provide powerful means of searching images by their visual content, these automated approaches are often computationally expensive and tend to be domain specific.

Indexing of visual images by textual descriptors is very time consuming and labor intensive. In addition, image collections pose problems of intellectual access because of the different meanings that images can convey to different users (Besser, 1990). Since images are often subject to a wide range of interpretations, textual descriptions can only begin to capture the richness and

complexity of a visual image. In addition, textual strategies are often unable to take advantage of modalities of searching and processing image information (Römer, 1995).

While textual descriptors have their limitations, they can be used in a way that allows users to employ their cognitive abilities to scan image content within sets of images to retrieve desired information. The SI Art Image browser research project used broad classification units to group image sets into meaningful categories that support browsing. The project work centered on activities needed to: 1) build a database of images, 2) design, build, and implement a system, 3) identify classification categories, and 4) evaluate the resulting system with real users using an iterative approach.

2. Building the Database

In developing an image database which contains data from content providers, and which will be available over the Internet, the project team found itself confronted with the issues of copyright, image quality, content depth and balance, and metadata control. Copyright constraints affected the image content depth and balance, while metadata control issues often slowed the development of the database and limited the categories which could be used in building the data fields. Copyright issues required that we work with a collection which could provide images cleared for use on the Internet, and we soon learned that ownership of images does not imply ownership of copyright. The providers of our image content supplied metadata which had been created for the collection over a number of years, and which had previously been used primarily on an internal basis. Faced with the prospect of their data being made public over the Internet to a world-wide audience, the data providers wanted the quality of the data to meet higher standards than had previously been necessary.

To develop a classification retrieval system for images from a given content domain, attention must be given to content balance and depth. Too few or too many images of a given type, style or genre will lead to an imbalance in the category content, and may tend to discourage comparison of images, or reduce its diversity and interest value to users.

3. Designing, Building and Implementing a System

The database was set up on a World Wide Web (WWW) server and can be viewed at: http://www.sils.umich.edu/Art_History/index.html This provided us considerable flexibility in providing access paths to information, and also allowed the interface to be tested and reconfigured in response to feedback from users.

The user has a choice of beginning with a browse or search approach. In choosing the browse approach, the user begins by making a selection from a menu organized by broad conceptual categories (artist, title, date, medium, object type, and subject), which takes the user to listings of subdivision groupings which in turn lead to sets of thumbnail images. The user can then click on the thumbnail to retrieve a document containing the thumbnail and textual information pertaining to the image. This document page also leads to full-page image.

At any point in the browse stage, users can choose a "search" option and execute a query in any of the category fields of artist, medium, date, etc. These queries can be used to generate image sets which can in turn be scanned. The project developed a search engine for searching textual descriptions and access points. The FTL retrieval application (Full Text Lexicographer) provides enhanced search capabilities and supports several kinds of searches: words and phrases, boolean operations, word and sentence proximity, and stemming. FTL allows searches on any of the fields in the database, and the user is able to ask, for example, what the collection has by a
given artist within a specific time span using a specific medium.

Since our system was developed with the needs of the end-user in mind, feedback from present and potential users played a key role. The project team used an iterative approach and questioned art history faculty and students—as well as librarians—to determine patterns of use and information searching practices.

We created prototype systems to solicit feedback from users and used this input to inform later versions of the system. The first prototype contained a database of 100 images from 20th century art. A second prototype focused on Italian architecture, and contained over 200 images. This version incorporated improvements suggested by the first group of consultants. The first pilot system was showcased to a group of about 100 students, faculty, and staff from the School of Information and other units of the University, and informal feedback was solicited from the viewers. Subsequent interviews with faculty from the History of Art Department, the School of Art, and the College of Architecture and Urban Planning elicited helpful feedback on the database structure. In addition, focus group sessions were held with students in an Art History class.

4. Selection of Classification Categories

Working with domain specialists in art and architecture, the project team investigated several existing models for classification or categorization, including the Library of Congress Classification, the Dewey Decimal Classification (DDC), the Fogg classification schemes, a modified Metropolitan Museum of Art classification system, and the Art & Architecture Thesaurus.

We decided that the first pilot database would be a good experimental platform on which to test the feasibility of using the DDC classification scheme as one means to guide browsers through a collection of digital images. Student team members set to work assigning images to DDC headings, and were attracted to its hierarchical structure and flexibility. However, most of the images in this database were 20th century art images, and the students found DDC to be inadequate in its ability to handle the description of abstract images.

Our initial plan was to provide one system for domain (art history) specialists, and another for general users. Our prototype database system allowed us to test whether this dual approach is appropriate, or whether one system can be modified to suit both sets of users. Responses from non-domain and domain users led us to conclude that a single system could suit both audiences. In discussions with students and faculty in art history and fine arts, we were told that the idea of using a broadly conceived browsing tool to create a subset of the image database, with the capability of bringing up a search engine in order to create a thumbnail gallery based on fairly specific criteria, would be an effective way to encourage use of the database by art history professionals.

As we moved into the creation of the second prototype database of architectural images, we looked at classification and description methods in greater detail. The Classification for Slides and Photographs: Western Art employed by the University of Michigan Department of the History of Art, had been found to be inadequate for our prototype of 20th-century images, but offered a more fully developed scheme for the classification of architecture, particularly Italian architecture. We also examined the art documentation and classification scheme employed by the Canadian Centre for Architecture, with which one of our team members had had prior experience. This classification scheme provided further input from which we developed categories and browsing hierarchies for the prototype of architecture images. Since our eventual database was broader than architecture, we adapted some categories for more general use.

The Art and Architecture Thesaurus (1990) served as our major source for broad categories to be used in the eventual system. It also served as a good source for the standardization of
descriptive terminology and for notes to assist non-domain browsers in understanding some discipline-specific terminology.

Eventually, the determination of the classification categories and structure was most heavily influenced by the characteristics of the data provided by the content providers, and we found that we had to work within the bounds of categories contained within the museum curators' cataloging data.

The lack of universally-accepted standards and vocabulary control for art documentation remains a barrier to sharing data about art works and cultural materials between institutions. As a result, the development of art object databases at most institutions has tended to evolve along fairly idiosyncratic lines, in accordance with the most immediate in-house data requirements. Unlike printed works which may exist in identical form in thousands of locations, art works are by their nature unique. As a result, descriptive standards may be viewed by some curators as unnecessarily restrictive.

Although there are certain kinds of information about artworks that are typically displayed with an image (e.g., artist, title, date, medium) a great deal of latitude remains in the selection of displayable data fields. The choices made will depend largely on assumptions made about the usefulness of certain fields to the intended user base, and on the richness of the data that comes from the image provider. Since the fields and field content of the data will vary to some extent between different institutions, the data provided will need to be normalized for display in the intended online image collection. In addition, there may be issues of incomplete or absent information that must be addressed. This may require that guidelines be established to set reasonable limits on the time and labor that will be invested to verify missing or incomplete data for art objects.

Working from the Art and Architecture Thesaurus (AAT), we selected seven facets for use in the browser. These broad conceptual categories serve as a starting point for users to begin their browsing. The categories also serve as fields within which the user can execute a search for a particular artist, title, etc. By combining terms at the point of the search, the broad categories can be assembled by the user to formulate more precise search terms to be used for more specific browsing or a specific known-item search.

We examined categories in the AAT (indicated by italics) with a view towards adopting or adapting them for use as categories in the browser:

- *Agents* was used for artist's names.
- *Space* was used to group images according to the nationality of the artist. A world map enables users to click on the continent of interest, and to be led from there to a breakdown of time categories within the geographic location.
- *Time* is used to break the content into chronological sub-divisions.
- *Materials* was used to designate medium; i.e., what the material is made of, or the process used to make it. *Objects* was used for type of artistic creation, e.g., painting, sculpture. The categories under *Objects* and *Materials* usually generated large sets which we wanted to further subdivide topically. Team members with backgrounds in History of Art were assigned the task of determining subcategories to break out the large sets into more manageable categories. Instructions were given not to spend inordinate amounts of time in determining the designations, since the purpose behind the research was to determine if quickly determined categories would be useful, and would in fact preferable to no subdivision. A number of taxonomies were created for this purpose. However, this effort proved too time-consuming to accomplish within the standards the team members had set for themselves and the deadline of the system production.

We added a *Subject* category to depict objects depicted within the image, but not necessarily the theme of the image. Since the content providers had not followed a standardized list in the assignment of subject terms, the vocabulary could not be controlled. One group within
the project team sorted out the list of subject terms from the original list to arrive at a small set of frequently occurring categories considered to be of general interest to users, for example, Animals, People, Places. This set of categories was determined by the non-expert group members, who may have felt less constrained than the domain members of the team in arriving at a general set of topical categories.

*Styles and periods* is an AAT category type that we had hoped to include. Stylistic period was used as a category in the second generation prototype. However, the museum curator content providers felt that these designations could not be assigned with confidence and accuracy, since many works had overlapping or ambiguous periods. An attempt may be made in future to map stylistic periods to chronology and location, and to tolerate a certain level of inaccuracy in the interests of a larger goal of providing a commonly understood frame of reference for the user.

5. Classification Culture Clash

The development of an organizational structure is of key importance to an online collection that offers a browsing mode for seeking images. Like the selection of data fields for display, the organizational scheme must be structured with its intended users' needs in mind. Ideally, one would build the classification scheme directly from the provided data, but this is not always possible. Most significantly, art object databases developed to meet the needs of museum curators often contain highly specialized information that is of no interest or value to the general user, such as information about inscriptions or watermark details.

Alternatively, information of proven interest to general users, such as information that would allow images to be grouped by stylistic periods, typically does not appear in databases developed by curators and other expert groups. Perhaps this is because art and art history experts are already familiar with what these stylistic periods are, know which works tend to typify which styles, and therefore do not typically record this information in their databases. Secondly, the art history community appears to disagree on the exact date ranges for these various stylistic periods, and individual curators may be reluctant to assign a given work to a stylistic category. In addition, the data was not originally created for the purpose of facilitating browsing in an online database. As a result, it may be necessary to supplement the data provided with added intellectual content that will facilitate classification, such as the very useful work done by student team members to categorize images by genre, such as landscapes, portraits, abstracts, etc. At issue is the amount of time and labor available to invest in this supplementary work, which can be considerable for large online collections. We would also need to address the issue of the usefulness and functionality of classification schemes where the attribution of images to various browsing categories is incomplete or inconsistent.

6. Evaluating the System

The SI Art Image Browser is being evaluated in three principle studies: 1) an examination of system usage by domain and non-domain experts, 2) an examination of human-computer Interactions, and 3) a survey of World Wide Web users. The evaluation of system usage compares the three primary image searching techniques—browsing, direct searching, and a combination of the two—and utilizes a quasi-experimental design. By evaluating how successfully different sample groups complete tasks using the three systems, inferences about system functions different user populations require can be made. The human-computer interaction study utilizes a variety of heuristics, cognitive walk-through, and key-stroke analysis to identify problems with the design of the user interface. An online survey of World Wide Web users will provide additional data on
general system use patterns.

7. Getting It Perfect Vs. Getting It

In the course of the project, a great deal of attention and time was spent in ensuring the consistency of the data that appeared in the various fields. Experience with the content providers contributing to the project revealed that art object cataloging that is carried out over a protracted period, by numerous individuals, and often leads to inconsistencies in the data that can be very difficult to eradicate. Unfortunately, there is no reliable automated means to make these corrections. At the present, there is no acceptable substitute for manual verification of each component of the data by team members with a good subject knowledge—a long, intensive and reiterative process. However, these data "housekeeping" issues were believed to be crucial to the accuracy and functionality of the image retrieval processes of the online collection, and the curators and student team members felt it could not be neglected without detriment to credibility of the project. Since content providers tend to be very conscious of their public image and reputation, we considered it important to satisfy their concerns about the accuracy of the data, even if the inconsistencies or errors originated with the content providers themselves.

However, the time-consuming efforts required to meet existing standards of some data providers will not be feasible in dealing with large-scale collections, and may result in many important collections remaining unclassified. It still remains to be seen if a middle ground between expert classification and no classification can be found in the process of determination of categories to enable users to browse image sets. The Dublin Core Metadata initiative to identify data elements which can be supplied by content providers of online resources is a useful example (Weibel, 1995). It remains to be seen if image content providers who are not professional catalogers or curators can be provided with a scheme they can apply to do their own quick classification and indexing for image sets in a distributed environment. This project has provided some lessons in the complexities of testing this premise.

8. From Art History to Earth and Space Science

The application domain for the Art Image Browser was art history. However, the strategy can serve as a model for intellectual access in image databases of other subject domains. The basic browsing concept used in the Art Image Browser project is being adapted, extended and tested in the University of Michigan's Digital Libraries Initiative sponsored by NSF/ARP A/NASA. This project will enable us to apply the concept of classification of images on a much larger scale and in the very different subject domain of earth and space sciences. It can be viewed at:

http://www.sils.umich.edu/Art_History/testarea/bin/vlad/htdocs/index.html

Acknowledgments

This research was sponsored by a grant from the U. S. Department of Education. The author is indebted to the student team members of the project, and especially to Anne Noakes, for her assistance in the writing of this paper.

References


Abstract: Recent research (Jörgensen, 1995) investigated image attributes as reported by participants in describing, sorting, and searching tasks with images and defined forty-six specific image attributes which were then organized into twelve major classes. Attributes were also grouped as being "perceptual" (directly stimulated by visual percepts), "interpretive" (requiring inference from visual percepts), and "reactive" (cognitive and affective responses to the images). This research describes the coverage of two image indexing and classification systems and one general classification system in relation to the previous findings and analyzes the extent to which components of these systems are capable of describing the range of image attributes as revealed by the previous research.

1. Introduction

Large collections of textual materials have existed for millennia, and the organization of textual materials for retrieval enjoys a long history and a repertoire of established techniques such as indexing and classification to facilitate the process. Today, with a number of recent technological innovations such as digital cameras and computer-based graphics packages, images and tools for their creation are widely accessible. Digitization and transmission of images over long distances is also now possible using both network and satellite technology. These capabilities have combined to produce a growing number of collections of images stored in digital format. These collections of images, just like their textual counterparts, require means for accessing the information stored within them.

However, indexing of images has lagged far behind that of indexing of textual materials. Images, by their very nature, are capable of eliciting such a number of individual interpretive and perceptual responses that any attempt at indexing them beyond a minimal, bibliographically-based level of description, seems to be an almost insurmountable task. Many authors have written about the complexity of image interpretations (Panofsky, 1962; Arnheim, 1974; Drabenstott, 1986) and the literature of image indexing suggests that it is a complex and poorly understood process (Lynch, 1991, Shatford, 1986; Turner, 1993).

While there have been several studies aimed towards understanding the types of queries formulated by different groups of image users (Enser, 1993; Bakewell, 1988; Keister, 1994), there has been little empirical investigation of those attributes of images which are typically described by humans in various types of tasks and interactions with images. Knowledge of the range and distribution of image attributes typically described in different types of situations can provide an empirical basis for choices as to which image elements are most usefully indexed for retrieval.

2. Results of Recent Research

A recent research project (Jörgensen, 1995) gathered evidence to address this gap in our
knowledge by asking subjects to describe images in several types of tasks. The researcher analyzed the image attributes in both verbal and written descriptions using the constant comparative technique to define attribute types and build higher level classes of attributes. For the purposes of this research, an image attribute was not limited to purely perceptual elements, but included a variety of cognitive, affective, or interpretive responses to the image as well; thus attributes included semantic, symbolic, or emotional characteristics of the image. There were a total of twelve attribute CLASSES and forty-six individual Attributes. The twelve CLASSES were LITERAL OBJECTS, PEOPLE, PEOPLE ATTRIBUTES, COLOR, CONTENT/STORY, LOCATION, DESCRIPTION, ART HISTORICAL INFORMATION, VISUAL ELEMENTS, ABSTRACT, VIEWER RESPONSE and EXTERNAL RELATION.

The research indicated that type of task (describing, sorting, or searching) has a major effect on the distribution of image attributes that are typically reported by subjects. Numerical distribution of attributes was not the only factor evaluated; term consistency and term level (superordinate, "basic level," and subordinate) in classes and attributes were also analyzed. Additionally, evaluation and analysis of the context in which image attributes occurred (term order, application of conceptual frameworks such as "Figure-Ground") added to the informativeness of the data. Suggestions were offered for ways in which researchers may apply the results to the development of heuristics for indexing of images for generalized access and drew upon both empirical results and theoretical considerations.

The tasks which provided the richest data were three describing tasks and a sorting task. Data for the three describing tasks were very similar. Those classes most typically reported across the three describing tasks are LITERAL OBJECTS and PEOPLE, followed by COLOR, CONTENT/STORY, LOCATION, and DESCRIPTION. The specific Attributes most typically commented upon within these classes are: Object, People, Body Part, Color, Activity, and Location - Specific. In the Sorting Task, both image descriptions and category names for groups of images were analyzed. Classes appearing most frequently in the descriptions were ART HISTORICAL INFORMATION, VIEWER RESPONSE, ABSTRACT, PEOPLE, and CONTENT/STORY. The researcher did not evaluate terms in these Classes for consistency, due to the variable format of the verbal protocols. For category names, the most typically reported attributes are ART HISTORICAL INFORMATION, ABSTRACT, and CONTENT/STORY. Category, Object, and People are the most consistently described attributes within category names.

<table>
<thead>
<tr>
<th>CLASS</th>
<th>Describing</th>
<th>Sorting</th>
</tr>
</thead>
<tbody>
<tr>
<td>LITERAL OBJECTS</td>
<td>29.3%</td>
<td>8.7%</td>
</tr>
<tr>
<td>PEOPLE</td>
<td>10.0%</td>
<td>8.6%</td>
</tr>
<tr>
<td>COLOR</td>
<td>9.3%</td>
<td>2.6%</td>
</tr>
<tr>
<td>CONTENT/STORY</td>
<td>9.2%</td>
<td>8.5%</td>
</tr>
<tr>
<td>LOCATION</td>
<td>8.9%</td>
<td>0.8%</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>8.0%</td>
<td>2.7%</td>
</tr>
<tr>
<td>VISUAL ELEMENTS</td>
<td>7.2%</td>
<td>3.8%</td>
</tr>
<tr>
<td>ART HISTORICAL</td>
<td>5.7%</td>
<td>23.8%</td>
</tr>
<tr>
<td>PEOPLE ATTRIBUTES</td>
<td>3.9%</td>
<td>4.0%</td>
</tr>
<tr>
<td>EXTERNAL RELATION</td>
<td>3.7%</td>
<td>8.7%</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>2.0%</td>
<td>13.8%</td>
</tr>
<tr>
<td>PERSONAL REACTION</td>
<td>2.9%</td>
<td>14.1%</td>
</tr>
</tbody>
</table>

Table 1: Comparison of classes for describing tasks (overall average) and sorting task
Three groups of attributes were typically noted by subjects across all the describing and sorting tasks:

- 1. PEOPLE—the human figure.
- 2. CONTENT/STORY—the "story" of a picture.
- 3. PEOPLE ATTRIBUTES—relationships, emotion, and mental activity.

The sorting task data indicate that similarity among images cannot be represented solely by "perceptual" attributes, but must take into account "interpretive" attributes (abstract concepts, symbols, and themes) as well. The variety of image attributes noted by the subjects reinforces the notion that images are capable of stimulating complex verbal and written descriptions and that creating an indexing tool for images is a challenging task. However, this variety also suggests that multiple approaches to indexing of images would be beneficial.

Existing classification systems have been described as "sparse" indexing languages for the visual content of images. However, the current research suggests the need to access other elements of images besides those of specific visual content. Therefore, one possible approach to indexing of images is not to create yet another indexing or classification system, but first to define those specific attributes to which access is needed and then to investigate existing indexing systems for their potential to contribute to this access. The current research evaluates the utility of two existing indexing and classification systems for images, the Library of Congress Thesaurus for Graphic Materials: Topical Terms for Subject Access (LCGTM) and the Thesaurus Iconographique (TI), and compares them to a general subject classification, the Dewey Decimal Classification. As each system has a different focus and is organized according to somewhat different principles, there is not an exact one-to-one correspondence between the named categories as reflected in the samples and the specific attributes or classes of attributes as revealed in the previous research. Therefore, data for each are reported in terms which reflect the source itself; this preserves the distinctiveness of each system but does not prevent meaningful comparison with the attribute data from the current research. The group terms from the indexing and classification systems analyzed reflect more closely the larger class level of the previous research rather than the individual attribute level.

3. Comparison with Existing Indexing and Classification Systems

The LC Thesaurus for Graphic Materials: Topical Terms for Subject Access (LCGTM) and the Thesaurus Iconographique, System Descriptif de Representations (TI) are two generalized systems in widespread usage which provide subject and content access to images, and while they both state the same general goals for their application as indexing and classification tools for images, they represent two contrasting approaches.

According to the introduction, the LCTGM "provides a substantial body of terms for subject indexing of pictures, particularly large general collections of historical images" and "offers catalogers a controlled vocabulary for describing a broad range of subjects, including activities, objects, and types of people, events, and places depicted in still pictures" (Parker, 1987, v). It does not include art historical and iconographical concepts but does supply terms for some abstract ideas. It provides for the inclusion of proper name terms but does not include them.

The LCTGM contains 3,567 authorized terms and almost as many cross-reference terms. The main source of terminology for the LCTGM is the Library of Congress Subject Headings, but terms are also drawn from the Art & Architecture Thesaurus, (terms for painting, drawing, architecture, and the decorative arts), the Legislative Indexing Vocabulary (contemporary political and social issues), and all terms found in the thesaurus Descriptive Terms for Graphic Materials: Genre and Physical Characteristic Headings (GMGPC), to be used in cases where types and
formats of graphic materials are the *subjects* of images (Parker, 1987, viii.).

The researcher analyzed a sample of terms from the LCGTM to determine which pictorial attributes the thesaurus includes. The researcher selected a subset of approximately 10% of the authorized terms (392) by sampling every tenth page of the thesaurus and evaluated these terms for the types of pictorial content which they address and for the extent to which they correspond to attributes described in the current research.

As can be seen from Table 2, the LCGTM is strongly reflective of the sources from which it draws its terms, especially previous LCSH subject headings, which themselves are strongly reflective of historical concerns such as industrial, political, and social development (these thesaurus terms account for almost 30% of the terms). The description of historical activities and events accounts for almost 12% of the terms. It is also interesting that such a high percentage of GMGPC terms (18.6%) is incorporated as subject terms; one could speculate that the number of pictorial materials depicting “Callotypes” or “Collodion dry plate photonegatives” as subjects would be rather small, except in a highly specialized collection dealing with the history of such processes.

<table>
<thead>
<tr>
<th>Subject Headings</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business/Industry</td>
<td>20.9%</td>
</tr>
<tr>
<td>Objects</td>
<td>19.1%</td>
</tr>
<tr>
<td>GMGPC</td>
<td>18.6%</td>
</tr>
<tr>
<td>Built Environment</td>
<td>5.3%</td>
</tr>
<tr>
<td>Activity</td>
<td>4.5%</td>
</tr>
<tr>
<td>Government</td>
<td>4.5%</td>
</tr>
<tr>
<td>People</td>
<td>4.1%</td>
</tr>
<tr>
<td>Abstract</td>
<td>3.5%</td>
</tr>
<tr>
<td>Event</td>
<td>3.5%</td>
</tr>
<tr>
<td>Health</td>
<td>3.5%</td>
</tr>
<tr>
<td>Society</td>
<td>2.8%</td>
</tr>
<tr>
<td>Theme</td>
<td>2.5%</td>
</tr>
<tr>
<td>Setting</td>
<td>2.2%</td>
</tr>
<tr>
<td>Natural History</td>
<td>1.7%</td>
</tr>
<tr>
<td>People-Related</td>
<td>1.5%</td>
</tr>
<tr>
<td>Organization</td>
<td>0.5%</td>
</tr>
<tr>
<td>Visual Component</td>
<td>0.2%</td>
</tr>
<tr>
<td>State</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

Table 2: Percentage distribution of subject headings from a sample of the LCTGM.

The LCGTM does include a high percentage of OBJECTS terms (19.1%), but contains no terms for location specification, and very few terms relating to VISUAL ELEMENT or ART HISTORICAL INFORMATION. It contains terms relating to abstract concepts and some thematic and “story” elements, but its high degree of precoordination creates a very narrow application of many of these terms. It serves its purpose of providing historical access to pictorial materials, especially those concerned with the history of the United States, but lacks utility for more general applicability to a wider variety of pictorial materials, in which other classes of attributes may be important.

The *Thesaurus Iconographique* (TI) explains its conceptual approach in detail. It takes the term “representation” in a large sense, treating the image as a document of general interest. Its primary goal is to serve a variety of research interests by completely representing the content.
of an image. The TI is composed of four parts, called rubrics, which are further subdivided into classes containing lists of hierarchical and non-hierarchical descriptors. The four rubrics are: general description, proper names, the precise text that inspired the representation, and the date of the contents of the representation, not its execution. The first rubric treats "the essential content" of the representation and is required for all descriptions.

The descriptors are divided into two fundamental groups, under the terms "Themes" and "Subjects." Themes describe elements and relations of general significance concerning human life and individual behavior or the deeds of society. The subject descriptors assist with situating pictorial elements historically and geographically.

Gamier (1984) is one of a very few published systems to deal with stylistic information such as the presence of a specific item in a picture, the nature of the image in relation to others, and the dominance of a particular technique such as relief, color, or perspective. There are two interesting aspects to this descriptive system in relationship to the current research: the incorporation of many different types of attribute information, and the emphasis on such aspects as meaning, themes, and "stories." The thesaurus section covering "General characteristics of the representation" provides terminology for describing such specific attribute information as the type of representation, the view, the type of figure, the position of the figure, the angle of view of a figure, the material situation and orientation, and "formal effects" such as line, form (color, surface, volume), effect of light, reflection, and superposition. Also under general characteristics are such factors as thematic relations (theme, narrative, symbol), "finality" (comic, instructive, moralistic, satiric), manner (antique, oriental), and "truthfulness" (anachronism, exaggeration, fantasy, supernatural).

An analysis of a ten percent sample of 463 terms from the TI shows the distribution of terms among the different classes of rubrics. Table 3 shows a distribution of terms with a somewhat different emphasis than the LCTGM. Nature, the life of man, and Biblical subjects compose the most frequently occurring terms in the TI, followed by social and religious life. The TI provides terminology for many of the image attributes revealed as being typically noted in the empirical research. For instance, it includes terms for many OBJECTS attributes, such as Clothing, Body Part, and numerous specific categories of objects such as cooking utensils under the various Thèmes. It contains terms for Abstract concepts and Emotion in such classes as Vie psychologique et morale. It contains many Theme and Event terms in both Thèmes and Sujets. Activity attributes are listed in Société et vie sociale. It provides extensive Symbol terms as well. For instance, an image with a frontal view of the head of a cow and a man's face side by side carries in addition to the physical description the terms "timidity," "baseness," and "laziness" (Gamier, 1984, 74). Many VISUAL ELEMENT attributes are covered in Effet formel.

There are some areas that are not covered, however, such as several of the ART HISTORICAL attributes, COLOR terms, Number, and LOCATION terms. Angle of view may be described, as well as position of humans or animals or the actual location of the representation within a larger work, but no descriptive location terms for items within the interior of an image are given.

The Dewey Decimal Classification (DDC), the most widely used library classification system in the world, is composed of basic classes organized by traditional academic disciplines or fields of study. As the introduction explains, the consequence of this is that there is likely to be no one single place for a given subject. Thus, a subject such as marriage has aspects relating to customs, law, religion, and other areas, and will be found under all these topics. Any subject which has been the focus of academic study should be represented.
Subject Headings %

<table>
<thead>
<tr>
<th>Subject Headings</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature</td>
<td>18.4%</td>
</tr>
<tr>
<td>The Body And Material Life</td>
<td>17.9%</td>
</tr>
<tr>
<td>Biblical Subjects</td>
<td>10.8%</td>
</tr>
<tr>
<td>Society And Life In Society</td>
<td>7.3%</td>
</tr>
<tr>
<td>Religious Life</td>
<td>7.1%</td>
</tr>
<tr>
<td>Arts And Performing Arts</td>
<td>6.0%</td>
</tr>
<tr>
<td>Energy/Industry/Trades/Business</td>
<td>5.2%</td>
</tr>
<tr>
<td>Transportation/Communication</td>
<td>4.5%</td>
</tr>
<tr>
<td>General Characteristics Of Images</td>
<td>4.1%</td>
</tr>
<tr>
<td>Political And Administrative Life</td>
<td>3.7%</td>
</tr>
<tr>
<td>Agriculture/Hunting/Fishing</td>
<td>3.2%</td>
</tr>
<tr>
<td>Intellectual/Scientific Life</td>
<td>3.2%</td>
</tr>
<tr>
<td>Psychological And Moral Life</td>
<td>3.0%</td>
</tr>
<tr>
<td>Imaginary Beings</td>
<td>2.6%</td>
</tr>
<tr>
<td>Armaments/Military</td>
<td>1.5%</td>
</tr>
<tr>
<td>Geography</td>
<td>0.4%</td>
</tr>
<tr>
<td>Historic Persons</td>
<td>0.4%</td>
</tr>
<tr>
<td>Names Of Groups</td>
<td>0.4%</td>
</tr>
<tr>
<td>Belief System</td>
<td>0.0%</td>
</tr>
<tr>
<td>Historical Period</td>
<td>0.0%</td>
</tr>
<tr>
<td>Imaginary Persons</td>
<td>0.0%</td>
</tr>
<tr>
<td>Mythology</td>
<td>0.0%</td>
</tr>
<tr>
<td>Ornamental Terms</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Table 3: Distribution of sample of headings from *Thesaurus Iconographique*

within the classification. While subjects are multifaceted, the notation representing a specific topic follows a hierarchical principle. The goal is to assign one number to each item, yet this number may be composed of smaller segments which represent more specific aspects of a subject, such as geographic area, period, or type of person or group.

A slightly different approach was taken for the analysis of the DDC. As the classification attempts to capture "all knowledge," it is much larger than the two previous classification systems concerned with image content. Analyzing the contents of the Relative Index, in which terms representing "subjects" are arranged alphabetically, did not seem to be particularly useful, as almost any of the terms in the relative index could potentially represent the subject matter of an image. A subject, according to the introduction, is the object of study and may be a "person, group of persons, a thing, a place, process, activity, abstraction or any combination of these," as opposed to a "discipline" an organized field of study or branch of learning under which subjects are subsumed, (Comaromi, 1989, lvi, liii). Therefore, each attribute as defined in the current research was evaluated for the extent of its representation in the DDC by checking the Schedules, Tables, and Relative Index for sections which could usefully represent the attribute.

The DDC fares quite well in its breadth of representation. The only class which was not represented in the DDC was LOCATION. However, the previous research indicated that the applicability of terms from this class may not be directly useful, and a different mechanism may be needed to adequately represent the conceptual basis of LOCATION terms. While a subject term for COLOR exists, there are no specific color terms and other color aspects such as Color Value are absent. Other attributes (and their classes) which could not be represented were Atmosphere and State (ABSTRACT), Number (DESCRIPTION), and Time Aspect (CONTENT/STORY). Table 4 contains a listing of attributes and a sample of Dewey Decimal Numbers which could be
used to represent these attributes.

<table>
<thead>
<tr>
<th>ATTRIBUTE</th>
<th>DDC</th>
<th>ATTRIBUTE</th>
<th>DDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>753-755</td>
<td>Object</td>
<td>many</td>
</tr>
<tr>
<td>Activity</td>
<td>790-799</td>
<td>Orientation</td>
<td>701, 741</td>
</tr>
<tr>
<td>Artist</td>
<td>PN</td>
<td>People</td>
<td>704</td>
</tr>
<tr>
<td>Atmosphere</td>
<td>611, 612</td>
<td>Perspective</td>
<td>70, 741</td>
</tr>
<tr>
<td>Body Part</td>
<td></td>
<td>Reference</td>
<td>PN</td>
</tr>
<tr>
<td>Category</td>
<td>750s</td>
<td>Relation</td>
<td>T1, 08, 305</td>
</tr>
<tr>
<td>Clothing</td>
<td>391</td>
<td>Representation</td>
<td>700s</td>
</tr>
<tr>
<td>Color</td>
<td>741, 747</td>
<td>Setting</td>
<td>710-745</td>
</tr>
<tr>
<td>Color Value</td>
<td></td>
<td>Shape</td>
<td>516</td>
</tr>
<tr>
<td>Composition</td>
<td>701</td>
<td>Social Status</td>
<td>T1, 08</td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td>State</td>
<td></td>
</tr>
<tr>
<td>Emotion</td>
<td>150s, 610s</td>
<td>Symbolic</td>
<td>709</td>
</tr>
<tr>
<td>Event</td>
<td>790-799</td>
<td>Style</td>
<td>704</td>
</tr>
<tr>
<td>Focal Point</td>
<td>701, 741</td>
<td>Time Aspect</td>
<td>702</td>
</tr>
<tr>
<td>Format</td>
<td>741</td>
<td>Technique</td>
<td></td>
</tr>
<tr>
<td>Location - General</td>
<td></td>
<td>Texture</td>
<td>620, 701</td>
</tr>
<tr>
<td>Location - Specific</td>
<td></td>
<td>Theme</td>
<td>many</td>
</tr>
<tr>
<td>Medium</td>
<td>741</td>
<td>Time Reference</td>
<td>709</td>
</tr>
<tr>
<td>Motion</td>
<td>701, 741</td>
<td>Type</td>
<td>704</td>
</tr>
<tr>
<td>Number</td>
<td></td>
<td>Text</td>
<td>745.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Visual Component</td>
<td>701, 741</td>
</tr>
</tbody>
</table>

Table 4: Attributes and selected Dewey Decimal Numbers which could be used to index these attributes (PN denotes proper noun)

As a general classification system for knowledge, the DDC has both breadth and depth. If the notational technique limiting application of one number only to an item is suspended, and a technique more similar to that of the TI is employed, then a faceted type of image classification could be built. The appeal of using an existing general system such as the DDC is that one search could bring up both textual and visual materials on the same topic.

4. Choice of Image Attributes to Index

Indexing of images is considered difficult precisely because there are so many possible attributes which could be indexed. Thus there are two important parts to any system of coding or indexing images. The first is a list of terms that is sufficient for describing the full range of attributes present in the image. However, just as important as this is a set of heuristics guiding the application of the coding or indexing system to a specific image. When there are many attributes present, it is important to have a framework that can guide the analysis and provide a foundation for choices concerning which attributes should have priority in indexing and to what depth this analysis should take place.

The introduction to the LCTGM (Parker, 1987, xiv-xv) discusses some of the multiple aspects of pictures which catalogers should consider when trying to decide which subjects to index, such as historical significance and novel aspects of the subject matter, whether the subject is better represented elsewhere, a particular point of view or message necessitating indexing of
context, and image content which is unique primary evidence of a particular time and place. In addition, indexing should indicate both the “Of” and “About” of images (e.g., “Richard Burton” and “Actors”), thus capturing both specific and generic levels of access.

Similarly, the TI proposes a logical order of description, first addressing the material disposition of the scenes, then chronological progression, then meanings in order of decreasing importance. While the TI requires a basic descriptive level of image content, example descriptions provided in the TI demonstrate the preponderance of terms describing meanings, symbols, and abstract concepts.

Fiction indexing research is another area that can be drawn upon for insight into providing access to more abstract, emotive, or interpretive elements of images. Those wishing to find fiction face many of the same problems that image searchers face: common retrieval points such as author and title often reveal little content information. Genre arrangements have proven useful for access to fiction (Baker, 1988; Peitersen, 1989), and Small (1991) states that access to genre scenes such as banqueting would be useful.

Beghtol (1991) notes that certain domains, such as fiction, may not be amenable to traditional classification approaches which emphasize mutually exclusive categories and exhaustivity. She describes some of the ways in which fiction varies from non-fiction. First, fiction may not be “about” a “subject” in the usual sense; a novel whose story takes place in a chemistry laboratory does not deal with chemistry as a “subject.” Second, fiction is not restricted to the “possible or the “real”; it can contain anything the imagination can invent and these imaginary things are considered “true” in the fictional world. Third, fiction authors do not eschew ambiguity in the same way authors of non-fiction are assumed to do; the details and outcome of a story may be deliberately unresolved by the author and unresolvable by the reader (Beghtol, 1991, 3).

These qualities are all qualities that pictorial materials share with fiction. Beghtol suggests that rather than using a summarization approach to fiction, it may be more fruitful to extract various kinds of data elements and to classify individual instances of these elements. She notes that characters, events, spaces, and times may be taken as fundamental elements in fiction; the current research (Jörgensen, 1995) indicates that these same elements are some of the more typically perceived attributes in pictorial images as well. Fiction and pictorial images also share the need to specify imaginary, non-existent, or impossible things, making controlled vocabulary difficult to establish. Some of the problems of ambiguity which Beghtol notes could be solved in both fiction and image indexing by use of the specific and generic levels of indexing, as discussed earlier. Researchers in image indexing could thus usefully draw upon some of the concepts put forth by those concerned with indexing fiction.

5. Conclusion and Future Research

This paper has described the components of three indexing and classification systems as they relate to recent research in image describing and sorting tasks. An analytical approach such as this should provide useful information for image indexers considering which system or systems to use as sources for image indexing terms. While none of the systems discussed in this paper can serve as a complete indexing tool for images, each can be usefully applied in relation to specific image attributes. Major attribute classes as defined in recent research suggest that a wide range of attributes need to be addressed in image indexing. Future research could draw upon a variety of existing indexing and classification tools and define which are most appropriate as sources for the various classes of image attributes. The comparison between image attributes revealed in recent empirical research and those represented in the Dewey Decimal Classification presents some
intriguing possibilities for future research into its potential contributions to the indexing of images and the linking of textual and visual materials.

References
Harold C. Steyer, Jr., Ana Flavia Fonseca, Diane D. Hopkins, Marc Nodell, Irene L. Travis, and William S. Wahl
World Bank

The World Bank's Information Management Architecture: A Blueprint for Building Institutional Information Services

Abstract: To facilitate information access and sharing, the World Bank began developing a Bankwide Enterprise Network in the early 1990s. The network, which was fully deployed at the Bank's headquarters campus and some field offices at the end of 1995, enables Bankwide services, such as electronic document management, directory services, and Internet access. As the Bank began development of these services, it became clear that it required a comprehensive model for its future information management infrastructure. The model was necessary to provide infrastructure that could be built in an orderly and logical way—one that would be manageable, modular, scaleable, and open. This paper discusses this model, developed by an interdisciplinary team, and its application to the Bank's evolving electronic document management system.

1. Introduction

The World Bank is an economic development organization owned by 178 member countries. It has about 9,100 staff in Washington, D.C., and more than 1,200 staff in 76 field offices world-wide. It is a significant example of a knowledge-based organization, and its major strategic asset is the staff's collective knowledge of economic development, acquired through 50 years of experience.

To facilitate information access and sharing, the Bank began to develop a Bankwide Enterprise Network in the early 1990s. The network, which was fully deployed at the Bank's headquarters campus and some field offices by the end of 1995, enables Bankwide services, such as electronic document management, directory services, and Internet access.

As the Bank began development of these services, it became clear that it required a comprehensive model for its future information management infrastructure. The model was necessary to provide infrastructure that could be built in an orderly and logical way—one that would be manageable, modular, scaleable, and open. Without this understanding, the Bank would face a chaotic environment of duplicative and incompatible efforts. Future development of information systems in support of Bankwide goals and business processes would be inefficient or even impossible.

2. Problems Addressed

It is helpful in understanding the project to look at a sample of the kinds of questions the Bank wanted to address. They include very basic questions like the following:

- How do we facilitate broad subject searches by staff for internal documents, external documents, experts, organizations, etc. on a topic of interest?
- How do we integrate all media in our planning, including structured databases, text, video, sound, multi-media?

• How can we guarantee the reliability and authenticity of an electronic document after it has been used in a business activity? For example, the cover memo, header or routing slip, any attachments, and subsequent comments must be permanently linked and unalterable.
• How are we going to allow stores of information to be accessed by many different applications?
• How are we going to support keeping distribution lists up to date?
• How are we going to provide access control when access in the Bank is based on position and function in the organization, not personal clearances?
• How are we going to ensure that our e-mail system will interface with our directories and our document management systems?
• When certain records are due for destruction, how can they be automatically identified and destroyed, leaving an appropriate record of their destruction? If non-record documents are stored in the system, how can they be weeded out?
• How are we going to be able to move documents physically without having to update their location information in multiple access systems?
• How can we ensure that collaborative work group support applications will be able to deposit work products into institutional stores, as well as retrieve information from them?
• How are we going to maintain information stores over time as technology changes?
• How do we keep the modules of the system independent and updatable?

Such questions have little to do with the subject organization of knowledge, but they are nonetheless critical to knowledge organization and management in a business context (Hopkins et al., 1995; Bearman, 1993).

3. The Development of the Information Management Architecture (IMA)

In the summer of 1993 the (then) Information, Technology, and Facilities Department brought together an interdisciplinary group of specialists in database management systems, search and retrieval systems, data administration, libraries, records management, archives, and electronic mail1 to address these questions. To handle the scope and complexity of the undertaking, the team employed well-known information engineering methodologies to link business information requirements to information services and to derive a set of implementation projects. The result of these efforts was the World Bank Information Management Architecture (IMA) (World Bank, Information, Technology and Facilities Dept., 1993), which has been guiding the development of the Bank’s Institutional Information Services for the past two years.

The biggest initial challenge for the interdisciplinary team was a common conceptual framework represented by a common vocabulary. In different information professions and contexts, terms such as "record," "document," "directory," "file," and "classification" have variant meanings that the group had to share and resolve before work could progress. The report contains a glossary of more than 70 terms that were defined in the context of this project. For example, one of the most important concepts adopted by the team—"document"—reflects the project’s organizational focus. The definition evolved to be "any identifiable unit of information that must be managed as one entity for use in a business process." A "document" can thus be, among other things, a data base record, a spreadsheet table, a videotaped presentation or a combination of any such objects to form a "compound document."

The team also used entity-relationship analysis and functional models to analyze the information management process into its key components. The components are listed and discussed below. The joint perspective developed through this exercise has been critical to the
success, not only of the IMA effort, but also to the projects based upon it.

4. An Overview of the IMA

There are two major components of the IMA: Information packages and Institutional Information Services—that is, roughly, the things being managed and the services to manage them. Under the model, documents, as defined above, are only the simplest unit of management. Information management in an organizational context, the team concluded, should be based on the "information package."

Information packages have four elements:
- A collection of documents;
- The directories that provide the information necessary to locate the documents;
- The semantics that provide definition, format, and display information about the contents of the documents; and
- The administration rules that govern the management processes, security, retention and archival integrity of documents in the collection.

For example, an information package might be:
- A collection of reports;
- The indexes to the report collection;
- The tagging that has been applied to define their contents and structure; and
- The rules regarding such things as who can update them, how long they are to be retained, and to whom they can be distributed.

These elements together provide the information necessary to manage any information resource.

In addition to the information packages, the other major components of the IMA are the six Institutional Information Services. Figure 1 shows the six services and their relationship to other system elements, such as standard human interfaces and business process support systems. These systems will include document management, correspondence tracking, desktop mail, database access, staff and expertise locators, and decision support, among others.

There is a high degree of integration and interdependence among the services. However, standard applications program interfaces (APIs) insulate the user interface and other business information applications from the underlying workings of the six services and any changes taking place in the technology that provides them. A brief description of the services, will be followed by an example of how information packages and services are utilized in building a document management application.

5. The Institutional Information Services

The six Institutional Information Services are as follows:
- Enterprise Directories;
- Information Creation and Storage;
- Information Retrieval;
- Information Administration;
- Messaging and Mail; and
- Business Service Requests.

If one service were to be selected as the core of this cluster, it would be the Enterprise Directory Service. The Directory Service is based on the X.500 standard. (See Chadwick,
Fig. 1: Six Institutional Information Services Will Provide the IMA Functions.

1994). Originally developed to support electronic mail, it can also be used for many other applications. The Bank has already migrated the "white" and "yellow" pages of its telephone directory to X.500 and is planning other applications, such as a database directory. Eventually, an X.500 directory will hold the location of all Bank documents and will be the single point at which such locations are updated.

The Information Creation and Storage Service, under development, will provide a standard method for electronic document management that enables individual staff, work groups, and business units to:
Manage the production of multi-part documents throughout the editorial and clearance phases;

- File work-in-progress and completed documents in a way that is compatible with paper-based files, and
- Meet their local records management obligations.

A staff member should be able to use any Bank standard word processor, spreadsheet, or other application to create, revise, and store electronic documents, from email messages to complex reports with multiple attachments. Control information provided by the document profile (description) will ensure the integrity and long-term availability of documents in electronic formats and support the retention and disposition functions.

The Information Retrieval Service is being designed to be completely independent of the Information Storage and Creation Service. Retrieval packages used will leave the documents in their native formats and use relative locations in their indexes. These features are widely available in commercial packages. However, the Bank is also working with vendors to accommodate taking document location information from the X.500 directory as a single source. It appears that this arrangement will be technically feasible. As shown in Figure 1, access to the service will also be independent of any particular interface. Users can then access many stores using one search interface, such as a Web browser. Alternatively, they can search through applications, such as groupware, that use application program interfaces (API’s) to integrate search engines into the service they provide.

The Information Administration Service provides facilities and procedures for information administration in order to maintain and make available the semantics and administrative rules. These rules govern collection maintenance processes, such as retention and disposition, but, more importantly, will allow business process support systems to access information from different sources. For structured data, the Service will concentrate on standardizing and supplying metadata for the various databases that must be shared.

The Messaging and Mail Service will provide the means to integrate messaging with other office and work processes, using the X.400 standard for message transfer.

The most important of the Business Service Request systems from the perspective of this paper is the Authorization Server. The server accepts and validates requests for access to documents, profiles or other objects, using a rule-based approach. The Authorization Server draws basic information from the Bank’s Staff Locator System and Organization Charts, which are accessible through the X.500 Directory Service. However, it provides much greater flexibility than can be provided through the staff lists alone. For instance, authorized managers or team leaders, whether of formal or informal groups, can specify the members of the work group. Those members can be staff under the managers’ administrative control, matrixed staff, or even short term consultants. Access can also be provided on the basis of organizational position. The server also supports rules for classes of objects, such as documents. An object class manager maintains the access rules for a particular object class. For instance, the Archivist is the object class manager for documents. This arrangement is both much more flexible and much more easily maintained than access control lists.

The server is implemented in a relational database management system (RDBMS) because, among other reasons, it provides referential integrity. The RDBMS can, in turn, feed work group information to the X.500 staff directories, which lack that feature. The output supplied and validated by the security server can greatly enhance staff access through the X.500 directory to information about current work groups and who is participating in them. Such information will provide a much-needed supplement to the more standard data available from organization charts or personnel records.
6. Information Packages

The four elements of information packages, as defined above, relay most heavily on the Creation and Storage Service and the Information Administration Service. The Creation and Storage Service maintains the documents themselves and their profiles. The Information Administration Service maintains administrative information and will also house such semantic definitions as standard document templates. Other services are used indirectly, as illustrated in Figure 1.

7. Applying the IMA

Over the past two years there have been a number of major integration and database construction efforts that directly relate to the IMA architecture. In this section we will look at how they are being brought together in a new business process support system, an image scanning station, which is scheduled to be available in for testing April 1996. As shown in Figure 2, the four infrastructure services are:

- A relational database that stores document profiles (descriptions) (Information Storage and Creation Service),
- A metadata manager tool that drives the profiling (Information Administration Service)
- The authorization server (Business Service Request Service), and
- A commercial storage and retrieval package already in use in the Bank that is primarily an image system, but also provides full-text search capabilities for imaged documents (Information Retrieval Service).

The new station will provide a number of enhancements to the commercial retrieval system, but the services that facilitate the enhancements are not within the system and are available to any other business support application that chooses to use them. There will be a common retrieval interface via a Web browser for profile-based searching and copy-and-paste from OCR versions of retrieved images.

The commercial product already allows images as well as ascii text be stored on the network, and provides a search engine that users have liked quite well. However, it has a number of short-comings. For instance, like many image systems, it provides only very limited profiling capabilities. In particular it has not been possible to tailor profiles in any intelligent way or to provide default values or pick lists for fields. These limitations mean that very little control information about documents can be collected without the process becoming an unacceptable burden on the users. It also does not provide document level security.

The new station will remedy these problems. The profile database, as part of the Storage and Creation Service, will store the profile data in such a way that it will be accessible to both the commercial system and others. However, perhaps even more importantly, the Metadata Manager in the Information Administration Service will allow information administrators to customize input screens according to the business context that first created or received the document. The rules in the Metadata Manager can control which attributes are presented, as well as the relevant default values and pick lists. For example, a staff member scanning a piece of correspondence will not be asked for a report number.

Document types, like memorandum and Staff Appraisal Report, are important to efficient records management and retention. To date the Bank has identified more than 400. Approximately 100 types are generic and thus may be used in the context of more than one business process (e.g., letter, memorandum, minutes). Other specific types are prescribed for individual business processes, like lending operations. The Document Types database is
maintained through the Metadata Manager. It supplies selected default values to minimize data entry when users are creating profiles. Individual staff need only become familiar with the specific types relevant to their business activities, plus a few generic types that repeat across profile type picklists.

While the scan station is an immediate product, the Bank is continuing integration of a more comprehensive document management system. Also built on elements of the IMA, it will cover documents as they are being created, as well as final copy.
8. Summary and Conclusions

The Bank is seeking to create information services that will provide integrated access for users, while being buildable, scaleable, and maintainable. In order to accomplish this goal, the Bank developed a general architecture specifying the various services associated with information systems—creation, storage, retrieval, management, dissemination, access. It also detailed how these services would relate to each other and the projects that would be needed to construct them. After two and a half years, this scheme is still providing a good foundation, and the services, as planned, are coming together to support the scan station and other business applications.

Notes

1. Members of the IMA team were: Harold Steyer (Team Leader); Ana Flavia Fonseca, Peter Gutterman, Ali Hashim, Diane Hopkins, Harold Islev-Petersen, Marc Nodell, Irene Travis, and William Wahl.

References


Abstract: Access to information—and classification techniques and methodologies that support that access—are already playing important roles in making information accessible to employees in the business environment. But most businesses are only beginning to solve the higher level requirements of enabling action, which is the domain of knowledge. The traditional document-oriented model itself is a stumbling block, because it focuses on the document as a large, inert information artifact, ignoring the ultimate business objectives of getting things done and generating competitive advantage. New models for organizational knowledge resources are needed, and classification approaches will still play a vital role, but such approaches must be highly adaptable, they must be formalized to accommodate technological implementations, and they must embody meta-principles for self-organizing knowledge resources in business environments.

1. Breaking Down the Boundaries of Technologies and Disciplines

The charter of the International Society for Knowledge Organization (ISKO) states its desire to function as "the connecting link between any and all institutions concerned with conceptual questions pertaining to the organization and processing of knowledge." That description resonated more than a little bit with us at Knowledge Transfer International, "the Knowledge Management Company," and it reminded us once again that the solutions we are seeking and applying for our general business audience come from unexpected quarters.

We keep being surprised at the expanding range of disciplines and technologies concerned with "knowledge": brainstorming software, hypertext, expert systems, electronic publishing, help desks and help systems, groupware, information and library science, object-oriented information modeling, information interchange standards, semantic networks, technical writing, and a variety of methods for classifying the content of bodies of information, including thesauri, back-of-the book indexing, and computer-processable AI ontologies.

However, we should not be surprised, because we know that knowledge itself has become organizational infrastructure in business, superseding line-of-command organization and even trendy "process re-engineering" management strategies as the persistent glue of business organizations. Knowledge is the only constant, even though knowledge itself is fluid.

2. Information Science—Access to Information Resources

The requirements of the business community do differ from those of libraries. The differences are not hard and fast, but they are still significant. Our perception of information and library science is that it is fundamentally about access to information. And the lessons of information science are already highly relevant to every Fortune 1000 company, because all of them are concerned with information as a corporate asset.

That may be most evident in the popularity of document management systems for
networked environments. Such systems make direct use of classification in the form of "document profiles." And every good help system for software applications uses keyword search. So the need for effective methods of mapping access to information could not be greater, and no one is better qualified to create those methods than information scientists.

But "information" is not "knowledge." Most of us in the traditional documentation and training community perceive the role of information science as ending when information seekers find what they are looking for. We think of ourselves as standing where words and images are converted into individual and organizational performance—where information becomes knowledge and is turned into action.

Information retrieval expert Elise Yoder summed up her recent study of World Wide Web search tools with the following observation: "... it is apparent that 'finding' and 'reading' are just two activities among a spectrum of activities that include identifying promising resources, assessing the value of things that have been found, reading for understanding, and using the information to do work" (Yoder, 1995).

Successful access and reading must be followed by understanding and integration of that knowledge into personal experience. That knowledge enables action. Our experiences applying that knowledge refine it further, synthesizing it into new knowledge, which in turn must be made accessible to others, completing the upward evolutionary spiral of organizational knowledge development, management, and transfer.

3. How Is the Business Environment Different?

The growth and incredible pace of change of technology mean that everything we deal with in our business and personal lives is evolving toward greater complexity and toward greater information content. Nevertheless, the business environment's needs are different of the typical community and university library environment—if not in purely qualitative terms, then at least in quantitative terms.

- Businesses continuously create their own new information products. If you are in business, you are in the business of creating new knowledge and creating information products, for both internal and external consumption.
- Increasingly, the intellectual holdings live in a virtual, shared space. Although the holdings of libraries increasingly take the form of electronic resources, companies are aggressively converting all of their information resources into centralized electronic repositories.
- The need to adapt to change is much higher. Bad or dated information itself must be rooted out and systematically killed, and classification systems themselves must change rapidly. For example, the category of Web-based document management systems did not exist a year ago, and new category represents a radical change in the distinctions among retrieval systems, document management systems, and other information-based World Wide Web technologies.
- Size of holdings is not always an advantage to the information seeker. Instead, it is often a disadvantage. The needs of the information seeker are immediate. Speed of access to solutions, not multiple documents, is vital. The corporate employee is not engaged in textual exegesis.
- The information needs are critical. Corporations are engaged in economic battle. Redundant, conflicting, or dated information must be reduced to "unique" kernels of true, relevant information. Lives and safety depend on it, too, in many cases.
4. How Do These Differences Affect the Organization of Knowledge in Business?

The business requirement is for knowledge, not just access to information. It would be
hard to verify that assertion if you take a look at most businesses, which are still counting
output of documents (or conversion of documents to electronic format) as a method of
measuring their “success” in organizing corporate knowledge resources. Yes, access is very
important, but it’s only part of the requirement, not all of it, because action does not take place
until information is converted into knowledge and performance.

The technology that is causing the problem—rampant, pervasive computerization and now
almost universal networking—will provide some of the important tools that allows us to break
the persistent constraints of the paper-based mindset. But we must do more new thinking
before implementation and, in particular, we must take full advantage of the experience and
solutions evolving in organizations like ISKO.

5. The Information Model Is Changing from "Document" to "Dialog"

We are already seeing a shift from the broadcast model of publishing—sending large,
comprehensive information objects ("documents") to those who request them—to allowing
information seekers to dial in and ferret out the specific information they need. The networked
computing environment makes this change possible, even encourages the transition to
information seeking as a process of dialog. The challenge facing the designers of information
repositories shifts from providing "all" answers to (1) shaping resources to be more effective
in providing knowledge and (2) telling information seekers how to ask the questions in a way
that provides them with the answers they need most expeditiously and reliably.

[This second change parallels the general business trend of shifting costs and burdens to
suppliers and consumers. The now omnipresent automated bank teller machines provide a
good example. ATMs are tireless and work all night, but they put the burden of specifying the
transaction on the bank customer instead of on a human teller.]

The business world’s current model of knowledge transfer still rests on books and
articles. But even in their snazzy electronic forms, books and articles are almost all still print-
based at heart. (Most current implementations of “hypertext” do little to change the model.)
The granularity of the resource is simply too large, and the burden of extracting knowledge is
placed almost exclusively on the knowledge seeker.

Deconstruction and formal modeling of the resource are the first part of a general
solution that will allow us to computerize the gathering and processing of knowledge resources,
enhance their effectiveness, and leverage their value. It is not simply a matter of breaking
books and articles down into smaller chunks—the approach taken by increasingly popular
document management systems and by many hypertext publications. Knowledge resources
must be “reverse engineered”—in effect inverting the processes we perform in assembling
traditional documents—in order to identify the component elements, relationships, and even
“behaviors” of those resources, accommodating the interaction between knowledge seeker and
a body of knowledge generated by subject-matter experts.

Information must also be modeled for more effective communication. "Irrelevant" detail
must be suppressed—or at least suppressible—so that we can read less and grasp important
information quickly and easily. But detail must remain accessible, because one person’s
“irrelevant details” are another’s critical kernels of information. Meaning and relationships of
meaning must be made explicit.
6. The Importance of Classification

Classification is the second major part of the solution. But classification must be driven down to the more granular level of document components—adding metadata to small document elements and even to the relationships among those elements.

When we reach this level of classification of discrete elements, we reap the benefits of distributed development, precise specification, controlled change, and flexible re-use, because explicit classification schemes help make information processable by computers as well as accessible to humans.

A second benefit of classification is that it can make meaning more explicit and precise, providing greater understanding, not just better access and processability. The linking of formal ontologies to deconstructed information elements may serve as a form of—or a complement to—a system of classification. Together, they may allow computers to extend our intellectual reach by supporting precision, mastering complexity, and providing unfailing memory for millions of facts and relationships.

7. Distribution of Responsibility for Development and Optimization

The traditional assumption of technical communication is that effective documents and training systems are assembled carefully and pro-actively from research by an individual or small group, and designed for maximum impact on a well-defined target audience.

But given the pace of change, the growing need for effective information, and the demand for timeliness and customization, it is not possible for technical communicators—or anyone else in the typical business organization—to pro-actively build such resources for each newly identified requirement. So the burden of building and managing knowledge resources has to be shifted to everyone who contributes to or uses those resources.

Deconstruction and classification of knowledge resources will make it possible to shift the burden of building (and contributing to the continuous evolution of) network-based knowledge resources from "librarians" and "technical writers" to other members of the organization—and even to customers. Thus the construction of the knowledge resource can become a distributed, manageable task, much as entry of data into fields in a relational database can be distributed among thousands of people if everyone understands and adheres to a consistent set of rules and guiding principles.

How are these evolving requirements for organizing knowledge in business reflected in today's practices?

8. Case Study #1: So Who Needs Classification???

In 1994, Morgan Stanley Trust Company (MSTC) found that its eleven globally dispersed functional groups lacked a complete set of operating policies and procedures. Many employees found themselves relying on fragmented and occasionally anecdotal information to do their jobs.

A prior corporate-level decision to create a paperless "electronic office" for propagating information efficiently and quickly had already put TCP/IP stacks in place, so MSTC chose World Wide Web architecture as logical mechanism for an organizational information resource.

Knowledge Transfer International (KTI), publishers of KM Briefs and KM Metazine, recommended developing an "intranet" that would serve as a "corporate World Wide Web" and provide internal e-mail, a bulletin board, and other functionality.

Policies and procedures at MSTC are inherently hierarchical—they encompass a series
of tasks and sub-tasks at the business unit, functional, and individual levels. Many are built around transaction processing systems, so ensuring that their presentation matches the hierarchy embedded in the work model is critical. Problems arise if a procedure isn't carried out according to the work flow embodied in a transaction system.

So how should MSTC's online information be classified for retrieval? The answer, at this time, is "Don't bother." The material is connected with structural hypertext links that directly reflect the work flow within the organization. In a sense, every document must be where it is because MSTC employees have well-defined job roles and correspondingly well-defined procedures. So who needs classification to make good use of the online documents? Certainly not those who use the system.

9. Case Study #2: Classification is Part of a $5 Million Difference


At the end of 1993, Digital had over 300 discrete internal databases of information about their products and services—helpful to their sales personnel and business partners, but not nearly as helpful as they might have been, because information isolated in business units prevented sales personnel themselves from presenting the full range of solutions offered by Digital. And at times, the product information in separate databases was contradictory, creating confusion among sales personnel and customers.

The solution was to create a central repository that replaced the separate databases. Of course, when you combine information once accessed by where it was found into one large, centralized database, you need to replace the organizational segmentation with enterprise-wide methods of identifying and extracting that information.

In a six-month span, the Information Repository group led by Deborah Bennett created a "common standard language" of Digital's products and services and built a library of keywords and synonyms—a rich classification system that reflected the logical organization of their promotional literature architecture. A team of 12 indexers schooled in information and library sciences had primary responsibility for building the library and assigning keywords to documents. During this period the group also evaluated the development efforts and audited the activities of employees using the new service.

What were the results? Even the development team was surprised by how rapidly Digital personnel took to using the system and by early demand for access to the system from Digital business partners. It may be hard to quantify the business benefits that rapid access and greater control over accuracy may bring, but Digital realized a direct savings of over $5 million dollars in dissolving the 300 separate databases.

10. Case Study #3: Classification in a Web Publishing Application

An electronic edition of the Communications of the ACM of August 1995—a special issue on hypermedia—is now available at http://space.njit.edu:5080/acm/overview.html. This effort under the direction of Michael Bieber at the New Jersey Institute of Technology employs two forms of classification: (1) metadata associated with each Web page and (2) semantic labeling of the hypertext links themselves—that is, you can view a meaningful description of the target of each hypertext link, and thereby decide whether you will actually traverse the link.

This Web publication is not a test bed for interface design, but it is a test of what value
these forms of explicit classification add to an online publication. The publication is still new, and no studies have been performed (at this writing, March, 1996) to analyze the results of this effort, but at Hypertext 96 Bieber provided at least one interesting observation about the authoring process: Most of the hypertext researchers in this collaborative effort had great difficulty generating good semantic labels for the links. This is surprising to me, because hypertext is a domain that is concerned with meaning and communication of meaning. Apparently, grasping the very notion of applying meaningful explicit labels was either very difficult for people without a formal understanding of the classification process or the participants lacked a commitment to doing it completely and well.

11. How Well Is Business Learning These Lessons?

Based on these and other cases, how well are business enterprises learning to organize knowledge? Not very well at all, it seems!

These case studies are still concerned with access to information, not with "knowledge" and its role in business organizations. Of course, raw access is a necessary first step, and some of the merits of improved access can be assessed by the initial savings they generate. But the need for access to knowledge that solves problems, enhances effective performance, and creates a competitive advantage is increasingly critical. Businesses will be forced to change their approaches, because we waste so much time if we cannot understand information and apply knowledge ... and ultimately re-use that critical asset.

Market forces should cause the change to happen. Think about the transformation of software application help systems over the past 10 years. They were once little more than online reference manuals. Now they contain rich hypertext relationships, they know your "context" in a limited way, and some of them even interact with you via "wizards."

It's impossible to draw statistically valid conclusions from three dissimilar case studies, but there may be some useful hints in this diverse sampling. The range of applications itself is instructive: (1) information tightly coupled with job functions, (2) information in a large central repository that requires sophisticated methods of classification and access, and (3) pinpoint location of the instances of an idea within a single publication.

Some conclusions that might be drawn from examining the case studies and the business requirements in general:

• **Downsizing will create opportunities.** There is a real, growing need to make information accessible, and there is money in it for information and library science professionals if the business requirement can be connected with information access and knowledge management.
• **Effectiveness will be evaluated.** If classification systems or practices fail to meet the audience requirements for effective access, the systems themselves will be changed on the fly. General principles for constructing adaptive classification systems will be vital.
• **Opportunities for experts in classification will exist, but they may center on design, not application.** Any time you can tell a business manager she can save $5 million, you are assured of an easy sell. (A positive sell on the quantitative merits of enhanced access may not be easy for some time yet.) But technology will replace application of classification systems by humans in many cases. For example, Digital will gradually replace its 12 indexers with selection of keywords by its own powerful Alta Vista Web search engine. And pattern recognition will supplement or even replace manual assignment of profiling information to bit-mapped images. Therefore, design and testing of classification systems, development of methodologies for applying them, and selection...
of supporting technology may be the primary opportunities.

- **Metadata associated with information elements will encompass patterns of creation and use, not just meaning.** For example, new products from the Xsoft division of Xerox Corporation support attaching “business attributes” to information objects—not just attributes that categorize objects by meaning. Information may be seen in isolation, but it may also be viewed in the context of its development, change, and successful applications—that is, in its context as knowledge in business processes.

- **We will experience a transition in emphasis from information access to creation, management, and extraction of knowledge.** The business requirement is not just accommodating how people find information but how do they derive knowledge and performance from it. New information models based on the dialog of information seekers with electronic subject matter experts will require thoughtful new solutions.

- **All organizational stakeholders will participate in the building of corporate knowledge assets.** It’s not possible to develop comprehensive top-down methods of managing all the knowledge resources in a company. There’s too much of it, it changes too fast, and the need for accuracy and effectiveness will not diminish. The alternative is to enable bottom-up development of those assets in a participatory environment. Formal meta-principles for self-organizing knowledge resources will be vital, because we must use computers to assist the processes of knowledge development.

Information and library science people will also have to make the transition to the knowledge-based business paradigm. If they understand that the requirement is “Solve my problem” or “Give me a competitive edge,” they can help lead that transition. The outlook for applying the lessons of information science—especially in the area of classification—should be good, because they transcend technology-driven solutions. But their success may depend in part on formal models that enable a wide range of supporting technical solutions.

References
Critical Notes on the Use of Knowledge in Knowledge Management

Extended Abstract

In recent years management and business studies have witnessed the emergence of a new field of activity, named Corporate Knowledge Management (CKM). The primary goal of this management discipline is to enhance and improve the use of knowledge as a corporate resource in organizations. Theoretical and practical initiatives in CKM, however, have met some difficulty regarding the conception of knowledge that should guide these efforts. This paper critically examines the contribution of Nonaka (1994) to this conceptual problem.

The emergence of CKM and the recognition of knowledge as an independent production factor next to labour, capital and natural resources have been prompted mainly by the information revolution, the explosive growth of the service sector and increased innovation-based competitiveness in a globalizing economy. In this context CKM is generally described as a consciously controlled and systematic effort to efficiently acquire, share, coordinate and deploy knowledge with a view to the optimization of value-added in intra- and interorganizational value chains. Some important problems regarding this effort centre around the intangibility of knowledge. Organizations especially find it difficult to keep track of the knowledge they have and to assess its quality and practical value, as the usefulness of individual knowledge units inevitably depends on the availability of other knowledge and their embeddedness in wider socio-cultural infrastructures and practices.

With Kuhn (1970, 4-5) we hold that effective professional practice builds on the adoption of a taxonomy which enables the articulation of problems, basic relations within the universe of discourse and preferred methods and standards to guide professional activity. An important difference between the approaches of philosophy and CKM to knowledge in this respect is that the responsibility of CKM cannot be restricted to knowledge that is of an incontestably scientific nature but must extend to all alleged knowledge a company accepts consciously or unconsciously as a capacity for corporate action. In terms of Poppers' three-world ontology (1972) CKM focuses not on knowledge as the objective, propositional content of thought irrespective of persons actually holding those thoughts (world 3), but on knowledge as a subjective mental state of persons having certain beliefs (world 2).

As one of the leading pioneers of CKM and to cope with the problem of continuous innovation be it technological, strategic or organizational Nonaka (1994) has proposed a "dynamic theory of organizational knowledge creation" in which he supports this subjectivist approach to knowledge. Basic to Nonaka's theory is the idea that to be innovative organizations need to actively tap and mobilize the tacit knowledge resources of their members through an elaborate process of "externalization". This process is fed by the commitment of individuals to share their experiences and confront their interpretative frameworks with one another in multi-functional "self-organizing teams", to conceptualize their ideas by use of metaphor and analogy into a model and consequently to act them out in a process of "crystallization", through which new product or organization concepts are refined and tested on their reality and
applicability by various departments of the company. Thus crystallization comprises the function of knowledge justification, which according to Nonaka is "the process of final convergence and screening, which determines the extent to which [new] knowledge . . . is truly worthwhile for the organization and society". Quality standards, Nonaka claims, generally include cost, efficiency, profit margin and the like as well as more aesthetic and "romantic" criteria related to a company's vision of the future and its own development (Nonaka, 1994, 26). Nonaka considers the determination of these standards a highly strategic task of company leaders, ultimately enabling "a truly 'humanistic' knowledge society beyond the limitations of mere "economic rationality"" (p. 34).

Apart from the rather vehement rhetoric Nonaka uses, our criticism of his views focuses on two central issues that are directly related to his subjectivist conception of knowledge. The attraction of subjectivism and the philosophically more familiar conventionalist view of science is that they prevent us from shifting the responsibility for the truth of our knowledge entirely to the object of research. This entails, however, the relativity of knowledge to particular practices or forms of life (Wittgenstein, 1953) and the concomitant possibility of the incommensurability of interpretative frameworks or paradigms (Kuhn, 1970). The concept of incommensurability refers to the absence of impartial standards or criteria to judge the cognitive and practical merit of rivaling frameworks and the knowledge claims involved in them.

Our first point of criticism is that Nonaka completely ignores the issue of incommensurability. This is especially regrettable for two reasons. Firstly, as Kuhn (1970) has pointed out, viewed diachronically fundamental breakthroughs in science by their very nature involve incommensurable frameworks. In R&D-management this is generally recognized in the claim that innovation as opposed to improvement is revolutionary. Secondly, viewed synchronically Nonaka misses the possibility of communication breakdowns and conflict in the organization due to the clash of incommensurable paradigms. As such incommensurability is not just an interesting philosophical problem but constitutes a serious practical problem for organizations comprising a multiplicity of cultures, institutional arrangements and professional backgrounds. Organizations cannot be conceived of a priori from a unitary cultural perspective.

Related to this, our second point of criticism focuses on the problem of knowledge justification. Again we take issue with two aspects of this problem. Firstly, by introducing the justification standards already mentioned, Nonaka seems to adopt an extreme relativism on the validation of knowledge claims, which is philosophically unwarranted even from a conventionalist perspective (cf. Kuhn 1970; Habermas 1981, 1984). We cannot simply believe what we like, based on the promise of a bright (financial) future. Secondly, Nonaka overlooks the possibility of fundamental conflict between standards as part of the incommensurability of rivaling frameworks. As quality standards according to Kuhn have to be taken as values, this opens up important moral-political ramifications of the problem of incommensurability in organizations (Bernstein, 1991). Nonaka apparently "solves" these issues by authoritarian means (the "strategic task of organizational leadership") or by implicit self-censorship within the company ("commitment"), which in our view might easily stifle and repress innovation.

In conclusion we claim that Nonaka paints far too rosy a picture of CKM, that naively ignores the fundamental moral-political issues that are at stake in CKM, by which the ideological sincerity of his tribute to a "truly humanistic knowledge society" is cast into doubt. While supporting a conventionalist conception of knowledge in CKM, we claim that it is essential to conceive of it as an integral part of the management of organizational culture and that CKM has to take into account the ideological ramifications of organizational knowledge creation. If money and power are going to be the dominant control-inspired criteria of the
field, our hopes for the future of CKM do not run very high.

References
Facilitating the Interaction of User and Public Knowledge Organization With User Profiles, User Views and User Education Plans

Abstract: This paper presents a new approach to bibliographic instruction which incorporates evaluation of the users needs. The model proposed creates a User Profile based on the depth of knowledge, subject area expertise, and information task of the population to receive instruction. The model uses this User Profile to posit a User View, an idealized “target” information pool which closely approximates the informational needs of the user. It then creates a heuristic search strategy and user education plan (UEP) based on this view. Suggested approaches to the application of this model and an example are provided.

1. The Problem

Reference librarians are not only the gatekeepers for access to online public knowledge organizations, they are also, more often than not, called upon to be the first line of instruction in their use. They find themselves instructing several different categories of users, with different levels of education, divergent needs, and various levels of computer literacy, on how to locate, access, and search the appropriate online resources. Some of this instruction is occurring in semi-formal and classroom situations, and many reference librarians, because they have often not received formal teacher training, are ill-equipped to deliver the right kind of information to the specific and varied audiences they must deal with. Even if the modern reference librarian has mastered the knowledge of on-line public knowledge resources and their software interfaces, the question remains, can he or she communicate that knowledge in an appropriate instructional package, to the user?

2. An Illustration

To illustrate the problem, let’s take a look at a real (and perhaps all too typical) example of a failure to deliver appropriate instruction. In the following scenario the librarian did not take the user or user group into appropriate account when she conducted a training session. As a result, she failed to meet any of the user’s needs and expectations.

Bibliographic instruction (BI) was presented to upper-level undergraduate nursing students. The bibliographic instructor, who was a senior reference librarian, had only 45 minutes to demonstrate how to search the electronic database CINAHL by OVID. OVID has two search modes: easy and full, the former operates with one function tool bar and the latter operates with two function tool bars. The bibliographic instructor did not conduct a reference interview with the nursing professor responsible for the class prior to the instruction, so she selected the example nursing topics at random, none were related to this particular course.

For the presentation, the librarian provided an in-depth explanation of every function of the database, for both easy and full mode. She then proceeded to demonstrate how to search in both modes, using two different nursing topics, both, as was mentioned previously, irrelevant to the course the students were taking. The students were then given an assignment,
but no student was able to complete it due to the confusion caused by the presentation. As a result, the students were overcome with anxiety because they could not complete the assignment and did not learn how to search the database for relevant topics. They left the instruction session with a negative attitude toward using the database, not to mention the library. A survey given at the end of the session also indicated that they did not feel comfortable approaching the librarian for further instruction. All in all, a deplorable situation.

3. Delivering the Informational Goods: User Profile and User View

Obviously the instructor failed to "deliver the informational goods" to the students. What went wrong? In a typical academic reference setting there are likely to be many different categories of information user. They may vary along several dimensions.

One dimension is depth of knowledge in subject area. Some users will be novices in their subject area, others will be more knowledgeable, and yet others will have a detailed knowledge of their subject matter. It is possible to develop a simple classification scheme for user types, which could be further refined in an actual application setting. This classification is based on domain knowledge:

A = General Users
superficial to nonexistent knowledge of domain, no knowledge of special vocabulary (terminology) of the domain

B = Intermediate Users
general domain knowledge, with intermediate familiarity with the terminology of the domain

C = Expert Users
high level of domain knowledge with a high degree of familiarity with the terminology of the domain including sophisticated concept relationships

Another dimension, which may intersect variably with subject area expertise, is extent of computer literacy. The user with the greatest depth of knowledge in his/her subject area may or may not be the most knowledgeable computer user. We can classify such users as follows:

1 = Novices
low level of experience with computers in general, little or no experience with searching and concepts of searching.

2 = Intermediates
some familiarity with computer operations; modest experience with searching and search concepts, e.g., keywords, etc.

3 = Experienced
high level of familiarity with computers and different software interfaces; significant experience with different types of searching and with advanced search concepts such as limiters, boolean operators, etc.

Further, each user is seeking reference assistance because of a particular information task they are trying to carry out. The information task dimension includes subject area, topic, type
of assignment, purpose of search relative to assignment, and so on. A simple classification scheme for information task might include:

\[ a = \text{simple (non-specific or general task)} \]

A simple task can be fulfilled with non-scholarly materials, e.g., undergraduate term paper on relatively broad topic not requiring in depth knowledge of subject area; general information in domain with broad scope, shallow depth, relevance factor: high recall/low precision

\[ b = \text{semi-complex} \]

A semi-complex task can be fulfilled with mixture of scholarly and non-scholarly materials, e.g., honors thesis or senior research paper; greater depth of search, scope is more focused, relevance factor: variable

\[ c = \text{complex} \]

A complex task can only be fulfilled with scholarly and research material, e.g., theses, dissertations and faculty research; searches of greatest conceptual depth and scope is highly focused. relevance factor: low recall/high precision

---

**SUBJECT AREA**

- GENERAL
- INTERMEDIATE
- EXPERT

**COMPUTER LITERACY**

- NOVICE
- INTERMEDIATE
- EXPERIENCED

**NATURE OF TASK**

- SIMPLE
- SEMI-COMPLEX
- COMPLEX

---

The combination of subject area preparation, computer literacy and specific information retrieval task creates a profile for each user. This user profile can be used to determine both what a user should need to access from an online resource and the most appropriate methods for accessing it. One of the most significant failures of our hapless bibliographic instructor was that she failed to make any attempt to develop a profile of her user audience. She had no sense of their topic (outside of the fact it was “nursing”), how much they knew about the topic and of the demands of their particular assignment (the information retrieval task). Further, she had no sense of how much they knew (or could absorb) about the online system itself. She failed to develop a user profile and, thus, could not develop an appropriate training strategy.
Using a concept borrowed from systems analysis, we can also say that she failed to determine the particular *user view* of the information that was called for. In most large organizations there is a stupendous amount of information potentially available.

But the engineer, the accounting clerk, and the salesperson in an automobile factory, for instance, do not need access to all the information available and they do not need access to the same information. Similarly the undergraduate biology major writing a term paper, the biochemistry graduate student writing a thesis and the full professor of biochemistry researching a scholarly book, each project a different need on the total available universe of biochemistry information. Each user has a particular task to carry out, thus only a selection of the available information needs to be given to them. This user-determined window on the data is a user view.

![User View Diagram](image)

**Fig. 2: User View is a Projection of User Profile**

The argument we want to make is that in order to be able to deliver the information goods to a user, the librarian has to act as a kind of systems analyst. The librarian ought to determine what the most appropriate user view on any public knowledge organization is and then configure an instructional package that teaches the user how to access the information selected by that view. Our example bibliographic instructor did not create a user profile, did not project a user view, and could not, therefore, meet her student/user's expectations for information needs with an appropriate training package.

4. **Summary of the Model**

User view is determined by several factors which arise from the user profile. These factors include, most importantly:

- the nature of the particular information task-at-hand (subject area/topic, depth and scope of information required, purpose of search, etc.)
- depth of knowledge of the subject (domain knowledge) including, particularly, knowledge of the special vocabulary (terminology) of the domain and the conceptual relationships
obtaining between them

- existing level of ability to access information, including existing knowledge of search strategy, existing patterns of information-seeking behavior, and understanding of on-line resources
- general level of computer literacy and ability to deal with and understand software interfaces

From the user profile the librarian/analyst identifies a user view. The user view is an idealized "target" information pool selected by the user profile. That is, the view is a pool of information the librarian projects will be the most appropriate pool for the user to access — given the task and its purposes, the existing level of user knowledge and the existing level of search ability. Included in the projected user view is a preliminary search strategy. This preliminary search strategy should be characterized by the following attributes:

- it is topic relevant (e.g., it should be about a topic or topics with which the student task is involved)
- it is task relevant (e.g., it should address some aspect of the task with which the students are faced)
- it is instructionally focused, (e.g., it should include those only those operations or techniques required for successful completion of the task at hand and should not try to cover all possible modes of access and searching.)
- it should utilize strategies of a scope and depth appropriate to the student levels of search ability and subject knowledge.

Once the librarian/analyst has determined a user view with its attendant heuristic search strategy then, especially in settings where, for instance, an instructor has brought a class to the library for a formal introduction to searching on-line resources, he or she must develop a user education plan (UEP). Unless the librarian intends to do the searching him or herself, the UEP is a necessary next step. The UEP is an educational package: a customized instructional set consisting of a heuristic search strategy, the instructions entailed by that strategy, and accompanying relevant explanations and examples. The entire package is derived from the user profile and its associated user view.

Summarizing the steps to the method:
- Phase 1: determine the user profile
- Phase 2: evaluate the user profile and determine a user view
- Phase 3: create a heuristic search strategy and adopt a user education plan

5. Applying the Model: Some Approaches

The model we presented can be partially implemented as a faceted classification scheme from which a series of formulaic potential user profiles can be constructed, for example: A1a, B2b, C3a etc. The formula B2b, for instance, would be an intermediate user (senior or first year graduate student) with a reasonably broad, but not very deep, knowledge of domain and its terminology, modest computer skills and experience of searching, and a semi-complex task which could be completed by accessing a mix of general and scholarly materials.

The task of the librarian as bibliographic instructor, would be to create such a faceted classification scheme as a means to identify generic profiles and user views existing in the user population he or she serves. These generic profiles could be attached to sets of prepared UEP's,
one UEP for each identified profile. The classification scheme presented earlier allows for 27 combinations, but it is possible to sophisticate the classification. The scheme should be as detailed as the makeup of the user community demands. Clearly it is possible to create a UEP for each user and each possible topic. And, indeed, this is what we could and should do in the best of all possible worlds and if we all had an unlimited amount of time.

In the real world we are more likely to succeed with a more modest approach. The bibliographic instruction community in a particular organization should undertake a kind of "ethnography" of the user community. This ethnography should identify the number of classifications required in each category. The ethnography should identify, perhaps by frequency, the subjects and topics most often requiring bibliographic instruction. From the ethnographic information a set of base UEP's could be constructed. This base set would provide user-appropriate training for as much as 80% of routine bibliographic instruction.

It should also be possible to combine these base, pre-prepared UEP's with interviews to allow for the inclusion of profile information that is highly variable, such as the specific topic of an assignment. In most user populations certain features of the user profile will remain relatively constant over time. It should be possible for a generic profile to accurately describe the characteristics of the "typical undergraduate English major with a moderate level of computer skill." Such a profile will remain relatively stable over a period of years, shifting only gradually. A UEP for such a user could be instituted in the nature of an open-ended training document. If stored in electronic form, a simple interview prior to the training session could be used fill in any variable data, such as the particular topic the English instructor wants this particular class to research.

Of even greater value would be a database of plans on various topics at the various classification levels stored on-line and retrievable by the BI prior to a session. Ultimately the goal should be for every bibliographic instruction session, even those with individuals, to be developed and recorded as an electronic UEP document stored in a database. Then users could gain access to the UEPs themselves and look at search strategies, explanations and examples on topics and at levels that match their own profiles.

6. An Example of a UEP linked to a User Profile

The following is a training plan, including a core search strategy, which reflects the needs of a User Profile classified as a B2b. The B2b comprises the subject area for intermediate users (B), the computer literacy intermediate level (2), and the nature of task as semi-complex (b). The B2b will be implemented for the user profile of nursing students, the example we mentioned previously. The UEP for the scenario is as follows:

The librarian consults with the nursing professor prior to the bibliographic instruction. During this interview, the librarian learns that this course is on geriatric nursing and it is the first course of the core nursing program. The librarian also discovers that the nursing professor is discussing the topic of Alzheimer's disease, and only wants her students to search for review articles from professional nursing journals.

It is from this interview with the nursing professor that the librarian determined the appropriate user profile and user view. For example, the student/user is upper division undergraduate students who have knowledge of nursing terminology, and the student/user has a specific objective when searching the CINAHL database. The only information not determined yet is the level of computer literacy. Here, the librarian can conclude that these students are upper division undergraduates who are probably familiar with using the computer, keeping in mind that this is a class instruction where some students have little or no computer skills, using the
intermediate level of computer literacy would be appropriate in this given situation. After determining the user profile and user view, the librarian is able to create a UEP tailored to student/user need's and expectations. The librarian designs the following UEP:

- introduction of CINAHL by OVID, explaining that for this B.I. session, only the easy mode will be used, begins demonstration
- explanation of easy mode function keys and the keys relevant for this B.I.
- demonstration of the search strategy, using the 'subject' function key for the topic of Alzheimer's disease, and explanation of topic relevance to the course
- explanation of how to limit the subject to age specification, using the 'limit' function key, since the course focuses on geriatric nursing
- explanation of how to limit the topic of Alzheimer's disease in the aged to review articles, as requested by the nursing professor

After completing the UEP portion of the B.I., the librarian can explain how to retrieve the citations and where to locate the actual journal review articles. This process should be completed in 25-30 minutes. The remaining time should concentrate on allowing the student/user to use a hands-on approach and practice the search themselves. This hands-on approach allows the student/user to become familiar with the search techniques and the database. As a result, the student/user is not overwhelmed by using the computer and will begin to explore various search strategies and techniques.

The UEP scenario, designed by the librarian for a specific situation, fulfills the requirements of the user profile and user view to tailor to instruction to the needs of the user, therefore, delivering the appropriate informational goods.

7. Conclusion

Adopting the user profile, user view, user education plan model should improve the ability of the librarian or bibliographic instructor to develop a training strategy that more closely meets the needs and matches the levels of ability and subject knowledge possessed by user groups. By improving the quality of instruction, we can better provide users with the means to successfully access and retrieve the information they need from public knowledge organizations.

Acknowledgments

I would like to thank Dr. Gregory M. Shreve, Professor and Director, Institute for Applied Linguistics, Kent State University—my friend and mentor—for so graciously contributing his time and energy as reviewer and consultant for my first paper to be presented.

References


Nohr, Holger. (1994). Technology assessment—a challenge to knowledge organization?


Selection of Search Terms as a
Meeting Place of Different Discourses

Abstract: Selection of search terms is considered a meeting place of different discourses. Discourse refers to the ways of talking and thinking about a certain topic. There exist concurrently different discourses on the same topic. Although the selection of search terms used to be described as a translation process, it could also be seen as a situation where searchers step into different discourses and select search terms from various sources on the basis of their own previous experience. Based on an empirical study, the paper describes the various sources of search terms. They are controlled vocabularies, assumptions about indexing, documents and their titles, clients’ words, databases, and the searchers’ own search experience. Because the searchers have different experience and are working in different types of work environments, they use these sources differently. The paper also outlines a model for understanding the selection of search terms as a meeting place of different discourses.

1. Introduction

The selection of search terms has usually been described as a translation process. We have learnt to think that during the process the concepts of a search request described in the client’s words will be translated into the words which can be used in a certain information system.

According to previous studies we know, however, that the selection of search terms is not only a translation process but also the seeking of search terms from different sources. Saracevic et al. (1991) investigated the sources of search terms and found that only 38% came from the written search request, 23% were generated by the client during the interaction between the client and the searcher and 39% came from sources other than the client (e.g., thesaurus and relevance feedback). In former studies it has been shown also that the searchers consider various alternatives for search terms on the basis of their own previous experience and their own work environments. Fidel (1986, 1991a, 1991b, 1991c) found that the searchers had certain routines and they used rules to support their selection of search terms. She also found that the searchers who worked in different types of search environments (practical, theoretical and general) selected search terms differently.

We can consider the selection of search terms a meeting place of different discourses. Discourse refers to the ways of talking and thinking about a certain topic. The discourse of welfare state refers to the ways of talking and thinking about welfare state. It is necessary to see that the different discourses of the same topic may exist concurrently. It is also necessary to understand that the discourses of one topic may change over time. The discourses have also their social functions; they have been produced in social practice (Abercrombie et al., 1994, 119-120).

When the searchers select the search terms to describe one certain search topic they have to step into several various discourses where this topic may be discussed and taught differently. The searchers have to try to understand the client’s talks and thoughts. They have to take into account the ways in which controlled vocabularies describe and organize the same topic. They have to meet the ways in which the authors of the documents talk and think about the same
topic. In addition, they have to consider the ways in which the topic may be discussed in the practice of indexing and databases. The searchers step into these various discourses with their own experience. Because the searchers' experience differs from the others' experience, they may meet and understand these discourses differently.

In this paper I will first describe the sources of search terms on the basis of my empirical study. I will show how strong the role of different sources is in the selection of search terms. After that I will outline a model for understanding the selection of search terms as a meeting place of different discourses.

2. Sources of Search Terms: an Empirical Study

In my empirical study related to the consistency in the selection of search concepts and search terms I also paid attention to the sources of search terms. A total of 24 professional searchers from three types of search environments and eight students of information studies participated in my study. They selected search terms for a search of the same 12 search requests. They also justified their selections. In the group of academically oriented search environments there were six searchers working in university libraries and one searcher working in an information bureau offering services to academics in the whole country (Finland). In the group of special service-oriented search environments there were, in addition to five searchers working in special libraries, two searchers working in university libraries offering, however, specific services to specific user groups in very narrow subject areas. In the group of public search environments there were eight searchers working in public libraries. None of the professional searchers were from the same place of work. The data of my study and different search environments are described in greater detail elsewhere (Iivonen, 1994, Iivonen, 1995).

On the basis of the searchers' explanations it was possible to see that the searchers got ideas for search terms from different sources (see Figure 1). The controlled vocabulary had a very important role as a source of search terms. Almost every searcher (30 searchers out of 32 = 93.8%) referred to it when they justified the selection of search terms. The controlled vocabulary is a language and it is in a very concrete way present in the selection of search terms. It offers a frame through which the world can be interpreted and described. As a discourse the controlled vocabulary offers ways to talk about a certain topic. At the same time, however, it restricts other alternatives very clearly and often also very strongly. In a controlled vocabulary there always exists the function of standardization; both the preferred terms and their relationships have been standardized. On the basis of the searchers' own justifications they seem to be very involved in the discourse of the controlled vocabulary.

Another fairly important source for search terms in my study was documents. Twenty three searchers (71.9%) referred to the documents and their titles in their justification of search terms. They thought of the ways in which the authors of the documents have talked about the topics of the search requests. One searcher who selected search terms "welfare" (in Finnish "hyvinvointi") and "welfare state" (in Finnish "hyvinvointivaltio") from a search request where social problems and people's everyday life in a society were discussed justified her selection as follows:

(\textit{Search request: Social problems, alienation from society, and structural change in society, and their influence on the control people have over their everyday lives in an affluent society.}) . . . I thought first that it (welfare) may occur in the title or in
The percentage of searchers who justified their selection of search terms by referring to the following sources of search terms:

1 = controlled vocabulary
2 = documents (e.g., the titles of the books)
3 = practice of indexing (assumptions about indexing)
4 = clients' words
5 = database
6 = own search experience in general

Fig. 1: Sources of search terms

In a certain way it is a challenge for the searchers to explore those other words with which the topic of a search request may be discussed in the literature. It is a real task to find a right context from the literature for the search request.

In my study it was not only the discourse of the controlled vocabulary in which the searchers were heavily involved. In addition they were involved in the practice of indexing. Several searchers (22/32 = 68.8%) mentioned assumptions related to the indexing process when they justified the selection of search terms. They thought how indexers would have used controlled terms in describing certain topics. The practice of indexing produces ways of talking about a certain topic, and these ways may differ from the ways produced by the authors of the documents. There are certain rules, policies and restrictions in the practice of indexing. Therefore it is understandable that the searchers take into account these ways when they select search terms.

The clients, when they describe the search requests with their own words, bring their own discourses to the search process. The clients' words are not necessarily very familiar to the professional searchers. They can, however, trust the clients' ways of talking about the topic and consider the clients' words as search terms. When the clients name a certain concept with
a certain term they tell more than only the topic of a search request. They also tell about the frame and tradition they come from because about the same topic is discussed with different words within different traditions. In my data a few searchers (14/32 = 43.8%) pointed out that they selected certain words for search terms because they were mentioned in a search request. One searcher described her confidence in the client’s words as follows:

‘This was difficult. I selected environmental movements and alternative activities. I would not have noticed that alternative activities... I think that if the search request could have been formulated in another way, so I could not have noticed it... I would have used environmental movements and nature conservation movements.’ (A searcher working in public search environments)

We might assume that the database for which the search terms were selected could have had a strong position as a frame in the selection of search terms. Surprisingly, it had not. Only ten searchers (31.3%) referred to the database when they justified their selection of search terms. They considered whether a certain search term could be or could not be useful in the database.

Similarly only 10 searchers (31.3%) referred to their own previous search experience when they justified the selection of search terms. However, 24 subjects out of 32 were professional searchers who had search experience. Of course they had their own experience although they did not explain it in words. The searchers always act on the basis of their own experience and they cannot exclude it from their action.

There were interesting differences in the use of different sources in the selection of search terms between the searchers working in the different types of search environments (see Table 1). Only the controlled vocabulary was referred to as a source of search terms equally and very strongly in each type of search environment. This tells about the strong position of controlled vocabularies in the selection of search terms.

The searchers working in academically oriented and special service-oriented search environments took into account documents and their titles more often than the students or the searchers working in public search environments. The collections of documents are more limited in academically oriented and special service-oriented search environments than in public search environments where all possible subject areas are included. It was therefore easier for the searchers in these environments to think about the discourse of documents. Possibly the restricted collections helped the searchers to understand and get a specific picture about the literature of the subject area of the search requests.

The clients’ words were referred to mostly by the searchers working in academically oriented search environments. The clients in this type of search environment were experts in their own domain. It is therefore easy to understand why the searchers trusted their clients’ own words, considered them and took them into account as search terms.

The professional searchers referred more often to the indexing process and to their own search experience than the students. The students did not have yet work experience— neither in the area of information storage nor in the area of information retrieval—and they could not use it in the selection of search terms. The searchers’ own work experience affects their decisions and when they lack this experience they cannot use it.

On the basis of the searchers’ justifications for the selection of search terms we can find that the searchers know the sources of possible search terms differently. Therefore, their
Source of search terms | Type of search environments | All search environments
---|---|---
---|---|---
controlled vocabulary | Special service-oriented search environment (n=9) (%) | n=32 (%) |
documents | 9 | 100.0 | 30 | 93.8 |
academic oriented search environment (n=7) (%) | 6 | 85.7 | 23 | 71.9 |
public search environment (n=8) (%) | 7 | 87.5 | 22 | 68.8 |
studying environment (n=8) (%) | 8 | 100.0 | 14 | 43.8 |
| | n= number of searchers who referred to this source of search terms |
| | Table 1: Sources of search terms mentioned by the searchers working in the different types of search environments |

abilities to take these sources into account vary. We cannot expect all searchers to behave similarly.

3. Meeting Place of Different Discourses

The selection of search terms is a meeting place of different discourses. The searchers step into these different discourses on the basis of their own experience and are therefore differently involved in these discourses. The searchers have to consider the ways of talking and thinking about the same topic inside the various discourses. It is obvious that the topic of the search request it has been talked about differently in the documents than in the controlled vocabulary. The controlled vocabulary may rule out alternatives of talking about the topic. It is also obvious that the practice of indexing may have some certain rules and ways which differ clearly from the practice of clients. Many times it is even difficult for clients to understand in which way and why just in this way the topic is described in indexing.

The searchers are (or at least they should be) conscious of these different discourses. They can think of alternatives for search terms from this point of view. The selection of search terms can be described in a model where the searchers consider the possible search terms stepping into the different discourses (see Figure 2).1

4. Discussion

The searchers are not satisfied with being only translators when they select search terms. They are more like explorers in different discourses. They consider different alternatives from different sources knowing that about the same topic may be talked of differently in various discourses. They also know that the ways of talking about one topic vary and change considerably across time and place.

- Although there are several different sources for search terms all of them are not as important as others and are not equally considered by all searchers. The controlled vocabulary has, however, a very strong position in the selection of search terms in general. Therefore, it is not a question of secondary importance which kind of discourse the controlled vocabulary offers.
Although there are several different sources for search terms all of them are not as important as others and are not equally considered by all searchers. The controlled vocabulary has, however, a very strong position in the selection of search terms in general. Therefore, it is not a question of secondary importance which kind of discourse the controlled vocabulary offers.

The understanding of the selection of search terms as a translation process may lead too easily to the replacement of one word (a client's word) with another (a word which can be used in a system), to replace one word with only another one. The understanding of the selection of search terms as a meeting place of different discourses leads to seek and explore search terms from different sources and to compare them. The fact is that there exist many alternative search terms for one topic and many various words can replace one word named in a search request. The fact also is that the same topic may be talked of differently in different discourses. For the searchers there exists the horn of plenty or a treasure of words to be used.

Notes

1. I would like to thank Dr. Diane Sonnenwald and Dr. Leah Leavrouw for the idea of this model. I got the idea for my model from their article (1996), which I translated into Finnish. They presented a provisional model of communication in the information system design process and showed the participants' interactions in design process in various phases. (See also Sonnenwald, 1993, 1995.)
References


Empowering Users for Improved Database Access and Analysis Through the Application of Knowledge Structure Views, Progressive Refinement Techniques and a Design Approach Driven by Usability

Abstract: An approach to accessing databases, which makes extensive use of structured thesaurus views for presenting and searching a large bibliographic database through progressive refinement, is being investigated in a research project which is seeking to apply usability criteria as the determinant for interface design. It aims to investigate measurements of retrieval performance in an application of graphical user interface techniques using multi-windowed layered attribute value aggregation, including those provided by a structured thesaurus, to access bibliographic databases. This paper reports on progress made in the first year of the project in respect of the application of usability methodologies, the revision of a preliminary design and the resulting prototype development. Implications for more widespread application are discussed.

1. Introduction

Research and development which seeks to improve access to databases through the increased utilisation of thesauri in the user interface has been undertaken at the University of Huddersfield since the early 1980s. The original work focused on an expert systems approach whereby selections from a display of Medical Subject Headings concerning cancer therapy were processed by rules which produced legal search statements for searching MEDLINE (Pollitt, 1987). Although the CANSEARCH system occasionally performed better than a human intermediary the expert systems approach proved difficult, if not impossible, to apply for other areas of biomedicine (notably biochemical genetics).

An alternative approach was devised which replaced the underpinning complex rule-base with a simple algorithm to produce a series of search statements similar to those applied in Quorum searching (Cleverdon, 1984). The user was presented with the results from a number of searches which applied different levels of co-ordination for the selected terms and concepts. The resulting system, MenUSE (Menu-based User Search Engine) demonstrated access to the complete MEDLINE database and was shown to be database and subject independent, relying on the thesaurus or controlled vocabulary as a means for the end-user to specify the subject matter of their search. (Pollitt, 1988).

Although this approach was applied to different databases, notably IEB’s INSPEC and the European Parliament’s EPOQUE database (Pollitt, Ellis and Hosch, 1994), a second dramatic change in approach lead to the adoption of what has been termed “view-based” searching. Most commercial interfaces for information retrieval incorporate thesaurus access to facilitate inclusion of search terms through a forms-based or command-line interaction. View-based searching, on the other hand, seeks to utilise the thesaurus display both as a means for

identifying relevant concepts and terms and to display the results of searching through the application of filters which afford an implicit ANDing of the presented terms (Pollitt, Smith and Ellis, 1994). These developments have clear parallels with the work on Graphical Boolean Expressions and Dynamic Queries carried out by Shneiderman and his colleagues at the University of Maryland. "Graphical selection of attribute values, progressive refinement of queries, and immediate recomputation and redisplay of the results appears to contribute many benefits" (Shneiderman, 1992 pp 424-425).

HIBrowse (High resolution Interface for Database Specific BROWSing and SEarching) uses multiple views modified through selection and progressive filtering (Pollitt, Ellis and Smith, 1994). An original HIBROWSE for EPOQUE design is presented in Figure 1. The HIBROWSE design was the subject of a research proposal to the British Library Research and Development Department which sought to examine issues of usability and performance. The resulting project was started in March, 1995 and is examining the application of the HIBROWSE design principles to EMBASE, Europe's most important biomedical bibliographic database. This paper reports on the progress made in the first year.

2. The MUSiC Method

Integral to the initial intentions of the HIBROWSE project was the use of the MUSiC methodology for usability design and evaluation. MUSiC, Metrics for Usability Standards in Computing, (Macleod and Bevan, 1993) is the result of a large European-funded project whose partners included the University of Cork in Ireland and the National Physical Laboratory in the UK. MUSiC provides a set of tools and methods which can be applied during the design of applications during phases of task and requirements analysis, design and evaluation. However, no particular development lifecycle model is assumed or advocated by the developers of MUSiC.

The suite of MUSiC tools consists of the Context Analysis questionnaire and report,
SANe, DRUM, and SUMI. The Context Analysis questionnaire and report are completed as the first stage of system development, the purpose of the context analysis is to identify the intended user population and the tasks they are seeking to undertake. The responses to the questionnaire are used to define the way in which formal usability evaluation and testing will be conducted later in development, and to set metrics on users' expected task performance. The context analysis will be discussed further below. SANe, the Skill Acquisition Network, is a computational cognitive architecture which may be programmed with a representation of the knowledge needed to perform tasks on the intended final system and with a model of the behaviour of the intended final system. SANe is able to reveal the success with which users may perform tasks on the final system and highlights problems that users may face, such as a stressful load on their working memory making tasks difficult to perform and mentally tiring. It is unlikely that SANe will be directly employed during the HIBROWSE project, although user-centred design will naturally require awareness of user's 'limitations' and mental capacities and abilities.

DRUM, Diagnostic Recorder for Usability Measurement, (Macleod and Rengger, 1993) is a video data logging and analysis tool. It is employed when a working prototype of the intended system is available and usability testing may be conducted. Usability testing within the MUSiC method is typically conducted in the laboratory where conditions may be controlled and the tasks that participants perform may be constrained to representative tasks determined during the context analysis. DRUM allows a video tape of user behaviour to be reviewed and relevant events and behaviours to be noted and logged on the computer. A small set of user behaviours are programmed directly into DRUM to note when users are engaged in productive or unproductive actions, or are reading system help. DRUM is then able to calculate a set of metrics from a small set of behaviours, such as the user's performance relative to an expert user, and the quality of their task performance. More sophisticated metrics may be analysed by programming DRUM with a larger set of events of note.

The tool that provides the last stage of analysis of the MUSiC suite is SUMI, Software Usability Measurement Inventory. SUMI is a psychometric questionnaire used to subjectively rate an application or prototype according to five factors: affect (how much the users like the application), efficiency, helpfulness, control and learnability. A program, SUMISCO, automates the scoring of completed SUMI questionnaires and, using a normalised database of SUMI scores for many types of applications, is able to place the averaged subjective rating of the application in the context of subjective opinions of many other applications. Comparing SUMI scores across applications of the same type, designers are able to comment on whether users prefer one application over another.

3. Context and User Needs

The first tool applied during system development from the MUSiC suite is the Context Analysis questionnaire and report. The context questionnaire is an aid to stakeholders in the design of the system when producing the context report. It is suggested that as a part of applying the MUSIC tools, the stakeholders in the system, such as system designers, programmers, representatives of the user population, project managers, product marketing staff and so on, meet in order to complete the context report. The report describes the intended final product, which can be as large as a new application, or as small as a revision of one application feature, the expected users, their abilities, likely level of domain knowledge, and possible physical and mental impairments. The report also describes the physical environment in which the product is likely to be used, the noise levels, the range of possible temperatures and
humidities, and the major tasks that users perform and the structure of larger sub-tasks that make up the major tasks that should be supported by the product or application being developed. Information such as a description of the physical environment in which the product will be used is typically obtained from field visits, but the remainder of the information needed to complete the context report is obtained during the context meeting. The completed context report then forms the basis of a plan of evaluation, the setting of formal usability evaluation sessions, the representative tasks that participants should perform, and the types of users that should be recruited as participants.

For many applications and products, completing a context report presents few problems. Typically an application will be written for an intended market where users' tasks are well-defined, or where competitor products already exist, or the product may be a revision of an organisation's existing product. Attempting to complete a context report for the HIBROWSE for EMBASE system has proved very difficult. A problem that confronts determining the context of use of an information retrieval system is the successful identification of the user population(s). In the case of EMBASE, the user population consists of everyone who has a need to access medical literature, which could range from research scientists and clinicians conceivably to members of the general public wishing to discover information on drugs they have been prescribed or seeking information after a health scare.

A useful way of describing the space of possible users for HIBROWSE for EMBASE has been a graph of domain knowledge plotted against search knowledge. Medical researchers are likely to have high domain knowledge and considerable database searching experience, whereas clinicians are likely to score lower on search expertise. The public will score very low on both indices. While useful for locating relative positions of different users, this classification does not aid determination of users' roles, educational qualifications or other skills. Clinicians undergoing training in hospitals and in general practice will often be expected to undertake research and publish, but will not have experienced the degree of research training of medical and drug researchers. Identifying the classes of user in sufficient detail to complete the MUSiC context study becomes impossible. Users of the system will have what Wittgenstein calls a family resemblance, rather than be classifiable.

The identification and description of tasks such that an evaluation plan can be prepared is proving difficult. If the user's existing tasks cannot be identified and described it becomes impossible to design the system to support tasks, a major aim in user interface design. In specifying the HIBROWSE for EMBASE system (Treglown, 1996a) in the task-based semi-formal User Action Notation, the tasks described, and the system behaviour when performing these tasks, were at the level of low-level features provided by the HIBROWSE system, such as selecting terms, displaying abstracts and scrolling windows. Attempting to identify larger tasks performed by users, described in terms of the smaller tasks specified, proves difficult.

Unlike most systems, information retrieval systems are required to meet the user need for relevant information rather than supporting users' work and tasks. Decision support systems, on the other hand, are directly associated with the work and task e.g. to determine the most efficacious dosage of a drug patient and route of administration for a particular. Bibliographic databases retrieval systems are undoubtedly used for decision support purposes but, by the very nature of our "open" systems of publication, cannot be as goal directed as the system designed to support a well defined task. Bibliographic sources are key to the preparation of guidelines for therapy which will be more task specific. More direct schemes for knowledge transfer, where research findings can be delivered to benefit both clinicians and patients, will replace the somewhat crude systems of today. In the meantime, improved access to the traditional databases will continue to be of widespread interest.
Considerable work has been undertaken by the stakeholders in trying to identify the tasks according to the guidelines for carrying out a context study, but it has only been possible to determine tasks at a very general level. It is clear that organised and formal task analysis should be undertaken when designing an information retrieval system, but the success of this approach may be limited. Considering the various approaches to obtaining task models from user behaviour (Diaper, 1989), computer logging of interaction with existing bibliographic database systems (where logging facilities exist) are likely only to produce a list of queries submitted to the database. This list will be useful in identifying the subject matter of user need, but will not help designers understand the process by which users articulate a goal in terms of actions or commands and then evaluate the system feedback in terms of their goals and plan further actions accordingly (Norman, 1984).

Traditional task observation methods may be more useful in the design process. A small field study observing an intermediary searching medical databases, conducted using the post-task walkthrough method, demonstrated the usefulness of that method, and larger-scale field work will be undertaken. The very small study revealed, however, the difficulty of extracting task models from observed data as much of the user's tasks were constrained by the user interface of the existing system. For example, queries had to be submitted to each portion of the database stored on a separate CD-ROM, rather than the user being able to submit a single query to the entire database.

Many of the task analysis methods employed in HCI, however, allow the designer to abstract away from the specific sequences of actions and commands that users must enter in order to perform tasks, so that only the larger tasks closer to the user need are considered. Future task analysis and comparison of task analysis methods will reveal whether information retrieval systems pose any special problems for usability-based design.

4. Testing the Preliminary Design

In the notional MUSiC development lifecycle, usability testing is conducted only when a working system or prototype is available. Usability testing can, though, be conducted at earlier stages of system development. Heuristic evaluation (Nielsen, 1993), where usability experts judge the quality of design against a set of heuristics, may be performed when the system exists only as storyboards or a paper-based description. In addition to heuristic evaluation a number of usability inspection methods have been developed (Neilson and Mack, 1994), including the Cognitive Walkthrough. These methods may be applied early in system development without the need for a working system. In the HIBROWSE project, we have explored the use of the Cognitive Walkthrough method to evaluate designs and to suggest design revisions.

Cognitive Walkthrough is analogous to the program walkthrough method where source code is stepped through line-by-line and the changes of variables noted in an attempt to identify and remove "bugs". There are two phases of activity, in the first realistic task scenarios are generated and participants are required to construct a sequence of actions and commands which, if they were performed by the intended user, would result in the task being carried out. In the second phase, the sequence of actions that the user needs to carry out in order to perform some task is stepped through. For each action, the designers judge the success or failure of the user's attempts to articulate the correct command to progress toward completion of their goals, and to judge the system feedback in terms of their goals and judgement of which action should be the next performed. The simplified Cognitive Walkthrough method (Wharton et al., 1994) has proved capable of being learned and applied by system designers who have no training in
design for usability (John and Packer, 1995). This result was replicated, and the walkthrough method was successfully taught in two short sessions and applied by the HIBROWSE designers to evaluating the usability of the interface when performing a representative task.

In order to evaluate the usability of the initial HIBROWSE for EMBASE design, a walkthrough was conducted (Treglown, 1996b). The task scenario explored was one performed by a professional search intermediary working at a large university medical library observed during a small-scale field study conducted by one of the authors. The scenario conducted by the intermediary prompted by the information needs of a clinician was: Search for papers on the pathology of the kidney in children up to the age of 16 going back 20 years. List only review papers if a large number of references is found.

The HIBROWSE designers taking part in the walkthrough were able to construct an action sequence to perform the task consisting of 30 actions in approximately an hour. Walking through the action sequence took approximately a further two and a half hours. As only a limited demonstration version of HIBROWSE for EMBASE was available (see figure 2), a semi-formal specification of the system developed in the User Action Notation (Hartson et al., 1990; Treglown, 1996a) by the HIBROWSE designers was used to predict the system's behaviour when performing the task.

A number of usability problems, of varying perceived severity, were predicted as a result of conducting this walkthrough. It was predicted that users would have problems knowing when they had to identify a set of documents to act as a filter through the selection of the numbers in the "Refs" column, as distinct from the selection of a term to browse the hierarchy without some form of appropriate training. This, and other usability problems, revealed that it was unrealistic to regard HIBROWSE as a 'walk-up-and-use' system similar to vending machines or tourist information systems which are intended to be successfully used by the general public without interface skills, or necessary domain knowledge.

The need to make HIBROWSE compliant with the design guidelines of the user interface toolkit used to implement prototypes was also revealed when it was noted that little interface feedback was given in response to many user actions. Users might expect more indication of current system state and availability of user actions denoted by the shape of the on-screen pointer than is provided by the current specification of HIBROWSE.

A major problem revealed was the lack of any feature to generate additional views onto different subject (thesaurus) related attribute values in EMBASE records, corresponding to facets in the user query. The need for further design work on mechanisms to generate new views was noted by the designers. The EMTREE hierarchy itself suggested usability problems. It was felt that users would have problems navigating and trying to find terms within the EMTREE hierarchy, the difficulty of arranging hierarchies to suit the concepts and conceptual hierarchy held by users make this a difficult problem to tackle (Mehlenbacher et al., 1989; Lakoff, 1987; Cochrane, 1992).

In the initial design of HIBROWSE for EMBASE, only one view onto each selected attribute was specified. The walkthrough predicted this to cause usability problems. As the user selects sets of documents from a number of branches in the EMTREE hierarchy, when they move from one branch (and window of information) to another branch (and a different window of data), the original set of documents, and set of terms by which those documents are indexed, selected will no longer be visible. The limited capacity of user's working memory suggests that the list of terms used to generate the current view will necessarily be forgotten as the list of terms grows. Users are likely to reach a point where they can no longer predict the behaviour of HIBROWSE when generating new views because their understanding of the current system state will be limited, system behaviour could then appear non-deterministic.
5. The Language of View-based Searching

It is difficult even within a team which interacts regularly to secure a common appreciation of concepts and issues addressed in the pursuit of a new paradigm in information retrieval. Experience within the team ranges from many years to a matter of months and residue from the MenUSE models of retrieval, which sought to gather terms as a prelude to searching, still persist. The language of design should seek to support this paradigm shift, some of this language is outlined here:

**View**—A window containing one or more attribute values for a specific attribute indicating the number of document references possessing those values (document sets). **Finding** (a view)—typing initial characters from an entry term to access views directly rather than through browsing. **Selecting** (e.g. by highlighting) one or more document sets and identifying attribute value within a view. **Refining** (a view)—the operation which produces a new view containing only selected attribute values and document sets. **Expanding** (within a view)—opening a folder (using a single mouse click) to examine more specific attribute values and document sets. **Filtering** (automatically across views)—the process of deriving the number of documents in a set which possess additional attribute values determined by the current views. **Widening** (a view)—including more general attribute values with the current view (applicable when a view is accessed directly rather than through browsing (i.e. expansion and refinement) given there is no "previous view"

6. A Revised Design

The following screens provide some indication of the nature of the revised interface design with a search for *Phase 2 clinical trials of vinorelbine for treating lung cancer*. The

<table>
<thead>
<tr>
<th>Ref</th>
<th>Title</th>
<th>Ref</th>
<th>Terms</th>
<th>Subject</th>
<th>Ref</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A DNA variant at the angiotensin-converting enzyme</td>
<td>1</td>
<td>27 Wilkinson Street</td>
<td>1819377</td>
<td>1828</td>
<td>A'Hern R.</td>
</tr>
<tr>
<td>2</td>
<td>24, Val Princesse</td>
<td>2</td>
<td>1959411</td>
<td>1922</td>
<td>2</td>
<td>A'Hern R.</td>
</tr>
<tr>
<td>3</td>
<td>UD Harley Street</td>
<td>3</td>
<td>2795904</td>
<td>5592</td>
<td>3</td>
<td>Ahls I.C.</td>
</tr>
<tr>
<td>4</td>
<td>ABC Epidemiology</td>
<td>4</td>
<td>2933774</td>
<td>12600</td>
<td>4</td>
<td>Abdulwahab R.</td>
</tr>
<tr>
<td>5</td>
<td>An Unit of Radiology</td>
<td>5</td>
<td>3422141</td>
<td>2268</td>
<td>5</td>
<td>Abraham C.</td>
</tr>
<tr>
<td>6</td>
<td>Academic Dept.</td>
<td>6</td>
<td>4096566</td>
<td>2016</td>
<td>6</td>
<td>Abrams P.R.</td>
</tr>
<tr>
<td>7</td>
<td>Academic Dept.</td>
<td>7</td>
<td>5257921</td>
<td>441</td>
<td>7</td>
<td>Acharya S.K.</td>
</tr>
<tr>
<td>8</td>
<td>Academic Dept.</td>
<td>8</td>
<td>555752</td>
<td>435</td>
<td>8</td>
<td>60312 1991</td>
</tr>
<tr>
<td>9</td>
<td>Academic Unit on</td>
<td>9</td>
<td>1590998</td>
<td>525</td>
<td>9</td>
<td>24782 1994</td>
</tr>
<tr>
<td>10</td>
<td>Action Res. Place</td>
<td>10</td>
<td>1664272</td>
<td>452</td>
<td>10</td>
<td>223620 1993</td>
</tr>
<tr>
<td>11</td>
<td>Anaesthetic Dept.</td>
<td>11</td>
<td>824001</td>
<td>97</td>
<td>11</td>
<td>344605 1992</td>
</tr>
<tr>
<td>12</td>
<td>Ref Type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>82371 review</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7861 short survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>519 editorial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9631 preliminary report</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>94701 abstract report</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>127076 book</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2: HIBROWSE for EMBASE—Initial design
choice as to whether to use term look-up and direct access is made available to the user on selecting a view. Figure 3 shows the buttons for the top level views in the EMTREE thesaurus, and the top level view for "Diseases". Note: The numbers of documents are to demonstrate the effect of progressive filtering and do not represent the results of actual searching. The user has chosen to expand lung carcinoma in the lung cancer view in figure 5. The initial characters of vinorelbine were typed into the term list and navelbine identified as the preferred drug name and selected as a single term view within antineoplastic alkaloids. Instead of viewing the term phase 2 clinical trial, the user has chosen the wider view covering all aspects of human experiment.

It is interesting to compare the view-based interface with the search statements for the same query when searching EMBASE on the KR Dialog, STN International and Data-Star host systems:

**KR Dialog**
- s navelbine/de
- s phase 2 clinical trial/de
- s dc=c6.610.75.760.520.510
- s sl and s2 and s3

**STN International**
- s navelbine/ct
- s phase 2 clinical trial/ct
- s lung cancer + nt/ct
- s L1 and L2 and L3

**Data-Star**
- nabelbine.de.
- phase-2-clinical-trial
- lung-cancer#
- 1 and 2 and 3

---

**Fig. 3: Top Level View of Diseases**

**Fig. 4: View of Lung Cancer**

**Fig. 5: Views of Lung Cancer AND Navelbine**
7. Planned Experiments

Once the revised design has been implemented with searching functionality it will be possible to conduct traditional usability evaluations and experiments. The uses and costs of a number of approaches to evaluation traditionally employed in Information Retrieval research have been identified (Paton et al., 1994), but this work is by no means comprehensive and seems to dismiss approaches when the costs become no higher than would seem reasonable or usual to a usability analyst. Approaches to making usability analysts aware of the needs of cost/benefit analysis and business practice are receiving increasing attention in the HCI community (Bias and Mayhew, 1994).

The problems of attempting to employ the MUSIC methodology to Information Retrieval systems have already been discussed, particularly in attempting to classify users. When conducting evaluation of a new system, the difficulty of creating a plan of evaluation according to the MUSIC guidelines suggest that it may not be possible to conduct sessions according to the MUSIC method. Work on locating and picking terms from menus, skills that users will rely on when using HIBROWSE, has revealed that formal educational qualifications, gender and many other factors are not determinants of users’ success, instead spatial reasoning ability is. It may be, therefore, that rather than group participants by the MUSIC criteria, participants in usability evaluation sessions should be grouped by spatial reasoning ability in order to understand results. Contributions to research more useful that the above speculations will hopefully be obtained from further work.

8. Applications in Other Contexts

These techniques are applicable wherever there is a structured vocabulary or classification scheme able to provide views onto a database. Of particular interest in the world of academic libraries is the potential for exploiting the Library of Congress Subject Headings (LCSH), Dewey and Universal Decimal Classification hierarchies in a HIBROWSE style OPAC. These techniques are already being examined as extensions for user interfaces providing GUI client interfaces for traditional command-line host searching systems.

Acknowledgements

This research receives its principal funding from the British Library Research and Development Department and is co-sponsored by Sun Microsystems, Software AG (UK), Elsevier Science B.V., Elonex and the University of Huddersfield.
References


_Journal of Information Science_, 20(6), 413-426.

Shneiderman, B. (1992). Designing the user interface: strategies for effective human-computer 
interaction. 2nd ed. Addison-Wesley.

Treglown, M. (1996a). A specification of HIBROWSE for EMBASE in the user action 
notation. School of Computing and Mathematics, The University of Huddersfield 
Research Report RR 9601

Treglown, M. (1996b). Results of applying the cognitive walkthrough method to the 
HIBROWSE for EMBASE System. School of Computing and Mathematics, The 
University of Huddersfield Research Report RR 9602.

Description in the Electronic Environment

Abstract: The significant differences that exist between the print and digital worlds are sometimes felt to diminish the need for bibliographic description in the electronic world. An analysis of these differences, especially with respect to (1) the control of production and distribution of documents and (2) the need for software intermediation, coupled with a discussion of the functions of bibliographic description in the task of document retrieval argue, however, for an increased role for bibliographic description in the electronic world.

1. Introduction

The bibliographic world has long recognized the necessary role played by description in implementing bibliographic control. For example, the entries of an analytical bibliography display a high degree of concern for descriptive detail, and even now, the only real international standard in the bibliographic control arena is one dealing with description. Description has yet to blossom, however, in the shorter history of the electronic environment; for example, search engines on the World Wide Web (e.g., Lycos, Yahoo, Excite, InfoSeek, WebCrawler, Open Text Search)—the Internet equivalents of library catalogs—give quick access to the resources of the Web, but are short on meaningful description. This development evidences a lack of understanding of the functions played by description, functions that are at least as critical to successful retrieval in the electronic environment as they are in the print environment. My intent here is to explore the need for descriptive surrogates of document-like objects\(^1\) in the digital world.

This exploration will proceed down three related paths. First, I will examine the role of bibliographic description in a general retrieval model for document-like objects. This will establish the default assumption that every retrieval system needs to incorporate bibliographic description. Second, I will introduce the general range of options respecting fullness of description used in retrieval systems for both print and electronic materials. This will show that the need for bibliographic description in the electronic environment is not universally accepted; indeed, the assumption evidenced by most current Internet search engines is that true bibliographic description is unnecessary. Third, I will examine some of the significant differences between print and electronic documents, exploring how they impact the need for bibliographic description, given the retrieval model previously introduced. This will demonstrate that the need for bibliographic description is heightened, not diminished, in the electronic environment.

2. The Role of Description in a General Model of Retrieval

In the most basic information retrieval model, two input streams—one representing users and their needs, the other representing knowledge and information sources—are compared so as to bring relevant texts (i.e., those that can effectively help respond to an information need, original or evolved) to the attention of the user. In order to identify these relevant texts, we form representations of the users and their needs on the one hand and of knowledge and information
sources on the other. These representations may involve natural language terms in user statements and in documents, terms assigned from controlled vocabularies to represent the aboutness of user queries and documents, or classes of related queries or documents, etc.

We portray the ideal retrieval system as retrieving all, but only, the documents that a user would select were he or she to personally inspect all documents available. In the past we tended to envision this ideal retrieval system operating in a single-pass mode, but increasingly we have come to recognize that truly effective information retrieval is both iterative and interactive. Rather than picture a one-step retrieval process in which the retrieval system does all the work (after user needs and documents are captured by some sort of representation), we must revise our model to show retrieval to be a multi-step process. In the initial step, the retrieval system identifies a set of potentially relevant documents. In a subsequent step, the user de-selects documents that, although they match certain criteria, appear not to be useful to that particular user, at that particular time, in that particular situation, for that particular need. Alternatively, in the second step, the user identifies especially relevant documents. Various characteristics of these documents become clues used in further searching.

Where does bibliographic description fit in this picture? In answering this question, I will take the term "description" to mean something a little broader than what we might normally assume bibliographic description to refer to. I will take description to be nothing more or less than a factual representation of something, especially if given in words. A bibliographic description is then, in its broadest sense, whatever representation of document-like objects is used in a retrieval system. The term is especially relevant when the representation is verbal. Further, it is assumed that the description is a level removed from the object of interest, since representation implies a substitution of one object for another. Thus, a document-like object cannot serve as a surrogate for itself; the representation/description is something other than the thing represented/described.

Although we often make a distinction between description and access—as when we distinguish between descriptive cataloging and subject cataloging—the above presentation should make it clear that the initial retrieval stage, in which various access points (e.g., name, title, subject) are used to retrieve potentially relevant documents, makes critical use of description. This is especially true regarding the use of subject access points, which are chosen to describe salient topics, issues, or needs addressed by a document. Perhaps we hesitate to speak of access points as involving description because we associate some degree of verbosity with description, while access points are the more efficient, the more compact they are. Nevertheless, access points do represent salient features of document-like objects and therefore qualify as elements of description. All access points stand as surrogates for their respective documents, quite in keeping with their representation here as descriptive elements.

A wider array of descriptive elements usually comes into play in the filtering stages, in an effort to enable the user to distinguish relevant documents from non-relevant documents within the set of retrieved documents. The contrast between the number of descriptive elements used for access purposes and for filtering purposes appears to be a function of the disparate roles played by different elements of description. Some, especially author, title, and subject, help define the potential relevance of a document. Many others, for example, date of publication, language, length of document, intended audience, help define why a particular document will not be useful (i.e., why the user may not find it relevant). Such descriptive elements are typically used to qualify searches, to screen out that which is unlikely to be relevant. In other words, when elements such as language or date of publication are used alongside name, title, and/or subject by the retrieval system, they almost always are used negatively, to screen out documents, not positively to retrieve
Within the contexts of retrieval and filtering, three particular functions of description are important: identification, selection, and the expression of relationships. Some retrieval operations are for known items. Those parts of the description involving identification of an item should enable the user to ascertain whether the item is or is not the known item being sought (which may in reality be only partially known). Other retrieval operations are undertaken without foreknowledge of the existence of relevant documents. In such a case, the user will need to be able to discern if a specific item retrieved is indeed a relevant item; where there are multiple relevant items, the user will also want to be able to judge which of them are most relevant. Both the identification and selection functions may be supplemented by the third function of expressing relationships between the item being described and other items. From the perspective of identification, the issue may be whether the item is the appropriate edition or version of the sought-for document; from the perspective of selection, the issue may be whether the item can be used in conjunction with other material (e.g., is the concordance to an available edition?; is the file in a format readable using available tools?).

3. Levels of Description

In implementing this general model of retrieval in the print world, different levels of bibliographic description appear to be warranted for different contexts. For example, in the world of AACR2/ISBD, a school library might adopt an abbreviated first level description as its default, while an academic library might adopt a second level description as the default for its general collection, but a full third level description for its special collections.

Bibliographic description in the electronic world similarly admits of varying levels of description. OCLC's InterCat project, which is building a database of bibliographic records for Internet resources, has adopted the traditional level of bibliographic description based on AACR2/ISBD and USMARC. The TEI community has gone a step further by encouraging the inclusion of a "header" file to accompany a digital text file, which includes the descriptive information from which an AACR2/ISBD-based description can be derived. These two phenomena represent the fullest level of description of electronic documents now in vogue. A middle level of description is represented by the "Dublin Core," a fairly small set of descriptive elements—author, title, subject, publisher, date, other agent, object type, form, relation, source, language, coverage—proposed for use in describing document-like objects (DLOs) on the Internet. A third position is advocated by those who regard electronic texts as self-documenting, such that no separate description is needed; the immediate availability of the text in the digital environment is considered as obviating the need for a document surrogate. Such a position underlies the numerous Internet search engines currently in operation and represents the de facto standard position on bibliographic description in the electronic world. In this view, access to documents can be effected entirely through surface characteristics of the document (and perhaps those of other documents). Thus, the time and effort that would otherwise go into generating bibliographic descriptions can be saved.

The seeming parallel between AACR2's definition of three levels of description and the presentation here of three approaches to level of description is deceiving. In the AACR2 world, the three levels of description exist along a single continuum, varying in terms of fullness of description. In the digital world, only the first two positions vary along the fullness continuum. The third position drinks from another trough, in that it views documents (or, as is currently the case with the typical Internet search engine, document incipits) as being able to achieve the same ends as bibliographic descriptions. The only sense of representation, and hence description, that
takes place here occurs when part of the document (the beginning, or incipit, of the document) stands for the whole document. The most universal attempts to catalog the Internet are currently based on the unexamined assumption that bibliographic description (surrogation) is unnecessary for effective retrieval when full access to document-like objects is readily available.

4. Significant Differences between the Print and Electronic Environments

The unspoken assumption behind Internet search engines (and other retrieval systems for the electronic world that eschew the use of document surrogation) is that the differences inherent in the change from print to electronic environments make the use of bibliographic description unnecessary. We can only see if this assumption makes sense by examining these differences.

The first difference is the great speed advantage maintained by the electronic world over the print world in affording access to document-like objects. Ironically, in the content of our retrieval model, the use of bibliographic descriptions in the filtering process, i.e., for the making of relevance judgments, currently has a practical speed advantage over the use of direct access to documents. If, for example, one performs a search through an Internet search engine such as Lycos and compares the amount of time it takes to access a next "description" (here a document incipit) in the Lycos output to the time it takes to follow a URL to the actual site, the average difference in access time would overwhelmingly favor the use of descriptions for the making of relevance judgments. Depending on network load, there can be a significant time difference in favor of seeing the next document surrogate over accessing the next document. We should note that in time the delay inherent in making connections across the Internet is likely to become shorter; the significant difference in access times that holds currently may eventually taper off into insignificance, but then again it may not. This discussion does not, of course, eliminate the real advantage the electronic world has in retrieving documents for use. Where the overall speed advantage lies—whether with the quicker retrieval times associated with making relevance judgments based on document surrogates or with the quicker retrieval times associated with actual access to documents—will depend not only on the difference between the speeds of those two operations in the two environments, but also on the number of documents retrieved (or at least the number of documents that the user deals with) and the precision of the retrieval (since the filtering out of irrelevant documents based on bibliographic descriptions eliminates the need to access the document at all).

Another area in which the electronic environment differs from the print environment concerns the production and distribution of documents. Production and distribution of documents operate in the print environment in such a way that document variants are fairly strictly controlled. Only a limited number of production points exist, at least in comparison to the number of production points in the electronic environment. Moreover, they distribute a limited array of types of documents. This puts a cap on sorts on the number of items to be controlled in the print environment. Document-like objects in the electronic environment are, however, currently proliferating like proverbial rabbits: over the course of the first three months of 1996, the number of URLs on Lycos nearly doubled; in the process of that growth, the Lycos collection of URLs surpassed the number of bibliographic records in the OCLC Online Union Catalog. Admittedly, the prototypical bibliographic record in OCLC represents a larger document than is represented by the prototypical URL in Lycos, but in one regard it is only the absolute number of document-like units being controlled that is at issue here.

This proliferation of document-like objects reflects the fact that not only will many more distinct resources be made available electronically than in print, but also that many more versions
of at least some resources will be made available. The differences between these versions, or variants, may be harder to discern than in the print world. If, for example, one has two print documents whose title pages agree with each other (including title, statements of responsibility, edition, and publication details), one can assume that their two texts will be identical. One cannot, however, make the same naive assumption with electronic documents, since they are so easily reproduced and altered.

The differences between the print and electronic worlds along the issue of reproduction and alteration are many. To start with, documents are generally not made widely available in the print world until their content is relatively stable. Drafts of paper documents, for instance, enjoy only limited circulation; not until the document has completed some number of rounds of revision will it be made widely available. Drafts of electronic documents, in contrast, may be made universally available much earlier in the revision process. Moreover, once a document has been made available in print, the issuance of a new version or edition generally awaits the accumulation of many changes. In the electronic world, however, new versions are routinely produced for even the most trivial of changes; since it is often easier to make a change to a digital file than to keep track that a change needs to be made to it, a new document version may frequently result from the need to make but a single change. Finally, when new print versions of documents are published, they usually come into existence alongside copies of older versions. The differences between the two can be discerned through direct comparison. While this pattern can also easily apply to the electronic world, as is the case when changes are made to copies of originals, it is just as likely that document revisions will result in newer versions altogether replacing older versions, not just in supplementing them. In summary, the revision of print documents is usually associated with stability, cumulativity of changes, and supplementation; the revision of electronic documents is much more likely to take place in the context of continuous change, in small increments, replacing previous versions. The process of revision thus typically impacts the print and electronic worlds in very different ways.

Another factor that differentiates between the print and electronic environments is the need to use special tools (including both hardware and software) to access and display documents in the digital world. This situation contrasts markedly with the print world, where eye-readable documents are instead the norm for both textual and visual materials. In other words, the document-as-stored in the print world prototypically exists in the same eye-readable sphere as the document-as-displayed. In the digital world, however, the document exists in dissimilar spheres: the base sphere is one of digital encoding, where the document-as-stored is a file of some number of bytes on a digital storage medium, while the modified sphere of the document-as-displayed may exist in time and/or space. The presence of software as an intermediary between the document-as-stored and the document-as-displayed in the electronic environment has as one possible result that the document-as-displayed may not be strictly equivalent to the content of the document-as-stored, as is the case when software is used, not merely for the purposes of accessing and displaying the digital document, but also for the purposes of manipulating it. For example, the document-as-stored may simply be a data set, but the document-as-displayed may be a map or a chart that communicates the data in graphic rather than textual form. The resultant visual image would be in some sense nothing other than another version of the data set, although the value added through the visual interpretation also suggests that it is more than the data set.

5. The Need for Description in the Electronic Environment

How do these differences between the print and electronic environments impact retrieval?
The proliferation of sources, many of which are interrelated by being versions of the same works, requires finer distinctions to be made if individual sources are to be uniquely identifiable. To borrow database terminology, this can either be done by assigning an arbitrary key attribute to all document-like objects, so that each object has a unique identifier for this attribute, or by recording further descriptive data for each object, the combination of which will be unique. While the first alternative is the more efficient for the purposes of identification, it overlooks the opportunity to supply meaningful identifying information that will show relationships between objects; moreover, it is unlikely to be a feasible solution, seeing we have been only partially successful in implementing such a system (i.e., ISBN) in the more controlled print environment. The identification issue is made even less manageable by the fact that we can regularly expect some percentage of documents to which descriptions refer to disappear completely, for example, through replacement. Unless descriptions are full, it may be impossible, given a description of an item that no longer exists and faced with an existing version of a variant of the item, to determine that the description does not match the item in hand.

The selection function of description is also more difficult in the electronic environment. Given the proliferation of electronic sources, it will be more and more common for searches to yield numerically overwhelming results. Whether the retrieval system is used to improve this condition—for example, through ranked output or through using additional search clues to narrow the results—or whether the user must filter out irrelevant material, a finer level of descriptive detail is needed than if retrieval results are not numerically overwhelming. Not only is this fuller description needed for sifting through the output to determine which objects are relevant, i.e., satisfactory to the task, but it is also needed—for identifying which document-like objects are best for the task.

Finally, the relationship function of description is also adversely affected by the transition to the digital world. The proliferation of variants complicates the task of expressing relationships across works. For example, does a document at one location exactly mirror a document that purports to be the same at another location? If not, how are they related? Further, is the one document identical to the document with the same name present at that location a day/week/month ago? The kind of proliferation that causes these kinds of complications arises from the lesser control on production and distribution in the electronic world. The presence of software intermediation in the digital world brings about another kind of proliferation that also puts strain on the expression of relationships. This is the kind of proliferation that results from taking a single text and representing it variously as, for example, an ASCII text, a PostScript file, a WordPerfect document, a troff document, etc. Incompatibilities between software products and across operating system platforms may require the generation of document variants at the document-as-stored level that are equivalent at the document-as-displayed level. Again, the expression of these types of relationships will require greater descriptive detail.

6. Conclusion

If it is as important in the electronic environment as in the print environment to be able to uniquely identify document-like objects, to be able to select from among numerous possibly relevant documents those most useful for the task at hand, and to know what, if any, relationship exists between two document-like objects, then bibliographic control should play as important a role in the electronic environment as in the traditional print world. An examination of salient differences between the two worlds suggests numerous ways in which the bibliographic control of digital objects is more complex than that of print objects. Hence, the need for bibliographic
description, which is at the heart of bibliographic control, is not diminished in the electronic world; rather the need for full description of electronic objects is elevated.

Notes

1. The use of the term "document-like object" is meant to suggest that the discussion throughout is more generally applicable than in the traditional text world. Although the term "document" alone will also be used in the paper, it is not meant to be distinct in meaning from "document-like object," but is used simply for purposes of diversity and efficiency. (See note 6 for the origin of the term.)

2. In giving this description of "description," I rely heavily on the dictionary's expertise, where the first meaning given for "description" is "an act of describing" and where the first meaning given for "describe" is "to represent or give an account of in words" (Merriam-Webster's Collegiate Dictionary, 10th ed., s.v. "describe" and s.v. "description").

3. The reader may counter that such elements are indeed often used positively, for example, to retrieve documents published after a certain date or to retrieve documents in a certain language, and I will counter in turn that the user rarely wants documents after a certain date or in a certain language per se, but wants to exclude documents from before that date or written in other languages; the fact that language or date qualifiers may be added to the search with AND rather than NOT is not necessarily indicative of whether they are being used positively or negatively, where positivity refers to including something for what it is and negativity refers to excluding for what it is not.


5. The TEI guidelines can be browsed at http://etext.virginia.edu/TEI.html. Material on the header is located in part 2, chapter 5.

Online Classification: Implications for Classifying and Document[-like Object] Retrieval

Abstract: Recent advances in automating library classification data are causing the makers and maintainers of DDC and LCC to seek enhancements to these schemes that will facilitate computer-assisted classifying and extend the usefulness of library classification to electronic collections. To explore the kinds of improvements that are needed to ensure that classification remains a viable subject retrieval tool, Dewey Decimal Classification numbers applied to over 40,000 records for Internet-accessible resources were analyzed. This analysis is used as a springboard for recommendations for improvements to DDC and LCC.

1. Introduction

Recent advances in automating library classification data which include the use of online editing systems by Forest Press and Library of Congress (LC) to produce recent editions of the Dewey Decimal Classification (DDC) and selected Library of Congress classification (LCC) schedules and the development of automated classification assistance tools like Dewey for Windows and LC’s Classification Plus are causing the makers and maintainers of these systems to seek enhancements to the schemes that will facilitate computer-assisted classifying and extend the usefulness of library classification to electronic collections. Mitchell (1996) describes how new uses of the DDC are driving structural and content changes to its underlying database. Even without enhancements, DDC, LCC, and UDC are increasingly being used to provide access to non-commercial Internet-accessible collections. An example of each is (1) Morton Grove Online WEBrary¹ (DDC), (2) CyberStacks(sm)² (LCC), and (3) NISS Information Gateway³ (UDC).

If additional motivation for supporting continuous improvements to traditional library classification schemes is needed, one need only look at the popularity of Yahoo and Infoseek, Internet search services that provide category-based subject approaches to vast quantities of electronic resources, to find it. These Internet-inspired classification schemes paired with powerful search engines have the potential to supplant library organizing techniques or at least to provide parallel approaches that become the preferred access method for electronic data leaving library classification to organize traditionally published materials.

To explore the kinds of improvements that are needed to ensure that DDC and LCC classification remain viable subject retrieval tools, Dewey Decimal Classification numbers applied to over 40,000 records for Internet-accessible resources were analyzed. This analysis is used as a springboard for recommendations about needed improvements to online classification data.

2. NetFirst Database

The source for the DDC numbers analyzed in this study is the OCLC NetFirst™ database. NetFirst records contain descriptive information, summaries, subject headings and Dewey classification numbers for over 40,000 Internet-accessible resources. The database coverage includes World Wide Web pages, Gopher servers, electronic discussion lists, library catalogs, and electronic publications in other formats. A typical NetFirst record is shown is Figure 1.
INTERCAT is a mailing list covering the topic of the OCLC Internet Cataloging project, of the Online Computer Library Center. Welcomes members' questions, discussion topics, examples, methods, and solutions relating to identification, selection, and cataloging of Internet resources. Membership in the group is available to anyone. The discussion archives cannot be searched. The organization responsible for running this group is OCLC, Inc., Dublin, Ohio USA. Their server name is OCLC.ORG. To join the group, send the message "sub INTERCAT your name" to listserv@oclc.org. You can reach a human list administrator for the group by emailing to intercat-request@oclc.org. To send mail to the entire group, send it to intercat@oclc.org.

Archives: can't be searched.
Number of participants: 2846.
Place of publication: Dublin, Ohio, USA
OCLC Online Computer Library Center, Inc.
Internet (Computer network).
Cataloging.

Fig. 1: A typical NetFirst Record

Note the use of multiple Dewey numbers in this record. Since Dewey is not being used to locate a document in a physical place, but to assign a document to a location within the DDC topic structure more than one number may be applied if appropriate. This means that a user presented with a Dewey topic browser as an interface to the NetFirst collection could discover the record for the Intercat mailing list through the Library and information sciences (025) hierarchy or through the Computer science (004) hierarchy. Care must still be taken to apply numbers consistently to like topics so that users are not confronted with the situation that the same topic is entered under different numbers at different times. Consistency in number application will also ensure that Dewey's hierarchical structure and established relationships can be used to maximum benefit when leading users to related subjects.

3. Use of Dewey in the NetFirst Database

To get a sense of the makeup of DDC number assignments in the NetFirst database a frequency distribution, by Dewey main class, of uniquely assigned class numbers with percentages was statistically generated (Table 1, columns 2 and 3). Next, the number and percentage of assignments per main class was tabulated (Table 1, columns 4 and 5). DDC classes 000 Generalities, 300 Social sciences, 500 Natural sciences & mathematics, and 600 Technology account for over 80% of unique classes and class number assignments in the NetFirst database.
A list of class numbers ordered by frequency of usage was also generated to aid in understanding DDC use in NetFirst records. The most frequently assigned Dewey number was 004.67 Wide-area networks with 545 assignments followed by 378 Higher education with 478 uses. Eight of the remaining 12 class numbers that had been applied more than 100 times were from main class 000, three from 300 and one from 600. With these data in mind class numbers from Dewey class 000 was analyzed more closely. Class numbers in the 004-006 range (Computer science) were found to account for over 42% (1,283) of unique class numbers assigned from class 000, but more significantly they account for about 6.1% of all unique classes and nearly 7% (3,126) of all class number assignments. These numbers found at the third summary level of Dewey were used more often than class numbers based on each of the main Dewey classes 100 (0.82%), 200 (1.48%), 400 (0.91%), 800 (1.13%), and 900 (4.59), and nearly as often as class numbers based on class 700 (7.09%).

4. Improving Online Classification Data for Classifiers and End Users

The uneven distribution of class numbers in the NetFirst database suggests that browsing interfaces that provide access to classified electronic collections need to be flexible or even dynamic in their methods for presenting classification system hierarchy structures. It is not difficult to envision an interface to the NetFirst database that promotes Computer science and Education to the top-level categories displayed to users. For databases of Internet resources, given the volatile nature of such objects, the interface could adjust the categories and subcategories displayed to users after querying the database for frequently occurring classes. Experimentation is need to determine optimal numbers of top, mid, and lower level categories and how much volatility users can tolerate in a subject-oriented interface.

In addition to guiding the display of topics in interfaces, statistical data on class number use can be quite valuable to help focus efforts to improve online classification data in other ways. When LCC and DDC are discussed as subject access tools the observation is often made that both

<table>
<thead>
<tr>
<th>DDC Main Class</th>
<th>No. of unique Classes</th>
<th>% of Unique Classes</th>
<th>No. of Uses</th>
<th>% of Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>000 Generalities</td>
<td>3,092</td>
<td>14.71</td>
<td>8,738</td>
<td>19.29</td>
</tr>
<tr>
<td>100 Philosophy &amp; psychology</td>
<td>255</td>
<td>1.21</td>
<td>373</td>
<td>0.82</td>
</tr>
<tr>
<td>200 Religion</td>
<td>430</td>
<td>2.05</td>
<td>672</td>
<td>1.48</td>
</tr>
<tr>
<td>300 Social sciences</td>
<td>6,929</td>
<td>32.96</td>
<td>15,612</td>
<td>34.47</td>
</tr>
<tr>
<td>400 Language</td>
<td>226</td>
<td>1.08</td>
<td>414</td>
<td>0.91</td>
</tr>
<tr>
<td>500 Natural sciences &amp; mathematics</td>
<td>2,327</td>
<td>11.07</td>
<td>4,591</td>
<td>10.14</td>
</tr>
<tr>
<td>600 Technology (Applied sciences)</td>
<td>4,666</td>
<td>22.20</td>
<td>9,093</td>
<td>20.08</td>
</tr>
<tr>
<td>700 The arts</td>
<td>1,698</td>
<td>8.08</td>
<td>3,209</td>
<td>7.09</td>
</tr>
<tr>
<td>800 Literature &amp; rhetoric</td>
<td>265</td>
<td>1.26</td>
<td>511</td>
<td>1.13</td>
</tr>
<tr>
<td>900 Geography &amp; history</td>
<td>1133</td>
<td>5.39</td>
<td>2,079</td>
<td>4.59</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>21,021</strong></td>
<td><strong>100.00</strong></td>
<td><strong>45,292</strong></td>
<td><strong>100.00</strong></td>
</tr>
</tbody>
</table>

Table 1: Frequency and % of Unique DDC numbers and Class Number Assignments in the NetFirst Database
systems contain caption headings that use general terminology or descriptive terminology more appropriate for classifiers than end users. Although it may not be practical or even desirable to evaluate all captions for expressiveness and currency, those captions that are associated with high use class numbers might be candidates for transformation. Figure 2 shows a prototype Internet browsing display with edited captions for portions of the Dewey 21 Education outline.

Fig. 2: Browse Display of Dewey 21 Education.

In this example, the term “Sororities” is used in place of the Dewey caption “Women’s Greek-letter social societies”. In this case, the substitute caption is a term from the Dewey Relative Index, but alternative captions need not and should not be limited to the index terminology of the classification system at hand as long as they are coded and linked appropriately in the underlying database. Following Barbara Tillett’s (1995) access control model, where catalog users are allowed customized choices concerning the form of name they wish to see, classifiers and end users could be presented with custom interfaces to classification data based on language, indexing terminology, or level of specificity required. This is particularly true for Dewey because it is available in full and abridged forms, has been translated into several languages, and contains links to other thesauri.

Mitchell (1996) describes how Library of Congress Subject Headings are being mapped to Dewey numbers in the DDC 21 database. These headings will be included in the Dewey for Windows CD-ROM database to serve as a source of additional entry vocabulary. In the future they may be used to support alternate or customized views of DDC. For LCC, explicit links between LC subject heading and class numbers occur in LC Subject Authority records that contain LC classification number fields. In an analysis of the LC Subject Authority file Vizine-Goetz and Markey (1989) found that about 43% of topical subject heading records (MARC tag 150) contain LC classification number fields. Science and technology classes account for almost half (47.72%) of the LC class numbers. Efforts to improve the index to LCC may also lead better links between LCC and LCSH.

In addition to mapping LCSH to DDC, the Dewey editors are experimenting with coding the nature of the relationship between the class number and subject heading(s). The ability to
accurately describe these relationships will determine whether successful classifier- and retrieval-assistance tools can be developed. A recent experience with an Internet search engine that provides suggestions of related topics to search under illustrates the importance of making accurate and reliable links between terms and classification structures. A search for “California red wine” retrieved many relevant documents, but also identified the names of sports teams and elementary and middle schools in California in the list of related topics. This example shows that no linkages are probably better than bad ones.

5. Conclusion
In this paper several suggestions for enhancements to online classification data have been made, including
1. Developing end user interfaces that employ adaptive methods for displaying classification hierarchy structures,
2. Transforming selected DDC and LCC captions to end user language, and
3. Using links to other subject thesauri, language versions, and full and abridged forms of classifications to provide custom views of online classification data.

By enacting enhancements of this type, the makers and maintainers online classification data move library classifications systems significantly forward in the effort to ensure that classifiers will be provided with advanced classifying tools and that library classification will play a major role in organizing existing and future document[-like object] collections.

Notes
3. http://www.niss.ac.uk

References
sequential and unified collection of World Wide Web (WWW) and other Internet resources, that uses the Library of Congress classification scheme as an organizational framework, was formally established on the home page server at Iowa State University in Fall 1995 (McKiernan, 1995).

The idea of creating an organized collection of selected Internet resources using an established classification system originated in a series of listserv and newsgroups postings that began in summer 1994 (McKiernan, 1996a, 2). Many of the issues raised in the associated discussions were coincidentally considered by Britten in his excellent review paper on building Internet collections prepared for the 1994 Charleston Conference (Britten, 1995). His articulation of the professional and institutional role that libraries and librarians have historically played in organizing information provides clear support for the benefit that a 'library-organized' virtual library could provide users seeking authoritative, significant and well-described Net resources:

"[T]he long-term institutional perspectives of libraries, librarians' expertise in information organization, and the library profession's organizational structure . . . would add both collection management expertise and permanence [to such a project]." (Britten, 1995, 247).

Britten's profile of the characteristics of networked information, namely the potential of a
network 'copy' serving the needs of many libraries simultaneously, and the transient and ephemeral nature of some sites and resources, highlights both the value that a centralized collection such as CyberStacks(sm) might generally offer to research and university libraries, and the potential role that librarians could play in the maintenance of access to these electronic sources (Britten, 1995, 249).

The establishment of a central Virtual Library Collection, coordinated collection development of Web and other Net resources using the Internet itself (Britten, 1995, 247) and the application of a widely-used library classification scheme are among the features derived from established library practices and procedures that have been incorporated within the CyberStacks(sm) model of a virtual library.

2. Selection

CyberStacks(sm) seeks not to be a digital library per se, but a true 'virtual' collection. The majority of resources selected for incorporation within its collection are monographic or serial works, files, databases or search services, and all of its selected resources are full-text, hypertext or, hypermedia, and of a research or scholarly nature. CyberStacks (sm) may indeed be considered the service envisioned by Tyckoson in his outstanding review article on the history and future of the catalog:

"The catalog of the future may not be merely an index to the world's collected information, but may actually provide a summary of the information for the user. A catalog that includes full-text works [emphasis added] on a variety of subjects as well as abstracts and citations to materials may serve as something more than an encyclopedia or index. Whereas the scholar of the eighteenth century used two tools to find out about his or her world, the scholar of the twenty-first century may need only one [emphasis added]." (Tyckoson, 1991, 25).

In its selection of resources for incorporation, CyberStacks(sm) has adopted many of the same philosophies and guidelines used in the selection of non-Internet resources by most librarians in the development of local collections, including authority of the source, accuracy of information, clarity of presentation, uniqueness within the total collection, recency or timeliness, favorable reviews, and community needs (American Library Association, Reference Collection Development and Evaluation Committee, 1992).

The foundation of the CyberStacks(sm) collection is a selection practice that seeks to identify, isolate, and select Internet resources that are discrete document-like objects. The philosophy of selection of resources for inclusion within the CyberStacks(sm) scheme has been well-articulated by Demas, McDonald and Lawrence in a recent article:

"Title by title selection of high quality resources is one of the most important values librarians can add in providing access to information resources, including those accessible via the Internet. A careful selection of resources is the touchstone of the electronic library" (Demas, McDonald and Lawrence, 1995, 280).

To facilitate the selection of such resources, in the initial phase of the project, we have intentionally sought to locate Internet resources that are the electronic equivalent or analog of standard Reference works, as well as resources that may be so considered. Abstracting and indexing services, bibliographies, biographical sources, dictionaries, directories, handbooks, guides
to the literature, maps, and standards, are but a few of these publication types. We believe that our intentional selection of sources of a Reference nature that serve to index, abstract, guide or review significant and relevant Internet resources, will mitigate some of the frustrating and distressing experiences of users who seek specific data and information via the Net. In general, through CyberStacks(sm) we have sought to provide the value-added service offered by libraries for generations in the careful selection and description of resources. In particular, we have made a systematic effort to identify and incorporate significant, yet elusive sources from significant Web sites (McKiernan and Ames, 1996).

The hypertext nature of the World Wide Web (WWW) and the various browsers used to navigate through it enable and facilitate the inclusion of an entire collection from a site, or only those selected resources considered most relevant. We recognize, as did Britten (Britten, 1995, 248), that the hypertext functionality of the Web permits one to incorporate selected appropriate portions of a site, or its entire collection, into a virtual library. While CyberStacks(sm) does include entire Web sites, only those that are considered well-organized collections of significant Internet resources are selected for incorporation. Such sites are viewed as either directories or encyclopedias, depending on their structure, organization and content, and are categorized and classified as such. Unlike most other media, an individual resource or site may be classified in more than one broad or specific category within a particular organizational scheme, presenting conventional as well as novel opportunities for enhanced access.

3. Library of Congress Classification System

The Library of Congress classification system is a well-established scheme that has been used by libraries worldwide for organizing a variety of publications and media for generations. Within its schedules, this classification system not only denotes subject coverage and content, but information format and conceptual relationships as well. Underlying the CyberStacks(sm) approach is the use of a hypertext outline of the Library of Congress classification scheme (LC Classification Outline, 1990) that allows users to scan broader or narrower subject categories at will and then, if desired, to connect directly to a selected resource.

While an increasing number of sites are adopting the Library of Congress classification scheme as a system for organizing WWW and other Internet resources, most notably the WWW Virtual Library, specific relationships to other resources are typically not indicated. Generally, selected resources are classified only at a broad level, and only listed alphabetically within particular classified groups. While a listing within a broad category does offer the user some assistance in identifying relevant resources, such an arrangement requires a user to review an entire list to be assured that an appropriate resource has not been overlooked.

Unlike most Web sites that have adopted the Library of Congress classification system, CyberStacks(sm) has applied the scheme at a deep level of specificity and description, incorporating resources within a hypertext framework that allows users to browse a collection of resources in order by an assigned classification number. In establishing CyberStacks(sm), we believed that a more comprehensive application of the Library of Congress classification system could offer the value-added structure, organization, context, and an appropriate level of specificity and description for enhancing the efficiency and effectiveness in identifying and using relevant and related WWW and other Internet resources. Indeed, the structure, organization and access provided by CyberStacks(sm) can be viewed as the convergence, merger and fusion of a subject-ordered catalog and an associated collection, where like materials metaphorically 'stand together' in browsable stacks.
4. Browsing

We believe that the use of a standard organizational scheme provides a framework that facilitates the predictable location of resources within an overall system and substantially lessens the cognitive load and psychological burden often associated with information overload and anxiety.

Within CyberStacks(sm) resources are categorized first within a broad classification, then within narrower subclasses, and finally listed under a specific classification range and associated subject description that best characterize the content and coverage of the resource. While we recognize the need to provide alternative methods of accessing relevant resources to suit the variety of information-seeking strategies of a diverse user population, browsing has been adopted as the primary means of identification and access in the initial implementation phase.

We agree with Chang and Rice that:

"Browsing is common... We all browse in various contexts to make sense of the world around us, such as when we read newspapers, scan television channels, go window shopping, or seek information in libraries [emphasis added]" (Chang and Rice, 1993, 232).

In citing the work of Herner, Bates, and others, they recognize the importance of browsing in human information seeking and highlight the need for information scientists to incorporate this approach in their system designs (Chang and Rice, 1993, 232-233).

Their summary of critiques of 'specific, direct searching' supports the need for browsable interfaces:

"[There are] some unrealistic assumptions about users and the nature of information seeking—e.g., that users have unbounded rationality, have static and well-defined information needs, know what they want and are output oriented..." (Chang and Rice, 1993, 233).

The clear articulation of the uncertainty that some users experience in the search process further supports this option:

"Users are often in an anomalous state of knowledge as they initiate a search..., do not have predefined search criteria, and may alter their interests during a search..." (Chang and Rice, 1993,233).

Their observations of end-user searching is particularly relevant to the decision to organize selected resources within a browsable, categorized scheme:

"Many end-user systems are difficult to use because they require training, knowledge of mechanical and conceptual aspects of searching, and a high cognitive load from end users... [A]ll of these may be overcome through browsing..." (Chang and Rice, 1993, 233).

Within CyberStacks(sm), users are presented with a variety of broad, narrow and related categories to browse. From a main menu, users may select from an array of primary Library of Congress classes (i.e., Q, R, S, T, U or V) or their associated subjects (i.e., Science, Medicine, Agriculture, Technology, Military Science or Naval Science). As a service to users, only classes which have subclasses with incorporated resources are hotlinked. Within a class (e.g., Science (Q)), users may select from a table of associated subclasses (e.g., Q, QA, QB, QC, QD), each with
an associated subject or topic (e.g., Science (General), Mathematics, Astronomy, Physics, Chemistry). For any subclass, users are presented with a table containing the range of associated class numbers (e.g., QB 1-139, QB 140-237, QB 275-343, etc.) and their associated subjects (e.g., Astronomy (General), Practical & Spherical Astronomy, Geodesy, etc.). At this latter level, users are presented with an option of browsing any (or all) of the related classes with a main class (e.g., Q, QA, QB, QC, QD) at the bottom of the table of class numbers.

At the top of every subclass table and resource profile is a horizontal menu of broad subject classes that permits users to access any class currently incorporated with the CyberStacks(sm) scheme (i.e., Q, R, S, T, U, and V).

6. Record Format

For each resource incorporated within CyberStacks(sm), a profile of its coverage, content and special features is provided. For each record, when and where possible, we have excerpted the summary data from the original resource to provide sufficient information about its subject coverage, and other features, to enable users to judge a resource's potential usefulness. Our focus is not to analyze a resource and delineate each of its potentially important data elements, but to characterize it. We believe that the meaning and value of a given resource can be conveyed as well, or better, by structure and organization. We believe that the CyberStacks(sm) scheme provides meaning through context; a context that is appropriate and relevant to users while seeking a resource which may satisfy an information need.

Although we have decided not to standardize the format of data within a record, many of the categories recommended for describing a resource within the CyberStacks(sm) scheme are identical or similar in function to the metadata elements of the Dublin Metadata Core Element Set, the Dublin Core. While the Dublin Core seeks to identify and isolate the data elements of an Internet resource, the summary information provided for a selected resource within CyberStacks(sm) is intended to describe the resource only to the extent that users can judge its potential value for meeting an information need.

While we prefer to describe resources within CyberStacks(sm) more holistically than analytically, we understand the desire to identify and define core data elements. Although we do not believe that an exhaustive, analyzed set is necessary for use within our scheme, we believe that a set such as the Dublin Core can provide the optimal level of descriptive and subject cataloging for World Wide Web (WWW) and other Internet resources that will facilitate their incorporation within emergent as well as conventional information and retrieval systems and services. We also do not believe that the holistic approach taken by CyberStacks(sm) and the more analytical approach represented by the Dublin Core are mutually exclusive. Indeed, many of the planned enhancements for CyberStacks(sm) will require the manipulation of delineated data elements to be realized.

7. Planned Enhancements

We recognize that there are inherent limitations in providing a browse function within CyberStacks(sm) with limited visual context. The addition of images and appropriate graphics could "further reduce cognitive load and memory tasks" as summarized by Chang and Rice (Chang and Rice, 1991, 242). Likewise, a "visual system that models a ... library's physical layout on the screen (including visual displays of book shelves [emphasis added], floor maps, and walking paths)"(Chang and Rice, 1991:242) would offer a familiar environment that could further facilitate
identification and selection of relevant resources by users. The Science Library Catalog Project (Borgman et al., 1995) not only has implications for the design of information retrieval systems for children, but offers an excellent model for the enhancement of other hypertext retrieval systems, including the CyberStacks(sm) scheme.

Although not as sophisticated as Borgman's efforts, we believe the recent incorporation of relevant icons to identify all broad subject categories as well as subclasses within the major CyberStacks(sm) classes provides additional visual information and context that will further facilitate selection and use of relevant Internet resources.

The application of virtual reality browsers (e.g., Pesce, 1995) to the CyberStacks(sm) scheme could literally provide the added dimension that would further facilitate use. The landmark experimental project directed by Arthur W. Rhyno, Systems Librarian, University of Windsor, Canada, that has applied VRML to 'navigate information space' at the Leddy Library, represents a major advance in user-friendly interfaces to the local online library catalog (OPAC) and has significant potential application for enhancing CyberStacks(sm).

8. Conclusion

In a recent review article on the pitfalls and promises of the virtual library, (Rooks, 1993, 22-29), Rooks provides a good general critical review of some of the major ramifications which the emergence of the virtual library may have on conventional library services. In concluding her article, she offers a balanced view on the future of libraries and clearly articulates the overall role that librarians should play in their development:

"The virtual library is not the ultimate answer to everyone's information needs. It is merely another step in a dynamic and evolutionary process. The traditional print library and traditional library services will not disappear. But, as librarians, we must accept and adapt to the introduction of new techniques and systems. We must recognize the enormous potential of the virtual library, address the issues involved in its creation, and take a leadership role in integrating these new systems and services into our libraries, for our own good and for our users" (Rooks, 1993, 28).

In creating the Virtual Library, librarians must look beyond individual and local needs, and begin to give fuller consideration to the ways in which our professional skills can be applied not just to the creation and maintenance of a special collection of Internet resources, but to the larger issue of centralized, integrated and unified national collections of Internet resources built through cooperative collection development projects in a networked environment. Within systems such as CyberStacks(sm), all well-managed local efforts potentially will not only benefit a targeted audience, but a broader, world user community as well.

References

<http://www.dlib.org/dlib/december95/briefings/12cyber.html>
<http://www.public.iastate.edu/~CYBERSTACKS/OCLC.htm>
Michiaki Iwazume, Hideaki Takeda and Toyoaki Nishida
Graduate School of Information Science, Nara Institute of Science and Technology

Ontology-based Information Capturing from the Internet

Abstract: In this paper, we present a system called IICA (Intelligent Information Collector and Analyzer) which gathers, classifies, and reorganizes information from the Internet. Ontology plays an important role in IICA. It specifies the common background knowledge shared by the user and IICA, allows IICA to make inexact match between the user's request and the candidates, and assigns user-oriented categories. IICA extracts information using a state transition network grammar and concept frames. We have implemented and evaluated IICA. The results shows the feasibility and robustness of the approach.

1. Introduction

As the number and diversity of information sources on the Internet is increasing rapidly, there is an increase demand for intelligent assistants which would help people search for desired information.

A number of tools are available to help people search for information on the Internet such as WWW Worm (McBryan, 1994), WebCrawler (Pinkerton, 1994). Unfortunately, existing tools are unable to interpret the content of information resources due to the lack of knowledge. We need more intelligent systems which facilitate personal activities of producing information such as surveying, writing papers and so on.

In this paper, we present IICA which gathers, classifies, and reorganizes information from heterogeneous information resources on the Internet. Ontology plays an important role in IICA. It specifies the common background knowledge shared by the user and IICA, allows IICA to make inexact match between the user's request and the candidates, and assigns user-oriented categories. Figure 1 shows the outline of IICA.

This system has the following functions: (1) Information Gathering: IICA gathers WWW pages on the Internet in response to user's requests. IICA uses ontologies to compute the similarity between the keywords given by the user and those extracted from candidate pages. (2) Information Categorizing: IICA categorizes the gathered pages by linking them with an ontology and (3) Information Reorganizing: IICA extracts specific information from pages using heuristics based on expression patterns and phrases (See Figure 2).

We tested IICA on the WWW. The results of the experiments suggests that the ontology-based approach enables us intensive use of heterogeneous information resources on the wide-area networks such as the Internet.

In Section 2, We explain an information gathering method using ontologies and heuristics. In Section 3, we explain a new method of text categorization using ontologies. In Section 4, we describe how IICA uses heuristics based on expression patterns and phrases to extract and reorganization specific information from pages. In Section 5, we describe the evaluation of the above three methods. In Section 6, we discuss the advantages of our approach and summarize this paper.

2. Ontology-based Intelligent Information Gathering

2.1 Ontology

An Ontology is specification of conceptualization which consists of a vocabulary and a theory (Gruber, 1991). The role of ontologies on information gathering is to provide knowledge for agents to infer information which is relevant to user's requests. Ontologies are often described in frame like languages and knowledge representation languages based on first-order predicate logic such as Ontolingua (Gruber, 1992). We decided to make use of weakly structured ontologies (Nishida, 1995) which is developed from existing terminologies, thesauruses (Iwazume, 1994). Weakly structured ontologies have only one type of associative relation between terms (see Figure 3).

2.2 Information Gathering on the WWW

IICA collects WWW pages by (1) accessing HTTP or (2) searching the archive of WWW pages. In the former case, IICA gets the specified page by sending a URL address to its socket modules and accessing the specified host. The gathered page is added to the archive. All pages in the archive are managed by IICA with its file table. In the latter case, IICA searches the archives using the file table.

The algorithm is basically breadth-first searching. The difference is that IICA evaluates gathered pages and decides which anchor to access next. Figure 4 shows an example of gathering pages on the WWW using the ontology-based method.

Suppose that the user's query consists of a keyword "knowledge base"and a scope parameter 4.0. IICA generate a set of related terms to the keyword using the ontology (See the upper right-hand side in the Figure 4). The distance between each related terms and the query keyword is within 4.0. In this example, The anchor "Knowledge Engineering" is given a weight 1.0 because it contains the pattern "knowledge". For detailed technical information, see Iwazume (1995).

2.3 Heuristics

When we search for information on the WWW, we use various heuristics such as empirical knowledge and common-sense. For example, the following heuristics seems reasonable when we search information on artificial intelligence: "the WWW page of institutes, laboratory often contains information about AI."

Such kind of heuristics can make the information gathering process more effective in cooperation with ontologies. For instance, the heuristics that "if search for information on AI, go pages of laboratory" is described as follows:

"artificial intelligence" $\rightarrow$ "laboratory"

IICA gives priority over the pages which contain term "laboratory" and access them by using the heuristics.

3. Ontology-based Text Categorization

Ontology-based text categorization is the classification of documents by using ontologies as category definition.

In our approach, the process of text categorization is twofold: (1) Text categorization by
calculating similarity between a feature vector and a category vector, (2) Modifying weights between terms in an ontology by calculating similarity between category vectors (see Figure 5).

A feature vector is a vector which represents feature of a document, while a category vector is a vector which represents the characteristic of a category. The feature vector is calculated from the term frequency and the inverse document frequency. The category vector is calculated from the feature vectors of the document assigned to the category.

We use vector space model commonly used in the information retrieval studies to weight terms and calculate feature vectors (Salton and McGill, 1983). The algorithm is as follows:

Step 1. Calculate the feature vectors of the gathered pages.
Step 2. Classify the gathered pages by calculated the feature vector.
Step 3. Calculate the category vectors from the classified pages.
Step 4. Repeat step 2 and step 3 until the category vectors converge.
Step 5. Calculate distance between the categories and renew weight between terms in the ontology.

Each initial category vector is calculated from the feature vector of the pages which is assigned to the category by matching keywords.

4. Information Extracting and Reorganization

This section describes information extracting and reorganization using heuristics. We collected and analyzed the sightseeing pages in Japanese. As the result, it was found that it is possible to extract and reorganize specific information from pages using heuristics based on expression patterns and phrases.

State diagram method. It is the method to analyze and extract specific items according to a state diagram. For example, in case of extracting information about transport facilities, IICA analyzes in such sequence as,

\[
\text{bus stop(point) } \rightarrow \text{bus } \rightarrow \text{bus stop(point) } \rightarrow \text{walk } \rightarrow \ldots
\]

Rule-based method. It is the method to extract specific items according to attributes and rules defined in ontologies. This method can be widely applied to various information on the WWW.

We describe the above two methods in detail.

4.1 State Diagram Method

The process of this method consists of three steps: (1) finding description, (2) extracting names of sightseeing places, and (3) analyzing description and extracting items using a diagram.

Figure 6 shows the process of analyzing description about transport facilities. The above state diagram in the Figure 6 is used for analysis and the bottom sentence is the target description. The thick curved line shows a sequence of states in the analysis. The analysis starts at the initial state.

The pattern "駅" ('station') turns out in the description; the current state changes to the state, 地点 ('point'). Since the first segment of the description includes "駅" ('station') which indicates "point," the current state changes to the "point" and the system gets the station name "河原町駅" ('the Kawaramachi Station').

Next, the pattern "バス" ('bus') is found; the current state changes to the state "バス" ('bus'), and it gets the name of the bus company "市バス" ('the City Bus'). Then, since the
264

expression pattern " 停 " ('bus stop' )" is found in the description, the current state changes to the
state 地点 ('point'), and it gets the bus stop name " 修学院離宮道場 " ('the Shugakuin Detached
 Palace Street Bus Stop'). It repeats the same process till the analysis reaches the end of the
description.

4.2 Rule-based Method
The descriptions such as 効能は神経痛である ('It takes effect on neuralgia') and 露天風呂がある
 ('There is an open-air bath') appear frequently in the pages about hot-springs. The
expression "痛 " means a 'pain' and the expression "風呂 " means a 'bath' in Japanese. Then we
use rules based on expression patterns and phrases like the above examples.

The process of describing extraction rules is twofold.

Definition of attributes. A specific item to extract is defined as an attribute of a class in the
ontology. For instance, attributes such as name, style, ingredient, effects are defined to extract
information related with hot-spring.

The following formula is the definition of hot-spring attributes. This formula means that
温泉 ('hot-spring') has the attribute called 温泉の名前 ('name') which take one value, the
attributes such as 風呂の種類 ('style'), 滑物 ('ingredient') and 効能 ('effects') which take some
values, and it is a 訪問地 ('tourist resort').

\[
\text{(define-pclass (温泉 (has-one 温泉の名前 )
(is-a 訪問地))}
(has-some 風呂の種類 )
(has-some 滑物 )
(has-some 効能 )))
\]

Describing extraction rules based on specific expression patterns. The following formulas
are the rules to extract effects of hot-springs. The first rule means that if the expression pattern
"効能 " ('effect') or "効力 " ('effective') appears with the concept 傷病 (sickness and injury) in
the same sentence, the pattern matching the concept 傷病 indicates 効能 ('effects'). The second
rule means that the expression patterns "+症 " or "+傷 " or "+痛 " turns out in the sentence, the
pattern indicates the concept 傷病 (sickness and injury). Here, the symbol "+" holds the same
meaning regular expression. For example, the expression "+痛 " ('pain') matches "関節痛 
('arthralgia'), "腰痛 " ('lumbago') and so on.

\[
\text{(define-concept (効能 (is 傷病 with (or "効能 >" "+ょう " "効果 " "効く ")))}
\text{(define-concept (傷病 (or "+症 >" "+傷 >" "+病 >")))}
\]

5. Evaluation

5.1 Evaluation of Gathering Information
We tested an ontology-based method for information gathering tasks on the WWW. We
evaluated our system by accuracy and efficiency.

5.1.1 Test of Accuracy
In order to evaluate its accuracy, we restricted 100 pages, and chose the 5 queries related
to AI in English and the 5 queries related to sightseeing in Japanese. Then we ran IICA on the WWW in the following ways.

1. *Breadth first search.* IICA doesn't use ontologies. It traces hyperlinks on the WWW using breadth first algorithm.


We evaluated the result of the experiment according to the standard as follows.

○: The collected page is directly related to user's queries.

△: The collected page is not directly related to user's queries, but it is related to user's interests.

×: The collected page is neither directly related to the user's queries nor related to user's interests.

Table 1 and Table 2 show the results.

<table>
<thead>
<tr>
<th>Search method</th>
<th>○ (%)</th>
<th>△ (%)</th>
<th>× (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Breadth first search</td>
<td>64.6</td>
<td>07.4</td>
<td>28.0</td>
</tr>
<tr>
<td>2. Ontology</td>
<td>66.6</td>
<td>11.6</td>
<td>21.8</td>
</tr>
<tr>
<td>3. Ontology + heuristics</td>
<td>67.8</td>
<td>10.6</td>
<td>21.6</td>
</tr>
</tbody>
</table>

Table 1: Evaluation of Gathering Pages Relevant to Artificial Intelligence

<table>
<thead>
<tr>
<th>Search method</th>
<th>○ (%)</th>
<th>△ (%)</th>
<th>× (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Breadth first search</td>
<td>57.4</td>
<td>08.4</td>
<td>34.2</td>
</tr>
<tr>
<td>2. Ontology</td>
<td>59.5</td>
<td>15.6</td>
<td>24.9</td>
</tr>
<tr>
<td>3. Ontology + heuristics</td>
<td>59.5</td>
<td>15.6</td>
<td>24.9</td>
</tr>
</tbody>
</table>

Table 2: Evaluation of Gathering Pages Relevant to Sightseeing

5.1.2 *Test of Efficiency*

We tested search efficiency of our method. We restricted 500 search steps and chose the 2 queries related to AI in English. Then we ran IICA on the WWW in the above three ways.

Table 3 shows the search result to the query consists of one keyword "knowledge base".
Table 4 shows the search result to the query which consists of two keywords "semantic network" and "production system". Here, the numbers in this table indicate numbers of pages.

5.2 *Evaluation of Information Categorizing*

We made an experiment of categorizing the about 500 pages concerned with AI in English and the about 800 pages concerned with sightseeing in Japanese. In order to evaluate our method, we calculated recall and precision. The result is shown in Table 5.
<table>
<thead>
<tr>
<th>Search method</th>
<th>O</th>
<th>Δ</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Breadth first search</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2. Ontology</td>
<td>21</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>3. Ontology + heuristics</td>
<td>44</td>
<td>13</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 3: Evaluation of Efficiency of Information Gathering—1 keyword ("knowledge base")

<table>
<thead>
<tr>
<th>Search method</th>
<th>O</th>
<th>Δ</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Breadth first search</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2. Ontology</td>
<td>10</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>3. Ontology + heuristics</td>
<td>18</td>
<td>23</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 4: Evaluation of Efficiency of Information Gathering—2 keywords ("semantic network" and "production system")

### 5.5 Evaluation of Extracting Information

The evaluation of two extracting methods is done with The targets were the WWW pages about sightseeing in Japanese. We tested our state diagram method for analyzing the 100 pages which contained description about transport facilities. Table 6 shows the results of the experiment. We also tested rule-based method for extracting information from the pages concerned with hot-spring, restaurant, and temples. Table 7 shows recall and precision results.

### 6. Conclusion

In this paper, we present a system called IICA which gathers, classifies, and reorganizes information from the Internet. We have implemented and evaluated IICA. We can conclude the following advantages of our approach from the results.

- Ontology and heuristics make accuracy and efficiency better in information gathering.
- IICA can understand which information is related to user's request using ontologies.
- IICA allows the user to search and reach the misclassified items by tracing ontological relations.
- It is possible to easily extract and reorganize specific information from very large text data by using heuristics based on expression patterns and phrases.
- The ontology-based approach enables us intensive use of heterogeneous information resources on the Internet such as the WWW.
<table>
<thead>
<tr>
<th></th>
<th>AI (English)</th>
<th>Sightseeing (Japanese)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>81.9</td>
<td>79.0</td>
</tr>
<tr>
<td>Recall</td>
<td>80.5</td>
<td>70.0</td>
</tr>
</tbody>
</table>

Table 5: Evaluation of Categorization of WWW Pages

1. A rate of pages that contain descriptions accurately found
2. A rate of pages that contain descriptions accurately analyzed and extracted

Table 6: Evaluation of Extraction of Traffic Information Using a State Diagram

<table>
<thead>
<tr>
<th>Domain</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot-springs</td>
<td>82.2%</td>
<td>61.2%</td>
</tr>
<tr>
<td>temples</td>
<td>72.2%</td>
<td>73.4%</td>
</tr>
<tr>
<td>restaurants</td>
<td>85.0%</td>
<td>41.0%</td>
</tr>
<tr>
<td>Average</td>
<td>79.8%</td>
<td>58.6%</td>
</tr>
</tbody>
</table>

Table 7: Recall and Precision of Extraction of Information Using Heuristics

References
Figure 1: Outline of IICA

Figure 2: An Example of Reorganization of Hot-Spring Information on the WWW
artificial intelligence; computational linguistics; production system / knowledge base management; multiple world; knowledge network; distributed knowledge base; knowledge base editor; logical database

pattern information processing; knowledge representation; knowledge base; semantic network

Figure 3: An Example of a Weakly Structured Ontology
keyword: knowledge base
scope parameter: 4.0

WWW page

- parse
- anchor-list
  - label
  - URL address
  - weight

add to the open-list / sort

open-list

Knowledge Engineering 1.0
Artificial Intelligence 2.0
Graduate School of Information Science 4.0

collect the top item of the open-list
repeat above process

Figure 4: An Example of Information Gathering on the WWW
Figure 5: Text Categorization Using an Ontology
Figure 6: An Example of Extraction of Traffic Information Using A State Diagram
With the burgeoning of client access to, and the proliferation of resources on, the Internet, and particularly, the World Wide Web (WWW), the need to have at least minimal "indexing" of available electronic materials is becoming more apparent (Udell, 1995). Keyword searching (through Archie, Veronica, and index-and-search engines for WAIS and WWW), while flexible and amendable to the natural language queries of individuals, yields, in many cases, either very high or zero recall. Some Internet clients might argue that too much information of dubious or no relevance is retrieved, and that a kind of "packaging" or more precise "labelling" of resources would result in more relevant and precise search results. While determining a "common language" across disciplines might prove linguistically challenging, understanding the nature, degree, and extent of both language uniqueness and terminological "drift" or "interchangeability" across several select subject disciplines, would offer a valuable first step in determining the need for, and potential applications and implications of a controlled vocabulary -- a metathesaurus -- to encompass those disciplines. Moreover, determining terminological uniqueness versus overlap would also support an evaluation of the suitability and applicability of existing general-level tools, such as the Library of Congress Subject Headings or general thesauri, for providing controlled vocabulary access to Internet resources.

A review of the literature reveals that while considerable informetric research has been conducted concerning 1) term frequencies and distributions, word clustering, and the broader context of content analysis within documents, or 2) the automatic selection of an indexing vocabulary suitable for describing documents within a more or less homogenous field of study, such as medicine, nuclear physics, or genetics (see, for example, Harter, 1975; Fedorowicz, 1982), there is a paucity of research applying the same or similar informetric techniques across subject areas. The exploratory study currently in progress is intended to address that gap, and to consider, in particular, patterns of terminological uniqueness versus co-occurrence across a sampling of representative disciplines within the humanities, the social sciences and science.

Assuming that the most important, most popular, or best established topics within a discipline will be most frequently represented in the vocabulary of documents within a discipline, and using programming already developed as part of a previous, related study of the application of frequency tables (Gignac and Howarth, 1994; Howarth and Gignac, in press), the research involves downloading records from a selection of CD-ROM databases and ranking frequencies of word counts from those records. Two hundred record titles, abstracts, descriptors, and source journal titles randomly selected from articles published in each of 1978, 1990, and 1993, are being derived from CD-ROM databases for Historical Abstracts, Library and Information Science Abstracts, PsychLit, the MLA Bibliography, Sociofile, ERIC, and MEDLINE. Discounting stop words, word frequency counts for the approximately 600 records per database are being calculated and ranked using a programme previously developed, and rates of terminological uniqueness versus co-occurrence -- interdisciplinary "drift" patterns across the eight-year time span -- are being inferentially compared using PC-SAS. Preliminary findings are being assessed to determine
whether there is any theoretical base for 1) considering the feasibility of an interdisciplinary metathesaurus, per se, and/or 2) evaluating existing general-level indexing tools (subject headings lists; thesauri) as to their potential suitability as interdisciplinary metathesauri.

References
The Evolution of a Concept System: Reflections on Case Studies of Scientific Research, Italian Literature and Humanities Computing

Abstract: By using rigorous arrangement criteria it is possible to organize concept systems whose structure can point out the following evolutionary phases: the rising through the characterization of the object and its typology; the features and the capacity to organize the knowledge domain within which it develops; the vitality produced by the impulses and motives of research which tend to increase the knowledge domain; the capacity to relate with other concept systems, identifying their osmotic relationships between knowledge fields; the generation of new knowledge fields; the solution through its integration in more complex systems in which new knowledge, correctly organized, converges.

1. Introduction

To perform the research described in this paper, we applied Dahlberg’s “Systematifier” category model for the analysis and concept structuring of the following domains: Scientific Research, Italian Literature, Cereals Cultivation and Humanities Computing. The application of the model to such diverse domains might highlight—according to the subjective viewpoint of the users of these principles—interesting aspects of the domains structured: that is, aspects in part common to all the domains examined and in part specific to each of them. It has been possible to note, for example, how a concept system representing the structure of a domain evolves at a rhythm which stresses the various phases of its life and characterizes a domain physiognomy, sometimes unidentified even by scholars of the subject.

The paper sets out to illustrate these aspects by focusing on three of the domains cited. Another paper considers the problems involved in the study of Cereals to stress how, applied to a specific disciplinary sector, analysis and inductive procedure may help to improve knowledge not only of the discipline inherent to the sector itself, but also of other disciplines.

2. The Concept System

A field of knowledge is identified and defined through a material or immaterial object taken from the reality and concepts which, on the basis of their statements, refer to the object itself. The organization of the structure of this field of knowledge may be determined by analysis of the characteristics of each concept and application of an inductive procedure. Analyzing the characteristics of the object and the concepts correlated to it, it is possible to highlight, through an inductive procedure, concepts of a higher order which ultimately identify with the application of categories to the field of knowledge in question.

The Systematifier category model allows us to identify these concepts of a higher order in each knowledge field, and to note relationships between them and all other concepts. We thus generate the general structure of the knowledge field, the framework of the system in which each category encompasses the concept cluster correlated to it. By establishing hierarchical relations between the single conceptual units which combine to form a cluster, the knowledge field is
ordered in a concept system represented by relationships between the single conceptual units.

The identification and definition of a field of knowledge are made according to forms and patterns which stem from two typologies. The first is a field of knowledge marked by an autonomous, independent object of reality described by its characteristics and the conceptual units which refer to it. Examples of this typology may be found in the fields of «Scientific Research» and «Italian Literature». The second typology considers a field of knowledge which emerges from the integration of different objects, each equipped with characteristics and conceptual units of its own. The new field which is thus determined is the product of a combination of different domains of knowledge and ultimately takes on a life of its own, represented by a structure of its own. An example is «Humanities Computing» which is determined through the integration of the domains of Humanities and Computer Science.

3. Identification and Definition of a Knowledge Field

The identification of a new field of knowledge takes place through scientific research activity which helps to increase knowledge of an object of reality. The new knowledge field is defined through the object and the concept cluster correlated to it. Even if the object is by its very nature integrated within a discipline—"cereal", say, in Agricultural Science, or a literary phenomenon in Literature—new concepts, which scientific research gradually adds to the field, progressively increase its development. The field of knowledge which has achieved a certain degree of development is thus autonomous—with often unapparent methodological characteristics of its own—disorderly and shapeless. If these concepts are ordered logically following a structure model, it is possible to use them and make them usable. It is thus necessary to organize the conceptual structure by attributing the proper classification in the system to each conceptual unit which belongs to the relevant knowledge field.

Through analysis and experimentation, we have seen how Systematifier embodies the simultaneous duality of formula and guide. As a formula, it provides rules and principles to which the right values have to be attributed in order to establish the structure of the system. As a guide, it directs the user towards rational choices in his own disciplinary field, offering him every opportunity to experiment and verify the principles themselves.

The Systematifier category model goes on to delineate and characterize the new field of knowledge, evidencing the object, its action, its theory and other characteristics.

3.1 Object

This category encompasses all concepts which identify the object and its properties, that is, 'what the object is' and how it appears.

The identification of the object is easy to imagine in some cases—in the study of "cereal", for example—but demands closer attention in others—the «Research» and «Italian Literature» domains, for example.

In the «Research» domain it was necessary to distinguish the "research" concept from that of "science". "Research" taken as the action carried out for the improvement of the process of knowledge and the process of application, is often considered a synonym of "science", a process for knowledge. Defined in this way, we tend to regard "science" as any scientific activity carried out irrespective of the results it obtains, as opposed to "science" as rigorous, methodical and systematic knowledge: that is, unified, true knowledge open to critical, conceptual and experimental verification. It was thus necessary to attribute a proper role to these terms within the system, eliminating any possible ambiguities. Likewise, it was necessary to attribute a proper
role to their derivatives: science policy, national and international science, forecasting and assessment of science and technology, big science and little science, science parks, science in the market and so on (Negrini, 1992, 296).

In the case of «Italian Literature», as an organizational criterion for the concept structure it was necessary to identify the ‘literary fact’ contained and illustrated in the literary work. The literary fact is the event which occurs at a certain place at a certain time in history, the product of the creative imagination of the author expressed using the means, forms and rules he decides to adopt (Negrini and Zozi, 1995, 127). The object of the «Italian Literature» system was identified as the “literary work” since, in so far as it contains the literary fact—that is, the instrument with which the author communicates his message—it is characterized by the criteria used by the author to express himself. These may encompass, for example, “form”—dialogue, drama, melodrama, novella, poetry, epic poetry, song, novel, satire etc—or “content”—diary, biography, historical work, obituary, collection of poems etc. In defining this particular system, it was possible to note how the categories which combine to form the structure of the system are also valid for a broader system covering «Literature» in general.

The “Humanities Computing” domain is a case apart, since it consists of the integration of two separate and autonomous knowledge fields, each of which enjoys an existence and independent development of its own. Since they benefit from powerful and accurate logical and technological instruments, the humanities receive a stimulus and an impulse to verify their methods of investigation and analysis, and thus make them more consistent. Hence the decision to identify the various sectors into which the humanities break down as the object of the new field of knowledge.

These reflections were indispensable to establish the contents of the domains and the demarcation of their boundaries—a task preliminary to the study of the formal structure of a field of knowledge—and to find inside it the proper classification of the concepts which form the domain.

3.2 Action and Process

Knowledge of the action made or undergone by the object, including the ways and forms with which the action takes place, represent the content of another Systematifier category. The concepts which express this action cast light on the purposes and methodologies of knowledge in the domain considered.

In the «Research» system, the action was viewed in terms of two distinct aspects: first, at “macrolevel” to record the activities generally carried out by public and private bodies and industry; secondly, at “microlevel” to analyze single research project and to define the methods and procedures needed to achieve innovative results at the theoretical, applied, experimental and productive levels.

For «Italian Literature», Literary Criticism was considered as the action performed on the literary work. It sets out in fact to study the work to investigate the literary fact and, from it, understand the creative image of the author. Criticism’s interpretative analysis of the literary fact is a process that cannot be arrested, reflecting the critic’s spirit and soul, but also the historical-cultural moment in which it manifests itself.

In «Humanities Computing» the action is considered as a method of scientific investigation of the object, its role being that of a “logical engine” capable of producing a consistent formalization of the research procedure and activity.
3.3 Theory

A fundamental category is that of theory about the object. By describing the object in question, the aim is to obtain a logically consistent formulation of the general principles of the knowledge field. Basically, theory consists of identifying and defining the principles which regulate and determine the existence of a knowledge field. It is possible to observe that the correct formulation of the theory of a domain emerges only at an advanced state of development of the knowledge field, and that, besides, theory also establishes the terminology to describe the objects or actions surveyed.

3.4 Other Characteristics

A criterion for identifying other characteristics which may help to structure the knowledge field may be provided by the result or product of the action performed or undergone by the object.

In the case of the «Research» system, in the context of the “microlevel” of the single research project, the result is presented in the drafting of a final report, a scientific paper describing the progress achieved and the procedures adopted to attain the set objectives. The measure of the scientific value of the product obtained is directly dependent on the product’s intrinsic capacity to furnish new knowledge. At the applied level, the product of research consists of an invention which, by virtue of its creation of new knowledge or utilization of phenomena already known, is susceptible to immediate practical employment in a different operating process. This is the process of innovation, a crucial factor in economic progress. The importance of an invention is thus proportional to the potential applied utility of the results achieved.

In the «Italian Literature» system, the result of the critical process consists of the conceptual units identified in the literary work, which we refer to as “elements” (esoteric, aesthetic, formal, literary, psychological etc). These concepts prove particularly significant in so far as they introduce the “added value” of the work as furnished by literary criticism—valuable new information for the researcher.

Another criterion which has proved common to these systems is the role of the physical person responsible for the creative process: a “scientist” or “engineer” for the «Research» system, an author, mediator (critic, actor, interpreter), reader and public for the «Italian Literature» system.

In the case of «Humanities Computing», it was possible to note that the result of the relationship established between the two fields of knowledge consists of an original research methodology and of the production of new tools for research which, in turn, are capable of determining new knowledge and new products. Hence the newly identified domain ultimately activates a process with autonomous fallouts, while simultaneously contributing also to the two original domains (Fig. 3). The tools consist principally of the application-oriented software developed: for example, the algorithmic formalization of rules for the analysis of the content and style of a text, for automatic translation or for the automatic composition of texts and music, as well as for the speech and music synthesis. As the product, we identified: data bases and expert systems in the various sectors of the humanities, landscape surveying and analysis systems (GIS), systems of computer-aided translation and automatic synthesis of sound and speech and, finally, systems for the automatic generation of language.

4. Dynamism of a Conceptual System

The application of Systematifier to define a concept system helps us to perceive in it a flow of knowledge which characterizes the system’s natural “motion”. It is produced by different impulses and aimed at improving new cognitive processes that are schematically representable.
The «Research» system is characterized by two different motions which originate in other sciences: in the *Science of Science*, which poses the «Research» system theoretical queries and produces culture; and in *Social and technological science*, which poses the «Research» system economic, technological, social and environmental queries. The results acquired in the «Research» system fall out over these sciences. The resulting flow of knowledge integrates the respective sciences which thus become a driving force for new cognitive processes (Fig. 1).

In the case of the «Italian Literature» system, criticism's action of investigation of the work which contains the literary fact highlights a centripetal action of interpretation of the literary fact. Criticism gives rise to different aspects of the literary fact which become the object of new studies and new critical surveys, thus producing the system's continuous motion of (Fig. 2a).

Analysis of an experimental sector such as Cereals Cultivation has likewise revealed a constant action in pursuit of new methods to improve products, and in analyzing their results for agriculture, food and industry. Yet this new knowledge is the point of departure for new studies and triggers a continuous process, which can be represented as a circular motion wherever the point of arrival is higher than the point of departure, and so on (Fig. 2b; process 1, 2 ... n).

In the «Humanities Computing» system, the novelties are the formalization of the research method and the potential of the multimedia approach. The humanities have to re-elaborate their study method constantly and, by revisiting its logical and philosophical roots, computer science reprocesses the theory of formal languages to allow constant investigation of the interaction between man and machine (Fig. 3).

5. The Relationship with Other Conceptual Systems

Not all the concepts that belong to a field of knowledge are of an endogenous nature. Many of them may be imported from other fields of learning with which they enter into would have been written differently. They are no less important for the researcher since osmosis. It was possible to note how concepts borrowed from other disciplines may become part of a system and, once inside it, acquire special meaning, hence a place of their own in the structure.

In the «Research» system, for example, criteria of evaluation and appraisal of scientific and technological research refer to statistical, economic, technological and bibliometric knowledge. Albeit acquired, concepts such as *Technology balance of payments (BTP)*, *Patents*, *BTP Indicators*, *Measures of research quality*, *Innovation statistics*, *Process Innovation*, *Knowhow* etc are to be regarded as new to the «Research» system in so far as they represent methodologies for evaluating research. They need to find their proper place in the system according to their own
special conceptual characteristics. In such cases a sort of "transdisciplinary" system emerges consisting chiefly of applications of the methodologies of one field of knowledge to others.

It may happen that concepts acquired from other disciplines maintain their original meaning and it is thus necessary to allocate them in such a way that reflects the correlations which they had in the ambit of their original discipline. The Systematifier model allows us to evidence such knowledge in a special category—exogenous flows vital for the concept system. An example of this second case is knowledge, which introduces to the new system phenomena or problems studied by the original system.

For «Research», this may consist of knowledge of a philosophical nature taken from the «Science of Science» domain (for instance, concepts such as logic, scientific method, theory, hypothesis etc) or of economic, social, health, technology or environment policy transferred from the respective systems.

The phenomenon is especially marked in «Literature» since not strictly literary terms serve to express the literary fact: concepts to do with Philosophy such as knowledge, aesthetics, idealism, modernism, nihilism, truth, nothingness, to do with Religion such as Christianity, cross, devil, hell, Franciscanism, Jesus Christ, to do with Psychology such as eroticism, imagination, madness, genius, melancholy, reason, sentiment, passion, dream, love, to do with Politics such as anarchy, antifascism, democracy, peace, war and so on for many other disciplines. These units are instrumental for the writer since they are essential for description of the event of the literary fact: that is, setting, character, mood, historical moment or anything else that may be molded by the author’s creative imagination. Without these elements the work would never have been written at all, or they are an essential element of the literary fact, fundamental for penetrating into the world of its author and understanding his message.

In the case of «Humanities Computing», the most important contributions come from Theoretical Computing, Cybernetics and Artificial Intelligence, but also from removed disciplines such as Statistics—although in this case the relationship would appear to be more strictly instrumental and application-oriented.

Fig. 2: Knowledge flow a) in the «Italian Literature» system and b) in an experimental sector system
6. The Proliferation of a Concept System

We have seen how an object of reality, even if it is traditionally part of one discipline, may through a set of items of knowledge referring to it acquire a connotation of its own and identify a particular domain. The same phenomenon may occur for a concept inside the structure of a concept system. The analysis and research carried out on its nature and evolutionary process make it a new object of knowledge and investigation which introduces a cluster of new concepts to the system, either acquired from other domains or totally new. Correlated to a new concept-object, this cluster of concepts invariably tends to expand and break away from the original system to achieve a structural autonomy of its own. This may happen, for example, with the study of a literary movement (Romanticism, Futurism etc) framed in the historical, artistic, intellectual and philosophical movement of a country or a culture and comparable with similar movements in other countries or cultures. It happens with studies on innovation and the characteristics and technological, social and occupational impact thereof.

The same trend may be noted in a more accentuated form and with a much more consolidated statute in Computational Linguistics, the crossroads where the original competencies of the two constituent domains of «Humanities Computing» meet. Similar phenomena may also be observed in sectors such as Teaching and Archaeology, which are most directly prepared to absorb the potential of multimedia and already show the first signs of original evolution.

Like embryonic cells, these clusters are elements of a new life with properties that force them to break away from the original system. They are flows of knowledge which nourish new domains, generating through their process and methodology the perpetual evolution of science.

7. Progressive Integration of a Concept System

Our knowledge is made up of disorderly sets of correlated, integrated and overlaid concepts. Systematized knowledge fields are limited parts of knowledge which have been analyzed and
organized, but which, in isolation, cannot represent all our knowledge. If these systems are founded on the same structural principles, then comparison between two of them might generate the presence of the same concept in both systems. This conceptual unit is common to the two structures and allows us to create a relationship between the two systems and study their rational combination and integration. "Concepts-objects" may appear that are hierarchically dependent (Science/Research: Literature/Italian Literature), or a concept may be the object of one system and present with a different allocation in a second system. Analysis of the relationships between these structures, and of the role of each concept within the categories, allows us to combine ordered clusters of concepts from the two systems and to integrate them rationally (Negrini, 1995). The crumbling of the original structures generates a system with a more articulated, complex form—a broader organized part of our knowledge. Without the analysis and ordering of original systems, "elements" of a universal system of knowledge, this broader "part" of the whole knowledge learning would not be achieved.

8. Conclusions

The application of Systematifier to identify and define knowledge fields makes it possible to understand particular aspects thereof and to note (depending the personal point of view of the user of the model) concept systems which evolve and change according to the evolution of knowledge processes.

The model represents constant criteria, valid for each domain, and allows us to create a stable structure in which each concept is allocated according to its characteristics. A distinctive feature of the model is that the progressive increase in the number of concepts does not modify the general structure but confirms basic principles, since it is easier to identify the role of each concept in the system, even if the allocation of the concept modifies the original hierarchical order inside the category. One fundamental result of the application of Systematifier needs to be stressed. The paradigmatic model allows us to record in a logical form complex concepts which, by virtue of the structure of the system may be "disassembled" and "reconstructed": for example, type of literary work, content, work criticism, product of criticism. The indexer can use this structure to analyze a subject and, by disassembling it, to identify in the categories concepts useful for the formulation of the object. The seeker can use the structure to find the single elements of a subject, with the aid of the defined categories. By these structure, he will be able to identify the concepts regarding those categories and to build up complex concepts.

Acknowledgments

The graphs were processed by Angelo Sensini of ISRDS-CNR.

References


mentation (K. G. Saur Verlag).

---


---


---


---


---


---


---


---


---


---

An Inductive Approach Towards the Integration of General Information Systems for Agriculture: The Case of CERETHES, with Particular Examples

Abstract: The present work tries to demonstrate, by concrete examples, that an inductive approach (through an object) instead of a deductive one (through a discipline), allows a field of knowledge to be organized, and creates a structure strong enough to systematize, rationally, all the concepts which belong to that field, using Dahlberg's category model, called "Systematifier". "Cereal" in this case is our "object", and its related thesaurus, which we called CERETHES, is described with some examples through the comparison with the most important general agricultural thesauri, as CAB-thesaurus and AGROVOC. CERETHES, as a result of its specificity and wideness, could be a tool of integration of agricultural sciences, nutrition, food industry and in other industrial sectors.

1. Introduction

A knowledge field consists of a cluster of knowledge units (concepts), which represents an "object" of reality with its predicates or related concepts. There are different dimensions of knowledge fields as well as different stages of development.

The research action, constantly and methodically carried out in a specific field referring to a given object, generates the contribution of new knowledge. It therefore happens that, if such a field, or part of it, is traditionally included in a discipline, it is very difficult that the new related knowledge becomes part of that discipline. The organization of new knowledge is therefore limited by the primitive structure of the discipline, in which new concepts do not find their correct allocation.

For the above mentioned reasons, it follows that acquired knowledge, without systematization, cannot be communicated properly, which means that those who have to use it and also whoever is still working on it, will lose. Such limits are present in every system, structured by disciplines.

The present work tries to demonstrate, by concrete examples, that an inductive approach (through an object) instead of a deductive one (through a discipline), allows a field of knowledge to be organized, and creates a structure strong enough to systematize, rationally, all the concepts which belong to that field.

A structure based on the "object" allows already acquired and new knowledge to be ordered; furthermore, it allows its future development to be predicted. Dahlberg's category model, called "Systematifier", gives safety and solidity to that structure. "Cereal" is the object of our work.

Research, which represents the first application of the category model on an experimental field, has been carried out by the Italian Ministry of Agricultural Resources, with the assistance of the Italian National Research Council (Institute for Studies on Research and Scientific Documentation) and the Experimental Institute on Cereals Cultivation.

2. Methodology

Thesauri are useful tools for managing indexing and information retrieval in large information systems, on a world-wide scale. The most important bibliographic agricultural data bases (CABI and AGRIS by FAO) were structured starting from the construction of two general thesauri, which they called CAB-thesaurus in the first case, and AGROVOC in the second.

The criteria used to select the key-words in the two above-mentioned thesauri were their frequency and usage in the world agricultural literature. In such a way, two general thesauri, built up by a statistical method, are now available.

The two information systems previously mentioned (together with their connected thesauri) are very efficient in the broad field of agriculture, but they cannot satisfy the need of specialized information, required by a researcher who works in a specific field and has adopted a specific terminology. The thesaurus (still being compiled), which we called CERETHES, has been created by adopting Systematifier, resulting in a good tool for knowledge organization and information retrieval in the cereal domain, looking at the specific needs of the researchers.

Two other characterizing elements of CERETHES are the presence of the definitions (attributed to each concept), and the image (attributed to the most significant one).

The principles contained in the Systematifier allows systematic order in a field of knowledge, and in particular, to those new arguments which spring from research, and progressively grow until they acquire a scientific connotation. Such principles involve researchers, who participate directly in the process of knowledge evolution.

A conceptual system is represented by the "structural relationship" between the single elements which are its components. To organize a conceptual system we need a structure.

The Systematifier is composed of nine categories. Each category, applied to a particular field, allows it to create the specific "facet", which is the fundamental element required in building clusters of concepts. The nine categories are divided into three groups, each of which contains three categories:

- the first group represents the core categories, i.e., fundamental principles, object and its definitions, activities and processes related to the object
- the second group identifies particular aspects regarding the object and its processes
- the third group includes the relationships between different conceptual systems

3. General Structure: the Facets of CERETHES

The focal point of research is "cereal". This is a living organism, with a specific life cycle, where (before its birth and also after its death) human beings have intervened to select the best varieties to be cultivated, to prepare the proper soil, to take care of its life, to modify its reactions to environmental conditions, to preserve it against diseases, to harvest and process it, and finally to store and, generally speaking, eat its products. The analyzed concepts are principally related to the cultivation of cereals, but during the definition of facets, the nutritional and industrial links, together with marketing of products, have also been considered.

To create facets we analyzed of the concepts extracted from the specialized experimental literature, through indexing.

Using Systematifier, the following nine facets were created:

3.1 First Group: Key Facets

Facet 1. Theoretical aspects of the object. Here are grouped concepts concerning the "scientific-philosophical aspects" related to the object—in other words the concepts that identify
the nature and essence of the object. This facet includes concepts regarding questions like: What is a cereal? Why do we cultivate them? Why are they so important to us? What implications do they have in relation to the development of human civilization?

Facet 2. Object and its morphology. In this facet we examine the "image" of our object, its typologies and its constituent parts. Concepts included are the phenotypical and botanical aspects of cereals, plant anatomy and structures. Examples are: wheat, rice, maize, spike, stalk, root, bract, anther, etc.

Facet 3. Life of the cereal. This important facet reflects the crucial fact that our object is a living organism. It was very difficult to group all the concepts (in a very broad meaning!) related to a living organism, and imagine its development, i.e., until its death. This facet contains all concepts related to biology in a broader meaning, including also diseases, which are an attack on the plant's life. Biological cycle of cereals (from birth to death) is described, including chemical compounds. Examples are: tillering, setting, starch, protein content, wheat leaf rust, oidium resistance, etc.

3.2 Second Group: Facets Identifying Particular Aspects

Here are grouped concepts which represent some particular activity or sector of study, partially connected with agricultural sciences, and disciplines completely different, such as industrial techniques of production, market economy, dietetic problems, marketing, etc. Most of the concepts derived from the indexing of specialized literature regards principally the cultivation and breeding of cereals. Later on, we will explore other related sectors, so as to acquire new concepts and gain more knowledge.

Facet 4. Production activities. Modern cultivation of cereals is a highly-skilled activity, which cannot be superficially carried out. All farm inputs are described, relative to cultivation, production, machinery and human activity. Concepts related to land suitability and exploitation are also mentioned and defined. Examples are: ploughing, dressing, disease control, irrigation, combine harvesting, sod seeding, etc.

Facet 5. Biotechnology. Genetics and biotechnology play a fundamental role in the life of (not only!) cereals. Cereal cultivation is related to all innovative theories and methods of plant breeding and plant biotechnology. Experimental methods, laboratory techniques, new varieties and genetic topics are also described and defined. The particular aspect brought out by this facet is Human Intervention (the artificial aspect in a biological context) in the life of cereal plants in order to obtain direct and indirect advantages. Examples are: genomic clone, type 1 callus, alternativity, somaclonal variation, genome mapping, transposon, etc.

Facet 6. Processing. There are many products derived from cereals which affect our lives. Industries process and trade in large quantities of cereal by-products all over the world. In this facet, concepts related to techniques and derived commodities, including laboratory analysis of food quality, storage, handling, scientific instruments, food products, drinks, and beauty treatment products (creams or shampoo) which use cereals as raw material, are described and defined. Examples are: bread-making aptitude, pasta, biscuits, parboiling, middling, beer, wheat protein shampoo, etc.

3.3 Third Group: Facets Identifying Relationships Between Systems

Facet 7. Socio-cultural and economic aspects. Cereals have been associated with civilization from our very beginnings; Thus their cultivation is linked to the development of industry, which is the basis of our western way of life. Here are grouped terms related to human
prehistory, agricultural development, urbanization, civilization, national cultures, diets and food habits; also rural sociology of the cultivation areas of cereals, market economy, problems of economic policy related to agricultural production and farms are included. Examples are: food habit, civilization, inflation, industrial production, land suitability, financing, etc.

Facet 8. Influences on the other sectors. This facet is related to the application of methodologies and techniques of the cereal domain to other domains. Only when these methodologies have been consolidated, will these concepts be diffused in other agricultural sectors, or other disciplines.

Facet 9. Dissemination of knowledge. This last facet is related to the diffusion of knowledge, including organization, teaching and information. Concepts and methods of extension activities and selective dissemination of information, information systems for agriculture, data bases and software involved with new technologies, international research organizations, are here grouped and defined. Examples are: AGRIS, CABI, CRIS, information systems, extension services, cymmit, irri, hypertext, etc.

The foregoing facets represent the fundamental structure of the conceptual system. To each facet, concepts referring to them have been attributed; thus we have created clusters of concepts in each single facet. Only subsequently can we establish hierarchic relations between single conceptual units belonging to each facet.

Concept attribution to a specific facet was made through a rational analysis of the characteristics of a concept in relation to the criterion reflected by the facet. It therefore happened that, especially in the first, uncertain phase of concept attribution, some concepts did not fully respond to the criteria required by the facets. Afterwards, with the presence of a growing number of concepts, which emphasized the content of the facets, we were able to modify the incorrect attributions.

Data is treated by CLASTHES, a software developed by the Italian National Research Council, able to manage an icon facet thesaurus, which allows the image of the object, if possible, to be shown. The software can manage not only hierarchies of terms, but also hierarchies of categories, and establish relationships between category (facet) and concepts.

4. Hierarchalization

In each cluster, we formulated the hierarchalization of the concepts. Due to the fact that the terms already belonged to a single cluster and that they had a logical link between them, the procedure of hierarchalization was relatively easy. Some facets contain more terms than others, but it depends on the kind of growth, as we said before, of some sectors rather than others in this phase of our work.

By formulating hierarchies of some concepts, we have created a significant structure, which is useful, we believe, not only for the seekers but for students and scholars too.

Some examples of concept structure are as follows:
The descriptor Spike demonstrates one of the most important goals of our work: to give researchers (but in this case also students or botanists) a good representation of a concept. All anatomical parts of spike are mentioned; in the world of cereals the concept of spike, as above described, is strictly related to the ripe plant of wheat, barley or rye; also the other five related terms are logically connected with the concept or description of spike. Through the adopted methodology (mainly by the presence of the facet "object and its morphology" previously defined) we have a good representation of important concepts. Another example is given:

SPIKE CHARACTER
NT1 beard presence
NT1 mucro length
NT1 spike color
NT1 spike compactness
NT1 spike fertility
NT1 spike length
NT1 spike shape
RT spike
RT varietal sheet
5. Comparison with Other Thesauri

An on-line research carried out a few months before the beginning of our work on CERETHES, showed us that several agricultural thesauri do exist, more or less specialized. All of them were built up through the traditional classification by disciplines system, and not one is devoted to the "world" of cereals. This was encouraging, and was also an occasion subsequently to compare our work with that of others.

In the agricultural sector there are several databases and thesauri, available all over the world, compiled by different methodologies. The two most important are CABI, with the connected thesaurus (monolingual with 60,000 terms) and AGRIS, with the connected thesaurus AGROVOC (implemented by FAO, multilingual with 22,000 terms).

The descriptor "spike" has the following structure in these two thesauri:

<table>
<thead>
<tr>
<th>CAB-thesaurus</th>
<th>AGROVOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPIKES</td>
<td>SPIKES</td>
</tr>
<tr>
<td>UF ears (plant)</td>
<td>BT1 inflorescence</td>
</tr>
<tr>
<td>BT1 inflorescences</td>
<td>BT2 plant reproductive organs</td>
</tr>
<tr>
<td>BT2 plant</td>
<td>BT3 plant anatomy</td>
</tr>
<tr>
<td>RT spikelets</td>
<td></td>
</tr>
</tbody>
</table>

These two structures show only "hynonyms" and, in particular, we stress in the AGROVOC structure, the presence of the concept "spikelet" as a related term. The definition we have found for spikelet is: "Part of the spike, constituted by a rachilla, on which are placed, in opposite positions, the bracts". On the basis of such a definition, we have inserted in CERETHES the concept "spikelet", as a narrower term of "spike", because "spikelet" is a constituent part of the spike. Similarly absent in CAB-thesaurus and AGROVOC are the spike characteristics, which are important in describing each variety.

Another example is represented by the concept "breeding", which in CERETHES has the following description applied to cereals:

<table>
<thead>
<tr>
<th>CEREAL BREEDING</th>
</tr>
</thead>
<tbody>
<tr>
<td>BT1 breeding</td>
</tr>
<tr>
<td>BT2 genetics</td>
</tr>
<tr>
<td>NT1 breeding methods</td>
</tr>
<tr>
<td>NT2 backcross method</td>
</tr>
<tr>
<td>NT2 bulk method</td>
</tr>
<tr>
<td>NT2 ear-to-row selection</td>
</tr>
<tr>
<td>NT2 haploid method</td>
</tr>
<tr>
<td>NT2 mass selection</td>
</tr>
<tr>
<td>NT2 mutation breeding</td>
</tr>
<tr>
<td>NT2 pedigree method</td>
</tr>
<tr>
<td>NT2 pure line selection</td>
</tr>
<tr>
<td>NT2 reciprocal recurrent selection</td>
</tr>
<tr>
<td>NT2 recurrent selection</td>
</tr>
<tr>
<td>NT2 recurrent selection for general combining ability</td>
</tr>
<tr>
<td>NT2 recurrent selection for specific combining ability</td>
</tr>
<tr>
<td>RT breeding design</td>
</tr>
<tr>
<td>RT breeding programmes</td>
</tr>
<tr>
<td>RT cereal</td>
</tr>
</tbody>
</table>
In CAB-thesaurus and AGROVOC the concept "breeding" is related in general to the plants as follows:

<table>
<thead>
<tr>
<th>CAB-thesaurus</th>
<th>AGROVOC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PLANT BREEDING</strong></td>
<td><strong>PLANT BREEDING</strong></td>
</tr>
<tr>
<td>UF breeding, plant</td>
<td>RT breeders rights</td>
</tr>
<tr>
<td>BT1 breeding</td>
<td>RT genetic inheritance</td>
</tr>
<tr>
<td>NT1 tree breeding</td>
<td>RT haplomethods</td>
</tr>
<tr>
<td>RT breeders rights</td>
<td>RT introduced varieties</td>
</tr>
<tr>
<td>RT gene banks</td>
<td>RT land varieties</td>
</tr>
<tr>
<td>RT germ plasm releases</td>
<td>RT plant biotechnology</td>
</tr>
<tr>
<td>RT inheritance</td>
<td>RT plant introduction</td>
</tr>
<tr>
<td>RT plant breeding methods</td>
<td></td>
</tr>
<tr>
<td>RT substitution lines</td>
<td></td>
</tr>
<tr>
<td>RT translocation lines</td>
<td></td>
</tr>
<tr>
<td>RT varieties</td>
<td></td>
</tr>
<tr>
<td>RT variety trials</td>
<td></td>
</tr>
</tbody>
</table>

In this case, AGROVOC has no narrower terms, CAB-thesaurus has only one narrower term (tree breeding), CERETHES, as shown before, and necessarily from our point of view, takes care of the concept "cereal breeding", describing its most important methods in narrower terms.

The final example concerning structures refers to the description of the concept "bread making", first in CERETHES, then in the other two thesauri:

<table>
<thead>
<tr>
<th>CERETHES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BREAD MAKING</strong></td>
</tr>
<tr>
<td>BT1 food processing</td>
</tr>
<tr>
<td>NT1 micro-bread making</td>
</tr>
<tr>
<td>RT bakers' yeast</td>
</tr>
<tr>
<td>RT bakery</td>
</tr>
<tr>
<td>RT baking</td>
</tr>
<tr>
<td>RT baking improver</td>
</tr>
<tr>
<td>RT bread</td>
</tr>
<tr>
<td>RT bread making aptitude</td>
</tr>
<tr>
<td>RT bread making quality</td>
</tr>
<tr>
<td>RT bread making trial</td>
</tr>
<tr>
<td>RT dough mixing stability</td>
</tr>
<tr>
<td>RT kneading</td>
</tr>
<tr>
<td>RT soft wheat flour</td>
</tr>
</tbody>
</table>

As can be seen, CERETHES has one narrower term, "micro-bread making", which refers to laboratory tests to define correctly the bread making aptitude of a flour, with related terms to describe other important concepts which have to be considered in this context.
The above examples are to illustrate how new knowledge is incorporated, and can be used in disciplines such as agricultural science, thanks to specialized "researchers oriented" thesauri.

The goal of our study is not only to provide a useful instrument for experts, but also to contribute to the various disciplines related to the whole context of cereals.

In our opinion, an in-depth integrative study of the structures of several thesauri greatly helps the development and classification of new knowledge.

Indeed, CABI, FAO, the National Agricultural Library, with the Department of Agriculture of the United States, together with other important world documentation centers, are presently working on a Unified Agricultural Thesaurus (UAT) which, at international level, is to harmonize agricultural terminology on the basis of experience with large information systems.

Integration projects of this kind would incorporate those methodologies and experiences about cereals already mentioned, which can then be applied to all sectors of agriculture: viticulture, olive culture and so on.

6. Conclusions

Our work is far from finished. Experience so far, though confined to the sector of cereals, has allowed studies of this "object" to extend and influence other sectors and disciplines, i.e., commercial production of pasta and bakery products: thus the study of our "object" (cereal) has demonstrated important lateral applications.

Using a categorical model, we could organize the structure of the system, building the general scheme in which concepts related to biological phenomena and artificial processes of cereals, their results, food and dietetic knowledge, etc., will be able to find a logical classification.

More analysis and research will be necessary, needing specialist assistance, which will allow us to gain more knowledge, and to increment and organize the clusters of concepts related to the defined facets.

We are also working on the classification of a sector which, as a result of its structure, will be able to integrate ordered knowledge in different disciplinary fields, further on in agriculture's future.

Notes

1. Definition has been taken from: Universal Italian Lexicon of language, arts, sciences and technique—Institute of Italian Enciclopedia, founded by G. Treccani—Rome, 1981.

References


Abstract: In response to a cultural imperative evoked by the AIDS pandemic, the authors have developed a controlled vocabulary specific to HIV/AIDS—encompassing the biomedical, legal, psychosocial, educational, historical, and literary domains—and accessible to communities heavily impacted by the epidemic. The work reflects both knowledge organization systems and environmental perspectives evidenced by the disease. It is intended to serve as a record of the lexicon—past, present, and future—and as a tool to organize the body of knowledge associated with the pandemic.

1. Introduction
Circumscribed by the same complexities as the malady itself, the body of knowledge that encompasses HIV/AIDS is diseased. Indeed, the vocabulary specific to the AIDS arena is riddled with signs and symptoms unique to this modern plague. The nature of the information devoted to the disease prohibits the use of existing organizational tools to describe effectively the multi-faceted concepts central to the epidemic. The work discussed in this paper represents an example of what has been called a defining role of information science: the organization of knowledge, or "knowledge work" (Florance and Matheson, 1993).

2. Background
The body of knowledge associated with HIV/AIDS is arguably more complex than that of any other disease. Moreover, the epidemic has altered the model of information production and consumption. In addition, HIV/AIDS has created its own vernacular that is representative of the diverse population of information consumers. Further compounding the intricate milieu, the body of information surrounding this malady continues to grow at an epidemic rate.

All information associated with biomedicine is complex; however, HIV infection is not only an extremely complicated disease process, but it also transcends the boundaries of biomedicine. In fact, HIV/AIDS transcends the boundaries of life and death themselves (Huber, 1993a). Included in the concerns associated with the disease are political, social, economic, legal, philosophical, psychological, and religious ramifications that go far beyond those encountered in other ailments. Consequently, the organizational schema must be broad in coverage yet specific in terminology so that the multidisciplinary, interdisciplinary, and cross-disciplinary nature of the epidemic is reflected.

In response to a perceived lack of support from the scientific community, the HIV/AIDS arena has fostered a non-traditional communications model where information is produced and consumed at the individual, local, national, and international level. Typically, information...
has been produced by the scientific community; forwarded to peer reviewed, clinical publication sources; and applied to greater or lesser degrees across the spectrum of care. The epidemic has changed this traditional model in that it is no longer the sole responsibility of researchers to produce information and of clinicians and administrators to apply that information (Michal-Johnson and Bowen, 1989). Rather, traditional consumers (i.e., patients) are often active producers of information within the AIDS arena, while traditional producers of information (i.e., researchers and clinicians) are often consumers of the information found in non-traditional sources such as underground newsletters and electronic bulletin boards. Yet, the traditional roles continue to exist as well. This paradigmatic shift confounds efforts to collect, organize, access, manage, and disseminate information concerning the illness and all its constituent elements.

Owing to the diverse—and unique—population of HIV/AIDS information producers and consumers, vernacular specific to this arena has arisen (Huber, 1993b). This vocabulary consists not only of scientific and biomedical terminology (Bierbaum, Brooks, and Brooks, 1992) but also includes language representative of the disciplines of education, anthropology, sociology, psychology, theology, and philosophy. Further differentiating the vernacular, terms common among communities affected by the disease are included. Partly because of the breadth of subject matters involved in the epidemic, yet the specificity required of the language used to describe it, traditional resources do not presently reflect the diversity of the vernacular.

Parallel to the number of documented cases of AIDS worldwide, the literature continues to grow at an epidemic rate. Several bibliometric studies have documented the rate of growth over time and compared it to the growth in numbers of cases (Forney, 1990; Gillaspy, 1995; Pratt, 1992; Self, Filardo, and Lancaster, 1989; Sengupta and Kumari, 1991). An organizational scheme useful for describing this literature and making it accessible to the producers and consumers of information across all boundaries is a helpful management tool when confronted with such a rapidly growing literature.

3. Statement of Problem

HIV disease and its end stage, AIDS, encompass a vast array of issues and concepts, crossing multiple disciplines. Furthermore, the language of HIV/AIDS has been produced by consumers of that body of knowledge. Existing vocabularies, while adequate for the areas they serve, do not reflect the full spectrum of the discourse engendered by the pandemic. Some examples serve to illustrate the variety of concepts specific to HIV/AIDS that are not explicitly represented in currently available lexicons. Two of these resources are general ones, while two are specific to some aspect of AIDS.

Library of Congress Subject Headings (LCSH) provides access to a variety of subjects. However, LCSH is not intended to describe biomedical concepts specific to HIV disease. For example, otorhinolaryngologic disorders—including otitis media, mastoiditis, rhinitis, and sinusitis—are a growing medical complication and serious concern among HIV-infected individuals; many publications assist physicians with managing these inflammations. Describing these works in such specific terms is not possible using LCSH.

While the National Library of Medicine’s Medical Subject Headings (MeSH) does allow for this degree of detail with regard to biomedical concepts, it does not provide the level of specificity necessary to describe various psychosocial and service-related aspects of the disease and its management. Services to HIV+ individuals and those living with AIDS constitute a significant component of the discourse associated with the epidemic. For example, buyers’ clubs, a concept not represented in MeSH, facilitate access to medications for the HIV-
infected. (Thalidomide—once prescribed as a sedative for pregnant women but found to produce deleterious side effects among their offspring—is currently a valid, approved treatment option for HIV-associated wasting syndrome, yet was until recently available only through buyers’ clubs.) As services to HIV-infected persons proliferate, this portion of the discourse continues to expand.

Even though two existing lexicons are HIV/AIDS-specific, they do not represent the entire breadth and depth of the discourse. The strength of the AIDS Information Network’s (AIN) controlled vocabulary lies in its focus on HIV/AIDS, but the bulk of the terminology concerns aspects of the epidemic void of biomedical concepts. Likewise, the National AIDS Clearinghouse’s Educational Materials Database Thesaurus provides access to only one domain of the body of knowledge.

Whether general or specific, all existing controlled vocabularies lack terminology representative of the entire spectrum of the global HIV epidemic. For example, one concept central to present understanding of HIV management is category of disease progression. Currently, patients who do not progress through the disease continuum (HIV infection, HIV+ and asymptomatic, HIV+ and symptomatic, AIDS) as it is currently understood today are classified as long term survivors, non-progressors, rapid progressors, or cleared. These specific terms, ubiquitous in the epidemic, are not found in other vocabularies. Yet another concept that transcends all boundaries of the epidemic is that of disclosure within the context of the disease: HIV infection itself, mode of transmission, sexuality, mortality. HIV/AIDS and HIV/AIDS-Related Terminology: A Means of Organizing the Body of Knowledge represents an effort to bridge the many disciplines, facets, and concepts included in the discourse specific to the pandemic.

4. Literature Review

Scientific communication has long interested information scientists and bibliometricians. Through the twentieth century, information science has matured as a discipline concurrent with an explosion in scientific knowledge and the vocabulary used to describe and document it. As the body of information has expanded, once-specific disciplines (e.g., biology and chemistry, astronomy and physics) have overlapped or even merged, creating entirely separate fields of study (i.e., biochemistry, astrophysics). Specialization has in some cases actually contributed to the dissolution of borders between disciplines (e.g., medical anthropology). As science informs increasing portions of human existence, disciplines from the social sciences and humanities have overlapped with scientific fields (e.g., philosophy of science). Griffith (1990) goes so far as to describe the aim of scientific study to be deciphering, comprehending, and interpreting human behavior.

At times, various aspects of a scientific problem are being dealt with in multiple disciplines. Theoreticians reflecting on the social study of science note that “building a commonly recognized body of knowledge is seen as the ultimate goal of scientific research” (Pierce, 1990, 52). McInnis (1995) asserts that as knowledge is organized, principles and theories—concepts—emerge. Increasingly, the bodies of knowledge, and therefore the concepts central to them, emerge from interdisciplinary foundations. As insights into the problem are gleaned from a multidisciplinary perspective, researchers in the information professions have the responsibility to synthesize the findings and package them for dissemination. Indeed, such packaging is essential for the information to be made available to a broader audience, so that work on the scientific problem may be further integrated and refined (Pierce, 1990).

This integration of disciplines and approaches was examined hypothetically by Karl
Popper. Popper theorized the existence of three worlds: World 1, composed of material, physical objects; World 2, composed of subjective, individualistic ideas; and World 3, composed of what he called "objective knowledge," which he defined as various products of the human mind expressed through artifacts of languages, the arts, and technologies (Popper, 1972). Brookes built on Popper's work and applied some of the concepts to the world of library and information science (LIS). He maintained that the theoreticians in LIS have the responsibility to (1) study the interaction between subjective ideas and objective knowledge; (2) describe and explain the interactions; and (3) organize the content rather than the containers of information, so that knowledge can be more effectively used (Soyibo and Aiyepeku, 1988). Florance and Matheson (1993), writing about the roles of health science libraries and librarians, emphasized the necessity for meeting information needs effectively in an environment of scientific communication, which they argue "revolve around the retrieval, creation, manipulation, management, and dissemination of new knowledge."

HIV/AIDS presents an exemplary model reflecting the synthesis of diverse social and scientific discourse. The body of knowledge concerning this disease continues to grow commensurate with the spread of the epidemic itself. The convergence of multiple disciplines, populations, and sociopolitical issues combine to form a discourse (and vocabulary) unique to this disease and its scientific and social constructs.

5. Description and Methodology

This controlled vocabulary was developed in response to a need identified within the AIDS service community for a means of organizing and making accessible the rapidly growing body of knowledge about the epidemic. Compounding the issue was the variety of formats in which the information was appearing, ranging from newsletters, to serials and monographs, to audiovisual materials, to Internet support groups. A further complication was the variety of users needing access and at the same time attempting to provide access to the information. This organizational scheme was designed for the latter group of individuals, most of whom are not trained librarians.

A test bed for creating and refining this work was provided initially by a large, community-based AIDS service organization. This field environment provided information services directly to health care providers, educational staff, care givers, and patients. The organization's resources included vertical files, videos, various serials and newsletters, and monographs. An iterative approach was employed, wherein alpha testing on each section took place; the system was refined; and the analysis was then repeated. As development progressed, gaps in coverage were identified and filled. From this ongoing effort, the vocabulary evolved.

Certain decisions made at the beginning of the process guided the development of the vocabulary. Central to the entire work was the notion of making the scheme as flexible as possible to suit the varying needs of potential users and to reflect the changing nature of the pandemic. It is anticipated that this tool will allow for the amorphous state of a vocabulary that is in constant flux. Within this context of a body of knowledge growing and changing at an epidemic rate, the following guiding principles were articulated:

1. Be specific yet general. Concepts should be specific to HIV or AIDS, yet broad in terms of illness, bodies of information, and approaches to organizing knowledge.

2. Use hierarchical system. Hierarchical arrangement allows for visual representation of the relationships of concepts within the domains. Further, since this vocabulary differs from other HIV-specific thesauri in its emphasis on biomedical terms, Medical Subject Headings (MeSH) provided a logical model for arrangement.
3. Be all-inclusive in terms of the various ramifications and implications of the disease. HIV disease is a multi-faceted condition that confounds efforts to define its boundaries. Therefore, the domains and hierarchies within them include broad concepts that cross disciplines including social science, religion, education, arts and letters, and biomedicine.

4. Provide for growth of the vocabulary. Many authors (Bierbaum and Brooks, 1995; Bierbaum, Brooks, and Brooks, 1992; Brooks, Bierbaum, and Brooks, 1990; Self, Filardo, and Lancaster, 1989; Sengupta and Kumari, 1991) have documented the epidemic growth of the vocabulary specific to HIV infections. Such growth continues unabated. Consequently, where possible, concepts such as "drug therapy" provide space to include names of specific pharmaceuticals as they become available. For purposes of consistency, no names of medications are provided in the work; rather, the "Instructions for Use" tell the target audience to add terms at their discretion.

5. Do not include names. Names of specific organizations and projects, historically or currently associated with the epidemic, are numerous and change frequently. As is true of drug terms, they may be added at users' discretion.

6. Adopt specific subheading system. In keeping with the use of MeSH as a model, terms that would clearly need to be used across domains are called "Universal Subdivisions." While not necessarily secondary in importance to other concepts, they may be appended to any term in the hierarchy to reflect a particular aspect of that idea. For example, age range categories serve to differentiate information about pediatric AIDS from adult AIDS.

7. Prefer plural forms of nouns, Anglicized rather than Latin terms, and normal word order. A desire for consistency, coupled with the needs of the intended audiences, determined these choices.

8. Provide large numbers of cross references. SEE and SEE ALSO references were included throughout to facilitate access to specific concepts.

Developing a tool for organizing the body of HIV/AIDS information involved numerous sequential tasks. Since the decision was made to create a hierarchical structure, the initial step involved identification of broad categories, or domains, representative of the concerns manifested by the disease. Domains served as the highest level in the hierarchical structure. Using the Dewey Decimal Classification as a model, ten inclusive classes were chosen. These domains were selected based on apparent foci within the body of knowledge concerning the disease. Domains were confirmed as the most important broad aspects of the body of knowledge by reviewing various HIV-specific resources, including medical textbooks, education curricula, personal narratives, and emerging electronic sources of information. Choice of the domains demanded adherence to certain criteria. Each domain had to be an "umbrella" concept that would succinctly encompass one aspect of the entire body of knowledge; be separate and distinct from all others; and be broad enough for categorical division.

Domains include Generalities; Epidemiology and Transmission; Education and Prevention; Clinical Manifestations of HIV and Complications, Malignancies, and Infections Associated with AIDS; Treatments, Therapies, and Medical Management of HIV Disease; Psychosocial and Religious Issues, Case Management; Legal, Ethical, Economic, and Political Aspects; Organizations, Funding Opportunities, and Health Policy; Fine Arts; Belle Lettres and Non-Fiction. Brevity of domain nomenclature was sacrificed for clarity, given the target audience and intended use. Domain groupings were consistent with the reality of the epidemic. For example, the three entities in the seventh domain—Organizations, Funding Opportunities, and Health Policy—are related in that most AIDS service organizations are dependent upon many levels and sources of funding and are bound by current health policy, policy and funding
that are often influenced by the very organizations they circumscribe.

Having discerned the highest levels of the hierarchical structure, the content within each domain was developed. HIV-specific concepts were identified using standard medical references including *The AIDS Knowledge Base* (Cohen, Sande, and Volberding, 1994), the *Textbook of AIDS Medicine* (Broder, Merigan, and Bolognesi, 1994), *AIDS: Etiology, Diagnosis, Treatment, and Prevention* (DeVita, Hellman, and Rosenberg, 1992), *Dictionary of AIDS Related Terminology* (Huber, 1993b), and *HIV/AIDS Curriculum* (Mountain-Plains Regional AIDS Education and Training Center, 1992). Specific terminology representing concepts was mapped to the domains and arranged hierarchically. At this time, the emerging vocabulary was compared to one of the only existing HIV-specific schemes. This early arrangement was developed by librarians from Philadelphia’s AIDS Information Network (AIN) and is non-hierarchical in structure. Concepts from the AIN vocabulary were added where appropriate. The National Library of Medicine’s *Medical Subject Headings* (MeSH) was used to supplement biomedical content development. The *Thesaurus of Educational Descriptors*, developed by information specialists at the National AIDS Clearinghouse, was used to verify completeness of the third domain, Education and Prevention. Subject specialists assisted with clarifying terminology in both the Religious Aspects and United States Government components of the scheme.

As work progressed, it became obvious that some concepts applied to more than one domain. Therefore, standard subdivisions (in this work, called *Universal Subdivisions*) were created. Some, such as age ranges, are generally applicable; while others, such as stages of infection, are HIV-specific. These terms reflect varied facets of main concepts and may be attached to them as needed. Categories of subheadings include Age Ranges, Sexual Orientation, Gender, Stages of Infection, Ethnic Groups, Geographic Names, At-Risk Populations, Religious Faiths, Signs and Symptoms, and Special Populations. (*At-Risk Populations* are groups of people whose behaviors or workplace tasks may place them at risk for HIV infection. *Special Populations* refer to groups of individuals directly affected by the epidemic.)

Finally, the authors generated an alphabetic listing of terms and wrote instructions for using the controlled vocabulary. The alphabetic index facilitates access to the hierarchical arrangement. Users unfamiliar with hierarchical schemes can go directly to the index and locate the term or terms they need to describe a given work. Cross references guide users to preferred terms within the vernacular. Instructions for use include examples for describing works using terms from the lexicon and combining them as needed to reflect various facets included.

6. Discussion

Discourse surrounding the AIDS phenomenon continues to evolve and requires that a nosological record be maintained in order to describe the collection of existing and emerging resources. Detailed description provides a key element necessary to organize the body of knowledge. As the epidemic matures, the continued accumulation and distillation of information will reflect a synthesis representative of the understanding of the disease and all its ramifications at a given point in time. This synthesis supports the depiction of accurate information that may allow for the iterative transformation to knowledge.

For example, early nomenclature used to describe the disease was tied directly to populations perceived to be most affected in the United States; i.e., gay-related immune deficiency (GRID), gay cancer, gay bowel syndrome, gay pneumonia. Eventually, the human
immunodeficiency virus (HIV) was identified as the causative agent of the constellation of signs and symptoms known today as the acquired immunodeficiency syndrome (AIDS). Over time, the pejorative terms were replaced with nomenclature descriptive of the diseases themselves rather than specific populations affected by them. This modified nomenclature is reflective of the expanded knowledge base that has developed over the course of the epidemic. Regardless of the standardization of the vocabulary, however, there is little consistency in its use (e.g., "AIDS virus" rather than HIV). This lack of consistency demonstrates that transformation from information to knowledge is dependent upon multiple factors and cannot be guaranteed.

Similarly, the legal realm offers evidence of an evolution in discourse. Since the beginning of the epidemic, legal discourse concerning disclosure issues, insurance coverage, adoption, medical malpractice, multiple end-of-life matters, and other topics has assumed a growing presence. Indeed, AIDS has generated more litigation than any other single disease in the history of the United States' legal system (Gostin, 1990). Despite the body of legal precedent regulating behaviors, practices, and procedures associated with HIV/AIDS, discrimination, abdication of responsibility, failure to adhere to guidelines, and other abuses continue to occur. Therefore, though the transformation from information to knowledge has begun (e.g., regulations governing blood donations), the process is impeded by the complex nature of the disease.

Owing to the basic science research performed as a result of the AIDS epidemic, the discourse of molecular biology, virology, immunology and other life science areas has expanded remarkably. In particular, scientists' understanding of retroviruses existed at a macro level prior to HIV/AIDS. Because of the compelling need to stem the spread of the disease, viral structures and their activity at the cellular level intrigued scientists world-wide to a new and urgent degree. While basic information about retroviruses was available, the epidemic served as an impetus to develop the knowledge base. Today scientists know how retroviruses insinuate themselves into cells, replicate, and mutate. This understanding theoretically facilitates the development of a vaccine, but the transformation from information to knowledge is not yet complete, since no vaccine yet exists to prevent HIV infection or to cure AIDS.

Indeed, no preventive vaccine exists to slow the epidemic, but prevention is possible through education (Osborne, 1989). The discourse necessary to implement this preventive measure varies from group to group, culture to culture. No matter the group or culture, however, a central tenet of the discourse focuses on what has come to be called "safe, safer sex." Since HIV is known to be transmissible through unprotected sexual activity, an entire discourse concerned with means of prevention has developed and continues to evolve. While safe sex prevention measures have been adopted by some individuals, HIV continues to be transmitted by unprotected sexual behaviors. Sadly, since the epidemic continues to claim more lives, the information available regarding this and other means of prevention has not yet been transformed into knowledge.

The body of knowledge surrounding HIV/AIDS continues to grow at an epidemic rate. Within the context of that body of knowledge, the transformation from information to knowledge has occurred in some instances, but not in others. The availability of a lexicon chronicling the evolution of the discourse facilitates maintenance of the nosological record and knowledge production. It also serves to record the social as well as the scientific constructs of the disease.
References


syndrome (AIDS) and the epidemic growth of its literature. *Scientometrics*, 17, 49-60.
Dewey Thinks Therefore He Is:  
The Epistemic Stance of Dewey and DDC

Abstract: The acceptance of a traditional Cartesian epistemology confined Melvil Dewey and his classification to a narrow consideration of knowledge while assuming the necessity of a universal language to describe it. The result is that the Dewey Decimal Classification marginalizes groups and topics outside of canonical knowledge. A feminist critique of Dewey's introductions to DDC and examples from The Electronic Dewey illustrate this problem. By taking a poststructural perspective, variation becomes theoretically possible and necessary for ethical practice.

1. Introduction

The Dewey Decimal Classification (DDC) structures the organization of knowledge in libraries according to its own world view. Editors of DDC address problems of marginalized areas, but because of the constantly changing nature of knowledge and local differences in focus their task is overwhelming. To address these problems globally and locally we must understand the underlying view of knowledge that has shaped DDC and consider other epistemic perspectives. In attempting this task I will first define the problem; second, explain my approach which is feminist and poststructural; third, identify Dewey's and DDC's epistemic stance; and fourth, suggest ways to develop strategies for ameliorative change.

2. The problem

DDC is a system for the organization of recorded knowledge primarily on the basis of subject covering the universe of knowledge in general as a universal language or code. From the perspective of DDC's user public, it is a narrowcast code—the signification of its notation is known to only a relative few. It is also an arbitrary code determined by a central authority. By its nature as a language or code it is a system with limits that define which subjects are centred and which are marginalized. DDC is not intended to cover the universe of knowledge exhaustively.

One parameter of marginalization is specificity—which subjects are treated in detail and which are not.

Another parameter is the de facto marginalization of subjects not fitting DDC's originating society. A.C. Foskett (1971) wrote that classificationists inevitably bring the bias of their times and themselves to the scheme. He critiqued DDC and other subject schemes with examples of women and sex. He later noted (1984) that despite significant changes the same generic problems continued to exist relative to "sensitive subjects." Jessica Milstead Harris and Doris Hargrett Clack (1979) discussed problems in both DDC and Library of Congress Classification (LCC) relative to Puerto Ricans, Chinese and Japanese Americans, Mexican Americans, Jews, Native Americans, Third World peoples, gays, teenagers, senior citizens, people with disabilities, and alternative lifestyles. Mowery (1995) stressed problems of application in the inconsistent use of LCC and DDC numbers for "Mexican American literature."

Problems of marginalization, exclusion and colonialist bias of Africana classification were addressed by H.O.M. Iwuji (1989) in religion, ethnography, social sciences, language and
literature, Philip Pacey (1989) in languages and literatures; and Michael Afolabi (1992) in African independent churches. The problems discussed on subject access to Africana are similar to those for other non-western literatures in terms of space and bias. Nabil Hamdy (1980) critiqued classification of Arabic materials, particularly relative to language, literature and religion. Fraiser McConnell (1984, 1985a, 1985b) investigated Melanesia in DDC’s Tables 2, 5 and 6 for geographic, racial, ethnic and national and language subdivision, describing inappropriate groupings and divisions, and the reflection of 19th century race classifications.

Steve Wolf documented (1972) how DDC lumped "gayness" with crime and sexual disorders, prostitution, pornography, character disorders, rapists, seducers, and perversions through context and references. Family and sex within marriage are set up as norms. Fran Steinberg (1974) noted improvement in the classification of women in DDC, but found many ghettoised classifications that implied a male norm. Ishbel Lochhead (1985) concluded that women continue to be treated as a minority with intellectual boundaries hindering classification of multidisciplinary women’s studies literature.

These and other issues have been addressed by the editors of DDC over the years with improvements. As a poststructuralist I see the task as not only infinite, but unachievable. However, it is possible to address the infinite problems of a finite system to produce manageable ameliorative change.

3. The Approach

My approach draws on feminist interpretations of deconstruction, the philosophy of language, and the politics of location. Feminist legal scholar Drucilla Cornell (1992) describes deconstruction as "the philosophy of the limit," recognizing any system's inherently constructed and exclusionary nature. From this recognition she concludes that replacing a system with another one is possible, but will still be exclusionary. Rather, we must make the system permeable, create spaces through which the voice of the Other outside of the limit has the opportunity to be heard inside it. Making these spaces requires strategies to address particular issues in particular situations. Such strategies can be devised for DDC.

Feminist philosopher Andrea Nye (1987) examined philosophers of language since Plato and concluded that all have believed the diversity and subjectivity of language must be standardized through a universal language. They ignore as unintelligible what is not or cannot be expressed in that language. Assuming a universal language is license to disenfranchise the marginalized. If we accept that DDC needs to be a universal language or code then we are free to ignore marginalized subjects.

Feminist theory has a manifold concern with situatedness: for example, the dichotomy of public sphere/private sphere; hierarchies that position some above others; the existence of women and other nondominant groups at the margins as opposed to the centre. Blunt and Rose (1994, 7) survey the development of feminist geography from spatial considerations of public/private (mainly relevant to white middle-class women) to an idealized global sisterhood to a postcolonial politics of location recognizing "that certain political projects construct spaces according to their strategic context and needs." This uncovers the previously transparent agenda behind spatial boundaries. "Mapping operates in hegemonic discourses as a form of mimetic representation—it textually represents the gaze through transparent space—but this form of mapping is contested in discourses of resistance. Mapping thus appears to be a spatial image that directly addresses the politics of representation as they are bound into the politics of location" (p. 8). Classification, especially in North American libraries with dictionary catalogues and open stacks, is a mapping of recorded knowledge. It physically (and electronically) structures collections for browsing,
determining proximity and distance, gathering and separation. It determines positionality.

4. Dewey's and DDC's Epistemic Stance

By epistemic stance I mean a view of or belief relating to the nature of knowledge and knowledge creation. Melvil Dewey and DDC both accept a concept of knowledge based in Cartesian epistemology. They accept that a single knowable reality exists and that we come to know it by discovering universal truth. If this assumption is so, then a universal language reflecting it is theoretically possible. Universality of language is an epistemic indicator of this tradition. The notational universal language of classification maps the topography of recorded knowledge structuring a space to reflect the single knowable reality and universal truth. Poststructuralism, on the other hand, rejects the possibility of universal explanations and accepts the possibility of multiple realities and multiple truths. If such a stance is employed, then DDC cannot be expected to be universal. Poststructuralism explains the exclusions and marginalizations of a "universal" language. The constructed system that is DDC reflects one or more realities and may, because it is constructed, be adapted to reflect others.

Did Melvil Dewey employ a traditional Cartesian epistemic stance in establishing his classification? Does DDC still tacitly accept this stance? I will seek Dewey's epistemic stance in his introductions to the first (1876—DDC1) and thirteenth (1932—DDC13) editions of the classification. Quotations are from DDC13 in Dewey's "simpler spelling." Parallel quotations in DDC1 are noted in square brackets. I will examine DDC in the form of the Electronic Dewey (EDewey, 1994).

Dewey manifested a traditional Cartesian epistemic stance in the necessity for a universal language:

Classification is a necessity if all material on any given subject is to be readily found. . . . By adopting the scheme in general use by libraries . . . numbers are in harmony with those of thousands of other catalogs and indexes in which the same number has the same meaning; . . . these numbers are the only international language of perfectly definite meaning among all civilized nations; and also cheapest and quickest in application. (DDC13, 43)

Some type of universal language is necessary for gathering all relevant documents. The classification brings order to documents by supplying a standard of sameness. It creates a one-to-one relationship between the number and its meaning. And as a universal language for use by the whole world it additionally achieves harmony and efficiency. For this result, Dewey seeks a one-to-one relationship with a place for each subject and each subject in its place. "All geometries are thus numbered 513, all mineralogies 549, and so throughout the library all books on any given subject bear the number of that subject in this scheme" (DDC13, 15; also DDC1, 3). The result is universality of location. Each book is treated the same as each other book having the same subject as an attribute—they are gathered together under one number always and under all conditions; not only a universality of location, but also a temporal universality. "Thus all books on any given subject stand together, and no additions or changes ever separate them" (DDC13, 22; also DDC1, 7). The subject will be gathered for all time.

Dewey's goal was to impose a universal language to overcome individual diversity. "Different librarians, or the same librarian at different times, classify the same or similar books in widely different places. Where one man did all the work for many years, there was a degree of uniformity; but even then there was danger of looking at the same book at different times from different viewpoints, thus causing confusion." (DDC13, 13) Difference is a threat to understand-
ing, resulting in contusion. A universal language is the answer to the confusion of diversity. This language is a notational language of base 10 numbers with the result that a relatively small range of options exists within each category. Because of this restriction some means of determining the relative importance of any given subject is required in building the classification and adding new subjects. Decisions have to be made to establish the lateral relationships between subjects: which subjects should be placed in proximity. The importance of subjects and their relationships to each other structure a body of knowledge, interpreting and constructing its meanings. Dewey used two means for structuring knowledge: canonical knowledge and expediency.

For canonical authority, Dewey began with the classes used by the St. Louis Public School Library: a reversed order of Francis Bacon's classification of knowledge (DDC 1, 10). Bacon identified three human faculties of memory, imagination and reason with the categories of history, poesy and philosophy. The Baconian modern scientist objectively distances himself from the object of his study, eschewing emotion (Keller, 1985, 33-42). Berwick Sayers (1926, 137-138) and John Comaromi (1976, 21-25) suggested that the relation between Bacon's classes and Dewey's is tenuous at best. However, Dewey follows Bacon's separation of reason and emotion, classifying memory/facts and philosophy/reason by topic, but imagination/poesy/emotion, the fine arts and literature, by the tangible aspect of form, not by the subjective attribute of topic. Dewey relied on contemporary experts to fill in his Baconian outline (DDC 13, 14). The result is a canonical map that reifies established perspectives—a hegemonic discourse in its politics of location, reflecting the authority of mainstream experts, who, like Bacon, contrast masculine science to feminine emotion.

Dewey's second mapping criterion is expediency which ultimately determines universal language. The order needed to avoid confusion must be given a structure that would meet other criteria. It must be comprehensible and efficient and, to achieve universality, it must have unlimited potential for expansion. To fulfill these criteria Dewey chose Arabic numerals in a base 10 or decimal arrangement. "Arabic numerals can be writn and found quicker and with les danjer of confuzion or mistake than any other simbols. . . . by exclusiv use of arabic numerals thruout shelvs, and indexes, catalogs and other records, there is secured the greatest accuracy, economy and convenience" (DDC 13, 26; also DDC 1, 8). Dewey accepted that this exclusivity introduces the possibility that the universal scheme does not appear to treat subjects equally. "Theoreticly division of every subject into just 9 parts is absurd. Practically it is desirabl to clas as minutely as posibl, without use of aded figures; and decimals, on which our skeme hinjes, allow 9 divisions as redily as fewer. This has proved wholly satisfactory in practis, tho apparently destroying proper coordination in sum places" (DDC 13, 16; also DDC 1, 4).

Dewey notes that the same theory, an absolute theory, will not work for every library. "Theory keeps numbers in strict sequence; but a hyer rule everywhere is 'sacrifice any theory for a substantial gain'. Practically there ar few libraries where it is not best to break order of clases." (DDC 13, 39) Local variation is not only possible, it is desirable, but only on Dewey's terms: "Even sum who hav uzed the sistem longest hav been misled into adopting chanjes which on tryal they wer compeld to reject, going bak to orijinal form at cost and confuzion of 2 chanjes. . . . The only safe rule is to make no chanjes or subdivisions without submitting them to the editor, who wil gladly advize on such matters without charj. . . . " (DDC 13, 34) Dewey's aversion to relocation has been adopted wholeheartedly by contemporary librarians who balk at change in the classification because of the cost of reclassification, regardless of the unquantifiable cost to the public of outmoded or inappropriate location.

Dewey uses two explanatory images of expediency: the businessman and the railroad. The businessman uses order to create meaning out of chaos.
A successful man is usually a classifier and chartmaker. . . . A large business or work unclassified or uncharted is not a worthy organization but mere material from which a clever brain may construct one. It differs in efficiency from the ideal as a mob of men differs from a well disciplined army. Piles of brick and mortar are not a temple any more than heaps of typ or Shakespeare's works, tho if 'classified' and set, each in right relation to the rest, the transformation is brought about (DDC13, 43-44).

With examples from business (and three other institutions based on canonical authority: the military, religion and literature) Dewey indicates that classification can overcome chaos: a "clever brain" can chart a classification and thereby create meaning.

Dewey accepts that universality and reason are necessary to make sense of chaotic diversity. He seeks to impose a universal language of classification on information characterized by mainstream Western scholarship. The resulting marginalized Others typically include non-Christian religions, non-European languages, and regions and countries with the least geopolitical influence. The politics of their location is justified by efficiency.

The scheme gives us for each topic, as it were, a case of 9 pigeonholes, with a large space at the top; and we use them as every practical business man uses such pigeonholes about his desk. If, as in 220, there are less than 9 main topics, it is often convenient to use the extra spaces for subdivisions. . . . Then in 280, having more than 9 topics, if we are using only 3 figures we put Congregational in same space with Presbyterian, and small denominations together in the last box, just as a business man puts his papers in his pigeonholes. If he insisted on having a different case made to order for each use, it would cost over twice as much; he could not group them together or inter-change them, and they would not fit on his shelves (DDC13, 21).

Expediency dictates structure. Dewey does not, however, discuss allocation of space for content in a broader sense (for example why does Christianity alone occupy 220-280 and all other religions together, 290?). He does not invoke literary warrant at any point. We are left with perceptions of how important a subject is in some absolute sense as the determinant of its pigeonholing, its spatial positioning.

Dewey's other image of efficiency is the railroad, another spatial image.

There has been perverse misapprehension of this feature, and critics oftenest stumble over 'procrustean'10'. In fact, this is an element of usefulness. A railroad also has the fault that it is procrustean in its path and in its times. It can not come to your door nor wait your convenience, as does the automobile; it can not go to the fields for its loads of produce; it can not turn out for obstacles; but because it is procrustean it can do its large-scale work much better and quicker and cheaper. The parallel could be fairly extended to many other cases, but any thoughtful mind will recognize that the economy and ease of working the Decimal system are largely dependent on its being procrustean (DDC13, 21).

Railroads map space differently than automobiles. They create stricter limits to access. Possible destinations are determined by economic discourse. Railroads flatten the landscape, cutting through hills, tunnelling through mountains, bridging valleys and avoiding curves. Automobiles are far more flexible. Streets and roads cover the landscape, leading to towns that have lost their rail service, to the front door. They, too, structure the map. Their economic discourse is related to vehicle ownership and they too have changed the landscape with suburban shopping malls, fast food "drive-thrus," and vast expanses of concrete and asphalt. Automobiles as an image are also
limiting. Either metaphor would be of a universal language based on the epistemic stance of a single knowable reality.

Does Dewey's stance continue in contemporary DDC? What does today's railroad look like? Probyn (1990, 178) offers a poststructural image of location as both defining and transient. Location orders knowledge according to established patterns, but it also allows that "any part of the world can be recreated or made to stand in for another... like in the movies as Canadian locals are dressed up in American location" (p. 183). In DDC ethnic and national identities are determined on the basis of a power politics of location. DDC Table 5: Racial, Ethnic, National Groups is described in the manual to the classification as preferring ethnic group over nationality. The exception is Canadian:

... for Canadians of French and British origin, the prescribed citation order is nationality first (T5--11 Canadians), then ethnic group: T5--112 for Canadians of British origin, T5--114 for Canadians of French origin... . . . In the absence of specific instructions to the contrary, however, use the citation order given at the beginning of the table, e.g., Canadians of Ukrainian descent T5--91791071 (not T5--11). Note that the same number is used for both Canadians of Ukrainian descent and persons of Ukrainian descent who are in Canada but not Canadian citizens. This lumping together of citizens and noncitizens is typical for Table 5 because of the low priority normally given to nationality; the developments for Canadians of British and French descent are atypical (Edewey).

This arrangement reifies the power of the federal government of Canada in defining nationality in relation to the "founding nations": Canadians of French or English descent. However, living as I do in a part of Canada in which Ukrainian Canadians are more numerous and prominent than French Canadians the example strikes me as made on arbitrary rather than sound historical grounds. Also, by privileging "founding nations" over "first nations," aboriginal peoples, it is a clear instance of Eurocentrism. English and French ethnicity have been "dressed up in [Canadian] location."

The manual in Edewey also describes my own ethnicity:

Normally the same number is used for both the majority ethnic group of a nation and the total population viewed as a national group, e.g., T5--94541 for both ethnic Finns and all citizens of Finland viewed as a national group. In such cases the question of priority between ethnic and national affiliation arises only for minority ethnic groups. Finnish citizens who are ethnic Swedes, a minority ethnic group, are classed in T5--39704897 (T5--397 Swedes + [notation from Table 2] T2--4897 Finland) because their ethnic group takes priority over their nationality (Edewey).

Finnish and Swedish are as much official languages of Finland as French and English are of Canada. Nonetheless, my "Swede-Finn" grandfather's history is classified separately from that of my other Finnish ancestors.

That DDC can accommodate variation is apparent not only from the Canadian example, but also from that for African Americans. "Special developments that allow expression of both ethnic and national affiliation are typically made only for the majority ethnic group in a nation, under the rubric 'national group'. . . An atypical development for United States Blacks (T5--96073) gives extra emphasis to nationality for a minority group, while still preserving the usual citation order of ethnic group before nationality... . . ." (Edewey) While African Americans are not grouped with the "majority" of US citizens, this exception allows for recognition of African-Amercanness in
African Americans who have relocated to other countries. Whether African Americans who have moved to another country prefer to be recognized primarily as of African descent via the US, or as citizens of the US who happen to be African American, or as Black people without national affiliation, is a question likely to produce a diverse answers. The uniform exception made here indicates the practical possibility of making exceptions where a need is recognized. These exceptions are to the generalization that the citizens of a country considered in general should be classified with the majority ethnic group of that country, an example of the tyranny of the majority that allows the majority opinion to override all others (Tocqueville, 1835/1966, 227-240).

In addressing the tyranny of the majority, the liberal approach of instituting equality has been followed more often than accommodating diversity. An instance is the general number for women in DDC:

<table>
<thead>
<tr>
<th>CLASS NUMBER:</th>
<th>305.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAPTION:</td>
<td>Women</td>
</tr>
<tr>
<td>NOTES:</td>
<td>Class here interdisciplinary works on women, on females</td>
</tr>
</tbody>
</table>

Add to base number 305.4 the numbers following 305.3 in 305.32-305.38 (Edewey).

The result is that women (305.42-305.48) are treated in the same way as men (305.32-305.38): the "men's liberation movement" becomes the model for the women's movement. The problem of equality is its homogenizing assumption that the same model will always apply. The men's movement, while sometimes simply a reaction to feminism, in its more productive form establishes that the gendered status quo is not a healthy situation for men. The women's movement has the inescapable difference that women are marginalized on the basis of sex, but in DDC the difference is erased. "The political notion of equality thus includes, indeed, depends on, an acknowledgment of the existence of difference. Demands for equality have rested on implicit and usually unrecognized arguments from difference; if individuals or groups were identical or the same there would be no need to ask for equality. Equality might well be defined as deliberate indifference to specific differences" (Scott, 1988, 44). Both men and women can be marginalized on the basis of other characteristics, and local situations govern the degree and nature of marginalization. "The sameness constructed on each side of the binary opposition hides the multiple play of differences and maintains their irrelevance and invisibility" (p. 46). The way gender operates in conjunction with other discourses of power is hidden by efforts at equal treatment.

Other examples of hiding the dichotomies follow from Dewey's suggestions on "viewpoints." He describes ways of using letters (not Arabic numerals) to add variations for local emphasis and convenience, and even a way to introduce what is clearly the viewpoint found in the "daily pres":

**Pro and con division of topics** It is very useful in many cases to separate books on a topic with strongly markt sides, so either set of views and arguments may be seen by itself. This has been dun in sum cases by subdivision, e.g. 337 Protection and free trade. In others it is equally useful, and can be indicated by an added mark, e.g. 324.3 Woman suffraj. The number may be uzed for jeneral works, giving facts etc. and advocates and opponents may be separated by + and - for positiv and negativ, or by p and c. (DDC13, 39)

The examples of women's issues and free trade make it clear that we cannot discount his views on the grounds that things were less complex in his time, or that he could not anticipate the kinds of problems faced in classification today. Protection and free trade are given separate subdivisions as major economic issues, while women's suffrage is given one number. The library must tinker
with the result to show different perspectives—unequal issues in his day as in ours.

Another way to consider the issue of equality and difference is to treat different groups differently. An example is the number for women workers: 331.4. The subdivision of 331.4 includes some of the problems peculiar to women in the paid workforce (such as maternity leave). There is no parallel section for men—men are the norm and women the exception. This section on workers also demonstrates that some differences are privileged over others. "331.3-331.6 Labor force by personal characteristics" instructs: "Unless other instructions are given, class complex subjects with aspects in two or more subdivisions of this schedule in the number coming first . . ." (EDewey). Age as a difference takes precedence over sex, sex (if female) over ability, and ability (if disability) over race. The young Chinese women workers (DDC's example) will always be allied with young people—not with other women or other Chinese workers. In this way, DDC creates a canon of differences.

Further examples showing the depth of DDC's dichotomizing occur in the sections dealing with metaphysics (110-119) and epistemology (120-129).

CLASS NUMBER: 111/.82
CAPTION: Unity
NOTES: Including part-whole relationships
DDC INDEX TERMS: Identity (Principle of); Part-whole relationships; Unity

CLASS NUMBER: 113/.8
CAPTION: Philosophy of life
NOTES: Origin and nature of life . . .
DDC INDEX TERMS: Death--philosophy; Life--philosophy; . . .

CLASS NUMBER: 122
CAPTION: Causation
NOTES: Class here chance versus cause . . .
DDC INDEX TERMS: Causation--philosophy; Cause--philosophy; Effect--philosophy

CLASS NUMBER: 128/.1
CAPTION: Soul
NOTES: For origin and destiny of individual souls, see 129
DDC INDEX TERMS: Body and soul; Soul; Soul--philosophy

CLASS NUMBER: 128/.2
CAPTION: Mind.
NOTES: Including mind-body relationship
DDC INDEX TERMS: Body and mind--philosophy; Mind--philosophy; Mind (EDewey; emphasis added)

Two things emerge from these five examples. First, they portray concepts as oppositions. Other entries in this range dichotomize "finite and infinite," "goodness and evil," "possibility and limits," "subjective and objective," and "determinism and indeterminism." Opposites are collapsed into one number regardless of standpoint. However, these five examples go further to hide the subordinate of the binary under the dominant. Works that treat the two together are spatially mixed with the dominant. There is no number for the philosophy of the body alone. Numbers for the body occur only in the 300s, 500s and 600s. The standard subdivision for philosophy can be added to these numbers, but the body is mapped far away from the mind and mind/body relations. Dewey's premise that "Not only are found together all books on subject sought, but most nearly allied subjects precede and follow, they in turn being preceded and followed by other allied subjects as far as practicable" (DDC13, 22 [DDC1, 7]) is observed in this situation only if mind/body relations are
considered allied to the mind and not to the body.

Other considerations might also be included in this analysis are much like the procrustean railroad keeping the classification on the rails instead of straying like the locomotive in the children's story, *Tootle*, to smell the flowers and play with the animals in a nearby meadow. These examples seem sufficient to point out that the Cartesian epistemology of a single reality knowable through reason—Dewey's stance—is still reflected in *DDC*.

5. Developing Strategies for Ameliorative Change

It is easy to critique a structure—in accepting a poststructural epistemology I reject the possibility of a universally applicable structure. This my epistemic stance does not reject the possibility of ameliorative change, but calls for changes that are local, partial and dynamic. To develop strategies for such changes I return to Cornell's philosophy of the limit. I seek ways of making the limits of the system open for the voices of marginalized Others, strategies for permeability, in two contexts: *DDC* and its local applications. These contexts are crossed by two types of action: deconstructive critique and reconstructive change. The result is a reflexive process occurring constantly and synchronically.

Global change includes the responsive adjustments that took women out of 396, lodged between etiquette and outcast races, and put us into 305.4 with other demographically-defined groups. A more encompassing type of change is that which sets up options for local definition. Some such provisions already occur in *DDC*, as in the 800s options for Canadian literature. Similar options could be readily devised for giving primacy to the histories of different countries (perhaps colonial powers and former colonies could trade places). More radical changes can be introduced gradually with libraries following the suggestion from *DDC20* (1989, xiii) to "implement the revised schedules and tables on an ongoing basis, instead of in more massive projects." Other changes might allow for mapping works about native peoples instead of dispersing them geographically by white people's geography. Different sequences could be used simultaneously, designated by letters or other symbols with precedent, again, in the 800s. Established methodologies such as citation analysis can inform remapping interdisciplinary territories, escaping the limits of mainstream disciplines and allowing in the voices of marginalized authors.

Local libraries could easily privilege different differences: gathering by race or ethnic origin is likely to be of more use in some libraries than by age. Young Chinese women workers could be located with other Chinese workers simply by a local policy giving precedence to nationality/ethnicity. Facets present only in the tables, such as gay and lesbian workers, could be given priority through a policy to place those facets first. Or precedent could be determined document by document according to the cataloguer's judgment (as in the use of 970.004 for specific American native peoples). However, judgment can be informed not only by the content of the document, but also with the library's users in mind. Each of these options requires libraries to reclassify existing collections and to edit copy for incoming materials. If libraries have a serious commitment to serving their users the workload imposed by such changes will be cost effective in terms of improved use.

Canadian history and literature variations of *LCC* are examples of successful adaptations for which libraries share copy, making application more economical. Such variants need to be restricted to geographical Otherness. Options could be devised such as rewriting the philosophy section to reflect the holistic, rather than dualistic, perspective common in North American native cultures.

Another departure would be to reintroduce a version of the classified catalogue to North
America. It would not include serendipity in physical browsing, but with the union catalogues that make up the "virtual library" multiple classified entries would give wide scope to electronic browsing.

Tootle, the naïve young locomotive, had to learn why the rails were necessary. He had to be tricked to stay on the tracks so the soup in the dining car wouldn't spill, so the wheels wouldn't get clogged with grass, so the destination would be reached on time. Our electronic setting could allow the locomotive to cross the meadow without spilling the soup. We might need to invest in reclassification as ongoing maintenance for browsing to periodically clean the grass out of our wheels. And perhaps we can learn to cope with being a bit late since no system can be perfect and we would rather have one that smells the flowers along the way.

References


Cataloging & Classification Quarterly, 5/3, 57-66.


Ontology for Knowledge Organization

Abstract: The first main thesis of the paper is that an ontology is not a catalogue of the world, a taxonomy, or a terminology. If anything, an ontology is the general framework within which catalogues, taxonomies, and terminologies may be given suitable organization. The second main thesis is that reality is organized into diverse levels and there are sophisticated dependencies among these levels and within them. An acceptable ontology should be able to model all these relations of dependence.

1. Introduction

In the literature, the term ‘ontology’ presents a variety of meanings and is used in many different ways. To verify this observation one need merely consult any dictionary of philosophy. As well as the variety of meanings entrenched in the philosophical tradition, new variants have been recently forthcoming from the database community and from certain sectors of natural language analysis. For example, Sowa states that, “The first step in designing a database, a knowledge base, or an object-oriented system is to select an appropriate collection of ontological categories” (Sowa, 1995). (For discussion of the relationships between artificial intelligence (AI) and ontology see Guarino and Poli (1995). More in general, on ontology and with particular regard to formal ontology, see Burckhard and Smith (1991), Poli (1992), and Poli and Simons (1996)). Apart from their multiplicity, many of these diverse meanings are accompanied by hidden assumptions and research strategies which are only occasionally made explicit. With regard to ontology, moreover, for some curious reason suggestions that the development of ontological analysis might be advisable often provoke extreme and contrasting reactions. While some regard ontology as a panacea for an extremely wide range of problems, others adopt entirely the opposite view and deny that ontological inquiry has any sense at all. If on the one hand there is often a desire to ‘capture’ the world in a definitive set of categories, on the other there is the opinion that the endeavor is impossible and indeed nonsensical. In fact, both points of view illegitimately overstretch the argument. The proposal advanced below seeks to avoid both baseless enthusiasm and delegitimating rejection. For simplicity’s sake, I shall proceed by advancing a number of theses.

2. A Number of General Theses on Ontology

THESIS 1. An ontology is not a catalogue of the world, a taxonomy, a terminology or a list of objects, things or whatever else. If anything, an ontology is the general framework (= structure) within which catalogues, taxonomies, terminologies may be given suitable organization. This means that somewhere a boundary must be drawn between ontology and taxonomy.

THESIS 2. An ontology is not reducible to pure cognitive analysis (in philosophical terms, it is not an epistemology or a theory of knowledge). Ontology represents the ‘objective’ side (= on the side of the object), and the theory of knowledge the subjective side (= on the side of the knowing subject) of reality. The two sides are obviously interdependent, but this is not to imply that they are the same (exactly like the front and rear of a coin). In order to conduct ontological analysis, it is necessary to ‘neutralize’, so to speak, the cognitive dimension, that is, to reduce it
to the default state. I assume that the default state is the descriptive one, where the dimensions of attention, of interest, etc., are as neutral as possible (‘natural’ attitude). It is of course possible to modify the default state and construct ontologies of the other cognitive states as well, but this involves modifications of the central structure.

**THESIS 3.** There is nothing to prevent the existence of several ontologies, in the plural. In this case too, ontological study is useful because, at the very least, it renders the top categories explicit and therefore enables verification of whether there are reasonable translation strategies and of which categorization can serve best to achieve certain objectives. This thesis is a subtle argument and requires some further comment. Let us assume that the general space of the ontology is a virtual space which is actualized in its concrete instances. The situation can be exemplified by citing the case of natural languages. The intended reference is obvious here and implies that the universal ontology is like the universal language. The least that one can say is that neither of them exists. Just as there is no universal language, so there is no universal ontology. What exists instead are individual natural languages, each of which is general; that is, it is able to say everything that it wishes to say. By the same token, there may not be a single universal ontology, but there may exist several general ontologies. Also in this case (which is the one that I regard as closest to the truth), structural studies of great interest are possible. For the sake of clarity, I shall continue with the example of languages. However different various languages may be, they display a number of highly significant uniformities. If we divide the structure of a language between lexicon (= content) and grammar (= structure), we note that there are dimensions which, with by and large limited variability, are systematically treated by the lexicon or by the grammar. For example, some languages have singular and plural, others add a dual form, others a trial form, others can grammatically indicate the many and the few. But that is all. No language has grammatical devices with which to denote the countable or 435. In these cases they resort to the lexicon. What I wish to say is that, for all their differences, languages display certain structural uniformities. The same should apply to ontologies. However different the various general ontologies may be, some structural uniformities among the top categories should nevertheless exist (Talmy, 1988). If this argument holds, two prime focuses of attention ensue from it:

(a) the construction of a general ontology to be used as a test framework;

(b) the theoretical and experimental analysis of the process of multiple categorization.

The first point is obvious. Without a general ontology we have no context of reference. The second is much more delicate. If the categorization is intended to be a passage from a substratum space to a target space (possibly discrete), the problem becomes that of translating objects categorized according to the categories of one particular target space into objects categorized according to the categories of another target space. One of the reasons for the difficulty of the problem is that we often find that we know only target spaces and not the substratum space whence they originate, of which we must hypothesize the characteristics—as far as we can—from the target space or spaces (Petitot, 1985).

**THESIS 4.** Reality is organized into diverse levels and there are highly sophisticated interdependencies among these levels and within them. For instance, an adequately sophisticated ontology must be able to perform the following two tasks at the very least:

(a) distinguish among such diverse objects as the things or inanimate concrete objects of the material world, organisms or animate concrete objects, the psychological objects of the mind, the social and institutional objects of social life, the abstract, ideal and fictitious objects of the sciences and arts.

(b) connect the various levels of reality by means of opportune forms of dependence. A promise as a social act with juridical force requires the existence of people with juridical capacity,
and that people do not fluctuate in some hyperspace but possess living bodies, which are in turn the result of particular supraformations of material entities (Hartmann, 1966).

3. Levels of Reality

The universal glue of any whatever articulated ontology is given by the network of dependencies that it creates and governs. We may distinguish at least five ontological levels: of the inanimate physical world, of the animate physical world, of the psychological world, of the social world and of the ideal world.

Specific forms of categorical and existential dependence exist among these levels. For example: a psychological object or event requires an animate physical object as its existential bearer. Should there be no person (and should there be no body of some such person), then neither will there be the correlative psychological states. Hypothesizing forms of existential dependence does not signify resorting to more or less overt hypotheses of reductionism. The various ontological levels may be existentially constrained without this implying that they are categorically constrained. The categories or properties which enable us to describe the world of psychological states are different from the categories that enable us to describe the world of animate objects and that of inanimate objects.

In more analytic terms, we may hypothesize that (a) new categories may intervene in every passage from one level to another, and (b) the lower-level categories that reappear at a higher level undergo transformations, they are ‘supraformed’.

The supraformation relationship is that which inheres, for example, in the passage from the inanimate material world to the animate material world. The categories of the former maintain their validity also for the second, but in a different form which reflects the specificity of the new level.

Differing from the supraformation relationship is the supraconstruction relationship, which holds, for example, in the passage from the animate material world and the psychological world. In this case, the higher level requires the lower one only as its external basis of existential support, but not as material to be supraformed (in cases like this one, the substratum of the higher level is not the material of the lower level).

There are several kinds of dependence relation among the various levels. For example, every higher level not only requires corresponding lower levels but also that these must have specific structural characteristics (of complexity, for instance).

In general, we may distinguish the general categories, which hold for all the ontological levels, from the regional categories, which apply only to certain levels. The general categories comprise, for example, object, event, substratum, substance, form, relation, determination, dependence, structure, part, whole, unity, multiplicity, dimension, continuum, discrete, internal, external, identity, diversity, possibility, actuality, necessity, change.

Examples of regional categories are the following. For the material world: space, time, cause, situation, reciprocal action, dynamic structure, dynamic equilibrium, becoming. For the animate world: organic structure, adaptation, end-directedness, material exchange, self-regulation, life of the species, degeneration of the species. For the psychological world: act, content, consciousness, unconsciousness, pleasure, displeasure. For the social world: social system, family, community, conflict, class, institution, integration. For the ideal world: the categories of the activities and products of knowledge, art and faith.

These lists are obviously provisional and their purpose is purely exemplary. A thoroughgoing ontology would have to refine them by making appropriate changes and studying the connections among the various categories.
This proposal for a stratified construction of ontology can be summed up in the following theses:

1. The general categories encompass all the various ontological layers. However, they display features at every level which are specific to that particular level (because they interact with the complex of categories at that level).

2. The categories of the lower ontological layers are the foundation for the higher ones, but they are independent from the higher categories.

3. The categories of the lower layers are stronger than the categories of the higher layers, but they have lesser structural power.

4. In the case of supraformation, the whole categorized by the lower categories helps to constitute the substance of the higher level.

5. In the case of supraconstruction, the whole categorized by the lower categories acts as the existential bearer of the higher level.

4. The Trap of Simplicity

An epistemological thesis which finds wide acceptance is that of theoretical simplicity. It is often said that, on the assumption that a procedure exists for calculating the complexity of a theory, of two theories T1 and T2 relative to the same domain the simpler one is preferable. However obvious this thesis may be, it is potentially highly dangerous for any enterprise of an ontological nature.

The reason is that the central problem of ontology is not so much the categorization of any domain whatever as the categorization of the maximal domain. What is needed, therefore, is not the simplest theory for any domain whatever, but the simplest theory for the maximal domain.

Accordingly, it is almost certain that the simplest theories for local domains are also those which most resist expansion to embrace broader ones. From an ontological point of view, the most useful theory is the one that can be expanded in the most economical manner possible.

We can reformulate the thesis by stating that it is advisable to spend more on constructing local theories because in this way less need be spent later when integrating and coordinating them.

If we take any two basis of data, there is no guarantee that their data will be interchangeable. The ontology (once constructed) must constitute the context into which the two data banks are merged, thereby modifying the data in order to render them interchangeable. The translation will be simpler, the more adequate is the structure of the initial basis of data.

5. A Word of Caution

In constructing an ontology it is important not to confuse ontological problems with those involved in their formal translation. In other words, care must be taken to distinguish between the ontological tree and the logical tree that should be its rigorous translation. I see at least three reasons for keeping the two trees distinct. First, the categories that constitute the two trees are not the same, and in passing from one tree to the other changes may occur of which one should be aware, also because there is nothing to guarantee the neutrality of the translation. For example, from the ontological point of view, 'property' and 'relation' are very different entities. From a logical point of view, however, it is obvious that the former is a monadic relation. This type of translation, therefore, may be anything but innocuous.

Another reason for keeping the level of ontological categories distinct from that of the logical categories is that there may be different logical translations of the same ontological structure. These different translations may prove to be entirely compatible; but they may equally
be incompatible and usable in different ways according to the particular purposes of the moment.

Thirdly, at the actual state of research, apart from the fact that there is no reason for believing in a universal ontology, there is nothing to guarantee that one single logic applies to the entire ontology. In other words, there is nothing to rule out that the different sections of the ontology may not be more efficaciously formalizable using different logics. In short, the ontology and the logic (or at least the formalism) which should give it formal rigor lie at different levels which should not be confused. Both must be addressed, but each \textit{iuxta propria principia}.

6. The Perspective of the Whole

On the basis of the foregoing discussion it is now possible to outline a hypothesis of ontology.

I assume the category of whole as the top level. Synonyms frequently used for this category are: entity, thing and object. I prefer whole because it highlights from the start a crucial aspect which is not stressed with equal emphasis by other proposals: being a whole means having boundaries. Boundaries which may be more or less rigid and of different kinds. The type of boundary that may be relevant on any particular occasion depends on the level of the whole. Although inanimate, animate, psychological, institutional, abstract, imaginary, etc. wholes may have different boundaries, they always have boundaries.

By virtue of possessing boundaries, a whole is something on the basis of which there is an interior and an exterior. Put in different but not alternative terms, we may also say that a whole is something which displays some form of independence with respect to an environment. Observing that, when analyzed at a sufficient level of detail, every whole vanishes into a continuum, or according to which every whole depends on something else, does not raise major difficulties. The fact that the boundaries of the whole are not absolute does not imply, in fact, that these are purely apparent boundaries. Wholes and their boundaries are realities which effectively operate at the appropriate level of granularity.

In dealing with wholes composed of other wholes, the problem arises of calibrating the 'weight' of the more general whole with respect to the 'weight' of the boundaries of its component wholes. For these cases I propose the following hypothesis: for wholes constituted by several wholes: the boundaries of the wholes of the ontologically lower layers are stronger (= more stable) than the boundaries of the wholes of the higher ontological layers. Both the sheep and the flock of sheep are wholes, but the boundaries of the individual sheep are stronger/more stable than those of the flock.

My analysis of the whole now proceeds by drawing on two groups of theories: those that give information about the whole as such, and those which give information about the various ways in which the whole can be considered.

7. The Whole as Such

Analysis of the whole as such examines level, location and amount.

By 'level' I mean the ontological layer of the whole and therefore whether it is an inanimate material object, an animate object, a psychological object, a social object or an ideal object. This classification activates the appropriate forms of existential dependence and the sets of specific categories which hold at that particular layer of reality.

By 'location' I mean the position of the whole in one of the semantic fields of space, time, matter and substratum (or existential bearer). These four semantic fields are appropriately activated by the level. For example, in the case of wholes of inanimate material type, one finds
that matter and substratum coincide; whereas certain ideal wholes do not have a temporal location, etc.

‘Amount’ tells us that we are dealing with only one whole, with several wholes, or with material denoted by mass terms. I assume that the default case is the one involving a single whole. If instead several wholes are given, then new categories are activated, like those of relation and force. By way of a further example: significant applications of the category ‘relation’ are those which concern the relations among wholes in terms of their location. In the case of two wholes, for example, relation < pattern < spatial < schematic < behind/in front; above/below; right/left; diagonal. Etc.

The appropriate categories are activated in the case of matter denoted by mass terms.

8. Modes of Considering the Whole

The whole has parts—which I distinguish into separable and non-separable—and a history. Let us first inspect the distinction between separable and non-separable parts. However difficult it may be to specify this distinction, a preliminary definition of the difference may be forthcoming from examination of the case of inanimate material objects. In this situation, we may call separable parts those which can be removed from the whole without anything else taking their place. We may therefore say that a leg is a separable part of a table. The removal of separable parts may have a destructive effect on the whole. Non-separable parts are instead those which can be recognized and distinguished but which cannot be removed (Husserl, 1970, 3rd Investigation). When they are ‘removed’ from the whole, a part of the same kind usually takes the place of the part that has been removed. In effect, what is really inseparable is not so much the individual part qua part as its genus. If I remove the color ‘red’ from a table, it becomes of another color or it assumes another color, but it is still in some way colored. For categories like color, weight, shape, size, consistency, etc., material objects are structured in such a way that they may display different instances of these categories in the course of their histories, but they nevertheless always have them. A table may be of different shape, color, weight, consistency, etc., but it will always have a certain specific shape, color, weight, consistency, etc.

With the appropriate variations, the difference between separable and non-separable parts holds for every type of whole.

The different ways in which separable parts are given enable us to provide a classification of the types of whole. We may accordingly speak of systems when the separable parts are given simultaneously and the boundaries of the parts are less strong than the boundaries of the whole. We may instead speak of aggregates when the separable parts are given simultaneously but the boundaries of the parts are stronger than the boundaries of the whole. We may then speak of events when the parts are given in succession and the boundaries of the parts are weaker than the boundary of the whole; reserving the term process for the case in which the parts are given in succession and the boundaries of the parts are stronger than the boundary of the whole.

As for the history of the whole, I distinguish between the history of the whole as such and the history of the location of the whole, according to the modification and/or substitution of its parts. The histories relative to the location of the whole and to the modification or substitution of its parts enable us to establish various criteria of identity (material, functional, etc.).

9. Conclusions

I have tried to show that constructing a general ontology requires the attention to be focused on—besides the problem of the theoretical and experimental analysis of multiple
categorization (§ 2)—the theory of dependence. My ontological proposal has been an attempt to articulate at least two of the many and diversified forms of dependence among wholes (as dependencies among ontological layers and among levels of one layer) and internally to one whole (as dependencies between the whole and its various kinds of parts, and between the whole and its history and that of its parts) (Ingarden, 1964).

References
L’Apparition du Computer: Epistemology and the Impact of Networked Computers on Society

Abstract: If we are to learn about the impact of computers and networks on society, it will be necessary to address issues from broad cultural-historical perspectives, such as has been done for print culture by those in the histoire du livre tradition. There are paradoxes faced by users of the Internet that have direct implications on their conceptions of the organization of knowledge. Perceptions of knowledge structures may play roles in searching habits or in deciding about the overall appropriateness of a Net search. The Net has been compared to a world brain and is here placed in the context of an early conception of a world brain.

1. Introduction

Human conceptions of knowledge are based on socio-cultural environments and structures of information sources. Some structures are institutional, some personal, others related to information retrieval systems. When confronted with the Internet and its various interfaces (known also as the Net, or through hypertextual or graphical interfaces as the World Wide Web, or simply the Web), users are forced into an almost chaotic realm of complex relationships between and among sources and into an unpredictable array of searching methods. That the entire process of research, if one can call it that, is disconcertingly difficult to evaluate should be clear to anyone who has used the Internet casually or formally. Whether a search has been exhaustive or even efficient remains ultimately unknowable. Closely related to evaluative measures are users’ perceptions of the nature of knowledge, in general and on the Internet; for some users, the Net may serve as their conceptual model of knowledge, its nature, form, likelihood of existing, organization, reliability—in short, a kind of personal epistemology. The “coming of the computer” has had radical effects on society’s perceptions of knowledge—perhaps not on the nature of knowledge itself.

There are several questions that should drive enquiry in this area and they are the issues that direct this paper. By stepping back from the positions of contributor or user, it would be desirable to examine the Internet from a broad perspective, perhaps on the order of two landmark works in the histoire du livre tradition (Febvre and Martin, 1956; Eisenstein, 1979). Second, it is important to view the advances in networked communication and computing as reflections of seemingly paradoxical realities, which may have direct effects on perceptions of the Net and which may help us understand more about the impact of networked computers on society, particularly on conceptions of the nature of knowledge. We are not necessarily dealing with knowledge itself here; rather we are concerned with people’s perceptions of it. Third, insofar as we may be concerned about the ability to "know" only that which is "true," we should be concerned with evaluative measures of Internet sources and the Net itself as an information retrieval system, or set of retrieval systems. This paper addresses some issues of information retrieval, particularly the feasibility of measurements of recall and precision. And lastly, it may be useful to place the Web in a context of one particularly perceptive—and early—approach.
toward the conception of a world brain. At the risk of paralipsis, a necessarily superficial
treatment of this topic, much of significance has to be omitted.

2. The Coming of the Computer and Networked Communication

To what degrees have computers and networks had an impact on society? In certain
arenas the effects have been profound, in others there has been no effect at all. To paraphrase
Febvre and Martin (1958, 12): The story is about something other than the history of a
technique. It has to do with the effect on late-twentieth-century culture of a new means of
communicating ideas within a society that has been essentially aristocratic, a society that has
accepted a culture and a tradition of learning which has been restricted to certain social groups.
An elite has once again been served.

The first users of information technologies have always been members of an elite of some
sort, whether from scribal cultures, the world of early printing in Europe, or in the Web
culture of today. Tied closely to literacy in each case, the elite also communicated in non-
vernacular tongues. The printing press was an agent of change in this regard: "The unified
Latin culture of Europe was finally dissolved by the rise of the vernacular languages which was
consolidated by the printing press" (Febvre and Martin, 1958, 332). In the case of computers,
networked or not, there has been said to be a lingua electronica—or perhaps several
(Compaine, 1983, 18) that was understood by a small group at first, then by more, to be
replaced ultimately by vernaculars.

In scribal cultures, oral transmission was highly valued by literate elites (Eisenstein,
1979, 11). Eisenstein maintains that in such an environment, learning "was governed by
reliance on the spoken word—producing a hybrid half-oral, half-literate culture that has no
precise counterpart today" (Eisenstein, 1979, 11). With the Web, the hybrid may have been
realized: great portions of the Internet consist of newsgroups, discussion rooms, bulletin
boards, e-mail -- mostly unfiltered, unedited, quasi-oral, and some quite illiterate. Other parts
represent the cutting edges of scientific and humanistic scholarship -- scrupulously edited,
carefully selected, and understandable by only the most literate of the relevant disciplines.

The relationships between the Internet and various levels of society would be as useful
to investigate as have been the relationships between printing and society. There are of course
many specific subtopics worthy of attention. An evaluation of the sources to which the Internet
is connected would be in order, but would be much more complicated than categorizing the
items listed in Hain's Repertorium bibliographicum or the Gesamtkatalog der Wiegendrucke
into disciplines: the sheer number is prohibitively large and growing; perhaps thousands defy
categorization. We have numerous studies of incunabula that tell us much about printed output
before 1500, but what have we for Web sites? Ephemeral "Internet Yellow Pages" may help.
To be sure, Hain carried out his research centuries after the period he studied. What will
remain of today's Web in 400 years, or in 10?

How has the Net preserved the moral, religious and literary heritage accumulated by its
predecessors from the eleventh to through the twentieth century? How has it ensured the
continuity of a tradition which has linked us with the past? Conversely, how successful has the
Internet been as an agent for the propagation of new thoughts? What role has the Net had in
the contagion of political ideas? How has it, by design or accident, helped spread English
around the world even more than it already has been? Are there other relationships between
language and the Net? What effect has the Net had on education, business, tourism, or
personal enrichment?

Cultural bibliographers have studied the geography of the book. An historical geography
of the Internet showing changes, perhaps on a yearly basis would be enlightening. An
historical Net demography, describing characteristics of the population of Net users and contributors would likewise be useful.

A reading history of the Internet could be fascinating. What literary forms might it claim responsibility for spreading? How has hypertext changed the way readers or authors think about linear texts? Which literary forms has the Net resisted? Clearly within bounds of cultural bibliography would be studies of Net economics. To what extent does trade exist on the Net? How have conceptions of profit changed since the early years of printing? Some scientific works with limited market appeal existed in manuscript long after printing was introduced. On the Internet, items with very low market value, some would say with no market value whatsoever, can be mounted easily by individuals. How can popularity be measured? For some printed editions it is possible to determine the size of print runs; it can be more difficult with Net sources. How often is a source accessed, read, or copied? Although some Web sites maintain usage statistics, many do not. Who has access to the Internet? Has it had an impact on all of society or only on parts? Printing had an effect on an elite with a subsequent effect on larger groups. Does the Internet reach larger groups in the same way? What are differences in age, gender, academic discipline, socio-economic status, and level of computer literacy that can tell us something about the relationship between the Net and society?

Modern researchers face a major problem not encountered by Febvre and Martin or Eisenstein: the Net has not been around a long time. Is it possible to gain an objective perspective when one is so close to one’s subject? For example, Febvre and Martin examined the relationship between printing and the spread of religious ideas between Catholics and Protestants, and other later groups: Freethinkers, Deists, Atheists, and Materialists. If such relationships exist between the Net and the spread of religious ideas today, they are difficult to observe. We are faced here with a conundrum: we do not have the perspective of time that can be so useful (conversely, we do have the advantage of being eye witnesses), yet it is all the more important to record these things now because of the ever-changing nature of the Net.

There is a certain ambivalence in the Web subculture toward the cultures it believes it has replaced. This follows a pattern from the Akkadians, who replaced the Sumerians, to newly-elected politicians of today, who often want to distance themselves from their predecessors. For this reason alone, it is not too early to call for an updated approach to cultural bibliography, a new incarnation of the *histoire du livre* movement, in order to come to terms with the impact of this revolutionary force on society.

3. Web Dialectics and Perceptions of Knowledge: Surprises and Disappointments

It would be unwarranted to adopt a metaphysical system that attributes an empirically sound reality to perceptions of objects or of environments, physical or electronic. Knowledge is different from perceived knowledge; the structures of recorded knowledge are not the same as perceptions of such structures. Yet it can still be a useful exercise to wrestle with paradoxes of the Internet that are present in a working conception of the Net, and likely present in an objective reality. The following propositions would seem absurd in most information retrieval systems:

- The Net or its sources are permanent and ephemeral;
- static and dynamic;
- intuitive and counterintuitive;
- quick and slow;
- predictable and unpredictable;
systematic and unsystematic (yet not random);
organized and entropic.

These dichotomies exist on the Net, are apparently mutually exclusive, but may in fact express truth. Some of these may influence the construction of users' mental models of the organization of knowledge. The Internet and the electronic sources to which it provides access are not knowledge; they do not know anything. The Net and the sources together amount to a large information retrieval system. Here we are interested in how recorded knowledge or communication have changed since the advent of the Internet and in how perceptions of recorded knowledge or communication have changed. We may be puzzled about this set of epistemological problems: the nature, scope, limits, and perceptions of human knowledge in an electronically networked society. There are at least two assumptions made in this paper about perceptions of Net knowledge. First, in using the Net we are limited to almost no "knowledge by acquaintance" except for direct knowledge of the Net and its interfaces. Second, while it is true that humans are limited to "true" knowledge because one cannot have "knowledge" of that which is not true, perceptions of knowledge, such as those that Net users may have, exist independently from the presence of truth. People hold beliefs for a variety of reasons, of which empirical thought is only one.

At early stages of some users' acquaintance with the Net, they may not have formulated a formal mental model of knowledge organization in that system. At this point they might describe the Net and its resources as disorienting, overwhelming, confusing, or amazing. In time, users may well continue to use these descriptors, but they will have acquired a more structured model of knowledge as represented on the Net.

Does such an orientation, rather than one of common conceptions of knowledge based on academic disciplines or on divisions within a company, have an effect on individuals, groups of individuals, or perhaps even a generation? Two people using the Web have different perceptions, which are due to differences in the structuring of what is experienced. How does this structuring take place? Any one person uses a small subset of the Net, and acquires over time a personal perception of its organization and contents. As most of us have acquired a knowledge of linear texts, some users of the Net have become comfortable with non-linear texts, with reactive or interactive texts. These can certainly be experienced in different ways. They are so designed. Some users may feel the Web is a random arrangement of items and acquire a particular sense of knowledge organization based on that perception. Like them, Gulliver ran across an encyclopedic frame with random words "in all moods, tenses, and declensions," which would ultimately produce "a complete body of all arts and sciences" and would work even faster "if the public would raise a fund for making and employing five hundred such frames (Swift, 1735; pt. III/V).

A "web" is an apt description of the network portion of the Internet. It is a useful conceptual model that reflects its hypertextuality; however, the analogy is somewhat inaccurate. The Web is not confined by a perimeter like a typical spider's web. Rather than being two dimensional, as many spiders' webs are, the Web is multi-dimensional. The Web is less static than that of a spider; it changes considerably over time. Web sources are not at all well-served by this model: at a given site, one may find practically nothing (perhaps a reference from an old to a new URL) or one may find the electronic resources of the Library of Congress.

One may wonder if the Net is randomly ordered or highly organized, but seemingly infinite. Indeed, conceptions of knowledge such as the random word frames observed by Gulliver or the infinite library of Borghes's "universe (which others call the Library)" (Borghes, 1956) can interpret knowledge as something with much, or without any, order. A similar
situation exists with some music from the middle of our century: that which is highly serialized (music in which durations, attacks, dynamics, and pitches are all rigorously organized) and that which is highly aleatoric (music in which many performance decisions are left to chance).

4. Evaluation

Perceptions of Net knowledge structures may play roles in searching habits or in deciding whether a Net search would be appropriate. How does one decide which way to search for information? In order to satisfy my curiosity about the ability of the Web to provide information on an obscure topic, I called upon one of the many search engines (Lycos in this case) to seek documents about Adalbert Blumenschein, an eighteenth-century Austrian librarian and priest who wrote a stupendous, but little-known, guide book to libraries of central Europe. Expecting to find nothing, I was pleasantly surprised to find two items: an abstract of an article in a scholarly journal at the journal’s Web site and an abstract of a paper delivered at a conference in Edinburgh at the Web site of the learned society sponsoring the conference. I don’t know of a library system that would be as successful. The results were also disturbing; it was only a happy coincidence that the journal and the scholarly society had Web sites and had this information available in a searchable format. Think of the thousands of publishers and other organizations that have no Web presence. It is not systematic in this regard.

The search engines (such as Lycos and many others) and classified arrangements (such as Yahoo and others) are the most important means of searching the Net. Both types of access were designed and made accessible many years after the Internet had existed. The Internet was not originally designed to be an information retrieval system. For this reason, it is not surprising that some established means of evaluation do not work well.

One measure of a system involves identifying and evaluating its sources of information. A survey of Web documents may tell us much about the system’s vastness. Evaluating the sources to which the Web leads is a monumental task perhaps best accomplished by focusing on specialized topics. Will either of these tell us how sources are actually used? No. Plays of Aeschelus are easy to find and it is always nice to know they are there; however, how many people actually read them in that form? The evaluation of individual Net sources is an overwhelming responsibility that has been aided by the presence of lists of recommended sites selected by librarians and other experts, as well as by Net source cataloging programs.

Evaluations of the Net, if they are applied at all, should be made not only to the search engines, but also to the classified systems, including some virtual reference libraries, the Yahoo index, and some self-described encyclopedias. I never felt the loathing of the Web’s Yahoo Index that Gulliver felt toward the Yahoos he met; being a former indexer, I succumbed immediately to the attraction of the Yahoo index, and would agree with Gulliver that I cannot deny that I am “a real Yahoo in every limb and feature” (Swift, 1735, pt. IV/I).

Are measurements of recall and precision feasible here? Recall, a ratio used to describe the ability of a system to retrieve a percent of relevant documents from all relevant documents in a system, is not useful here because the total number of relevant documents is not knowable. Documents are very fluid and changing. Being part of such an unsystematic system, the presence and quality many sources depend on the sustained interest of an individual or organization. For instance, in a reference class, I had students use the Web to locate reproductions of “The Scream” (or “The Cry”) by Edvard Munch. One particular site was especially fruitful: a person, obviously fascinated by the different forms of this image, had created a “Scream” site, which provided links to many digital versions of the work. This is the work of a volunteer who may be excited about Munch this moment, but who may not even
have a Net account tomorrow or in ten years. There are thousands of such passionate collectors. Although students found some images easily, it is impossible to know how many they missed.

Precision, which describes the ability of a system not to retrieve irrelevant documents, may be a more satisfying measure. Many users have already waded through a considerable amount of Web "trash" and have already carried out informal measurements of precision. Another complication, commonly encountered in recall/precision measurements, is that it is difficult to define relevance because it has always been very personal. Measures of recall and precision depend on relevance, which can be so unpredictable, so subjective, that it is difficult or impossible to verify. For one individual, a document or information source may be "close enough" to a subject or "good enough" for a particular use, even though it is not the best or even close to the best. If a fee is required, as they are for some of the highest quality Net resources, will a given user be less likely to use it? If it is necessary to register, will a user be less likely to use a very good source, even if no fee is involved? Will a frustrated user switch topics rather than carry out an exhaustive Internet search? Will a novice user take the time to learn how the different search engines work? Several Web search engines allow for Boolean searching, but in at least one case, the default operator is "OR," which of course has the potential of delivering results that are hardly precise. These problems are addressed in part by variations of the Principle of Least Effort: a solution will be judged by a user to be satisfactory if it is easily found, even if it is not the best available solution, and perhaps even if it is not a solution at all.

5. Conclusion: The Web as World Brain

The "hook" of a recent article in a popular information technology magazine reads, "Given that the Web itself is becoming the sum of the world’s knowledge, isn’t putting the *Encyclopaedia Britannica* online a spectacularly useless thing to do?" (Rossney, 1995). The author demonstrates that an online *Britannica* is indeed an especially useful resource, but treats as axiomatic the "given" that the Internet is assuming the role of world brain, a useful model that has been proposed from time to time.

Almost sixty years ago, H.G. Wells proposed the establishment of a world brain (Wells, 1938). He could not have predicted the existence of cyberspace or hypertext documents or e-journals; he would not have been able to imagine search engines of various capabilities. Yet his proposal seems in some ways to have been realized. Wells's proposal and the Internet go beyond the centuries-old idea of universal bibliography. Bibliographers of the past, such as Konrad Gesner, would surely be astounded by the breadth and depth of the Internet; yet some of them would be indignant over the haphazard way it has developed and about the low editorial standards of many of the sources it provides. What might take the "universal bibliographers" a time to realize is that the success and rapid evolution of the Internet are due in great part to the relatively disorganized way it has been formed and to a practice that has allowed any of millions of users themselves to contribute texts, graphics, other datafiles, collections of materials, and correspondence and to gain intellectual access to them in powerful ways. Increasing attention is being directed toward quality control, and equally importantly, to organized intellectual access.

In his discussions of a "world brain," Wells expresses a dream that is currently being realized. His words may be considered verbatim (1938, 20-21):
This World Encyclopaedia would be the mental background of every intelligent man in the world. It would be alive and growing and changing continually under revision, extension and replacement from the original thinkers in the world everywhere. Every university and research institution should be feeding it. Every fresh mind should be brought into contact with its standing editorial organization. . . . It would do just what our scattered and disoriented intellectual organizations of today fall short of doing. It would hold the world together mentally.

Wells saw the world brain as a tool in scholarly communication, presently one of the most heralded functions of the Internet (1938, 24):

To [the specialist] even more than to the common intelligent man World Encyclopaedia is going to be of value because it is going to afford him an intelligible statement of what is being done by workers parallel with himself. And further it will be giving him the general statement of his own subject that is being made to the world at large. He can watch that closely. . . . He will be able to criticize the presentation of his subject, to suggest amendments and re-statements.

Likewise, and in response to Wells's proposal, Smith agreed that "in a way characteristic of the spirit of the whole, the Living Encyclopaedia would turn the intellectual organization of whatever unit of society it had come to serve into an organic community activity rich and joyous with the spirit of mutuality" (1941, 60).

As concerned about the organization and structure of the world brain as we are about the topography of the Internet, Wells suggested that an Encyclopaedia Society be formed to survey the available material, which he considered to be in "a state of impotent diffusion" and to assemble authoritative subject bibliographies, perhaps even a master bibliography, and to form a general editorial board and departmental boards (1938, 27-28). At the same time, he proposed that the project represent the entire world (1938, 74):

So that while I believe that ultimately the knowledge systems of the world must be concentrated in this world brain, this permanent central Encyclopaedic organization with a local habitat and a world-wide range . . . nevertheless I suggest that to begin with, the evocation of this World Encyclopaedia may begin at divergent points and will be all the better for beginning at divergent points.

It is an understatement to say that "divergent points" characterize the Internet. Smith elaborated by suggesting that in order to avoid the destruction of intellectual freedom in such a vast and diverse plan, "the users themselves must in the ultimate be the controllers" in a situation in which contributors would heed the users to a greater degree than they would heed their readers in the traditional publishing world (1941, 61).

Very well aware of the developments of documentation projects in the 1930s, Wells emphasized the availability of a variety of materials for a wide readership (1938, 76-77):

It seems possible that in the near future, we shall have microscopic libraries of record, in which a photograph of every important book and document in the world will be stowed away and made easily available for the inspection of the student. . . . The time is close at hand when any student, in any part of the world, will be able to sit with his
projector in his own study at his or her convenience to examine any book, any document, in an exact replica.

Wells's proposal has of course not been completely realized. He believed that the new encyclopedism "should consist of selections, extracts, quotations, very carefully assembled with the approval of outstanding authorities in each subject, carefully collated and edited and critically presented" (1938, 20). While there is certainly a large and growing mass of texts and other sources, they are not always carefully assembled, uneven or no attention is given to textual quality, and subjects are not systematically represented. Further, Wells suggests that the system "would not be a miscellany, but a concentration, a clarification and a synthesis" (1938, 20). As it now stands, the Internet—arguably for the better—is a monumental miscellany, is regularly diluted, is by no means clear, and synthesizes nothing.

The Web may resemble human thought more completely than Wells could have envisioned. Like humans, the Web possesses long-term and short-term memory. On it can be found the unorganized rantings of hysterical individuals and the well-selected and sublimely written and organized texts of some of the best minds the world has seen. Also represented is the dark side of the human psyche. After carrying out a search or two—for purely clinical purposes, of course—I can conclude that there are a considerable number of links that guide one to the nether regions of human depravity. I was not completely ignorant about certain aspects of human behavior, but now I feel considerably less naive. The Net is indeed an educational tool.

Many people consider the Internet to be a living, growing, world brain-like organism with a life of its own. Upon reflection, it may well be living and growing, but its life, which is not self-sustaining, depends on many factors. We know its health will be dependent on the existence of conscientious hosts and users; we know less about the future of its political, economic, educational, social, and cultural life. The need for research related to electronic networking and networked resources is manifold. It is important to develop storage, retrieval, and communications technologies. It is as vital that we understand the organization of the networks themselves as it is to come to terms with the range of sources present on them. There exits a social imperative for us to manage the Internet and its successors within the contexts of its economic and political environments. Likewise, we are obliged to understand the related issues of accessibility to networks. Many of these needs fall into the categories of applied and theoretical research. Several journals, including Internet Research: Electronic Networking Applications and Policy, reflect the need for serious scholarly attention. Especially needed are perceptive examinations from historical and philosophical perspectives.

References
Abstract: An automatic indexing system using the tools and techniques of Artificial Intelligence is described. The paper presents the various components of the system like the parser, grammar formalism, lexicon, and the frame based knowledge representation for semantic representation. The semantic representation is based on the Ranganathan School of thought, especially that of Deep Structure of Subject Indexing Languages enunciated by Bhattacharyya. It is attempted to demonstrate the various steps in indexing by providing an illustration.

1. Introduction

The present work is an attempt to build an automatic indexing system using the tools and techniques of Artificial Intelligence. Information is presented to the indexer in natural language form and it warrants that any attempt in building an automatic indexing system should incorporate Natural Language Processing (NLP) tools.

Though, natural language syntax presents many problems, the real crux of NLP is semantics. Natural language is knowledge dependent. An NLP system should contain the knowledge of vocabulary, knowledge of syntax and also the knowledge of the specific domain.

2. Objectives

- To design and develop an automatic subject indexing system.
- To demonstrate that NLP approach has more potential in subject indexing as it attempts at semantics, unlike statistical techniques.

3. Hypotheses

The general theory of Subject Indexing Languages (SIL) forms the basis for this work:

1. The deep structure of SIL could be used as a semantic structure for expressive titles. In other words, the deep structure of subject indexing languages could serve as a Meaning Representation Language (MRL) for a natural language processing system.

2. Conceptual dependency grammars and many other grammar formalisms are heavily verb oriented. However, in the case of NLP systems for subject indexing, they should be noun phrase oriented.

3. Noun phrases that occur in a particular content description (expressive title) have a fixed role to play and have fixed relation to each other, irrespective of the noun phrase position in a particular sentence.

4. A frame based knowledge representation model is well suited for arranging the roles of Noun phrases and their relation to other Noun phrases.

4. Overview of Prometheus

Prometheus includes a bottom-up parser. The lexicon and syntax are built using.

unification based grammars where linguistic features are expressed in PATR notation. The meaning representation language is based on the Ranganathan (1967) school of thought especially the Deep Structure of Subject Indexing Languages (Bhattacharyya, 1980), which he introduced in his Postulate based Permuted Subject Indexing (POPSI).

Prometheus takes expressive titles of documents as input. The natural language parser checks the syntax of input statements and generates noun phrases. The basic approach in subject indexing emphasizes the importance of noun phrases (Vickery and Vickery, 1992) unlike the Conceptual Dependency grammars (Schank, 1984) which is verb phrase oriented. The noun phrases that are generated by the parser are passed on to the meaning representation system, which is basically a knowledge representation model based on the deep structure of subject indexing languages.

4. 1 The Parser

The purpose of a parser is to compute syntactic structures out of the given natural language statements. A grammar formalism as a declarative description of language does not specify how syntactic structure is to be computed. The critical decision is to choose an appropriate one out of many parsing algorithms. There are three basic issues involved in parsing --- whether the parser should be bottom-up or top-down; whether it should be depth-first or breadth-first and the issue of storing the intermediate results. Prometheus has used left-corner, bottom-up parser (Gazdar and Mellish, 1989).

4. 2 Syntax

One of the significant changes in the programming field is the shift from procedural definition of semantics of programming languages to declarative description of denotational semantics. The change holds true even with natural language processing.

The declarative paradigm that is adopted is Context-Free Phrase Structure Grammar (CF-PSG) and to be specific the Definite Clause Grammars and PATR. PATR in particular has become potential lingua franca for NLP work and many other grammar formalism can be expressed in it.

Modern theories of grammar use features. The extent and sophistication of feature based grammars have grown massively in 1980's. In feature theoretic syntax, atomic categories such as NP and V are replaced by sets of feature specifications. Feature specifications such as case (nominative, accusative), number (singular, plural), person (first second, third), etc., can be added to the grammar more elegantly.

In case of expressive titles of documents, it should be noted that they may not be complete English sentences. In other words, a set of syntactic rules are developed in order to accept titles of documents. One such rule expressed in terms of linguistic features in PATR notation is:

Rule NP --> [NP1, CONJ, NP2]:-
    NP: cat == np,
    NP1: cat == np,
    NP2: cat == np,
    CONJ: cat == conj,
    NP: np1 == NP1:np,
    NP: np2 == NP2:np.
4.3 Lexicon

Feature structures are now widely used to represent morphologic, syntactic and semantic information. Feature structures along with lexicon have permitted computational linguists to adopt very simple and compact rule system at the cost of passing almost all syntactic facts about language into the lexicon.

Word determination:
\[ W:cat = n, \quad W:n = \text{determination}. \]

Word diagnosis:
\[ W:cat = n, \quad W:n = \text{diagnosis}. \]

4.4 Semantic Representation

In order to represent the meaning of given expressive titles, Prometheus uses a frame-based knowledge representation model based on the deep structure of indexing languages.

Each noun phrase that is passed on to the semantic system is analyzed in order to find out to which category it belongs. In other words, it is essential to recognize whether an isolate belongs to either entity, or property or action or speciator etc. In addition, the system performs standardization, modulation, assignment of indicators, generation of subject index entries following the syntax of the deep structure of subject indexing languages.

5. Formulation of Subject Index Entries

Prometheus adopts the following steps in order to arrive at the deep structure of subject indexing languages:

Step 1: Identification of syntactic categories of terms in an expressive title. Exclusion of articles, conjunctions, etc.

Step 2: Generation of syntactic structure.

Step 3: Construction of noun phrases which include uniterms, compound and composite terms.

Step 4: Standardization, identification of elementary categories for each noun phrase and assignment of indicators for each elementary category.

Step 5: Modulation

Step 6: Generation of subject entries using the syntax of deep structure subject indexing languages.

Illustrated Example:

The best way to understand how Prometheus generates deep structure of a given expressive title is to demonstrate it with the following title:


Step 1: Identification of syntactic categories.

The parser of Prometheus is a bottom-up parser i.e. it is data driven. The parser reads every word in the expressive title and looks for each word in the lexicon. In the lexicon, each
word is mentioned in PATR notation, in the following way:

Word determination :
W : cat = = = n,
W : sem = = = determination.

Word of :
W : cat = = = prep.

Word depth :
W : cat = = = adj,
W : sem = = = depth.

Word dose :
W : cat = = = n,
W : sem = = = dose.

In the lexicon, each word has an entry, describing the syntactic features and also, the word it should pass for the final syntactic structure. This is mentioned under the feature sem.

Step 2: Generation of syntactic structure

The parser of Prometheus attempts to apply the grammar rules specified in PATR notation. Thus, the adjectival phrases depth dose, satisfies the following syntactical rules

Rule NP —> [ADJ, NP1] :
NP : cat = = = np,
ADJ : cat = = = adj,
NP1 : cat = = = np,
NP : np : adj = = = ADJ : adj,
NP : np : n = = = NP1 : np.

Rule NP —> [NP] :
NP : cat = = = np,
N : cat = = = n,
NP : np = = = N : n.

These two rules state that a noun-phrase can contain an adjective followed by a noun. The rule for adjectival phrases is recursive, so that it can accept any number of adjectives to be proceeded before a noun as is the case in X-ray rotation therapy. Following is the syntactic structure produced by Prometheus for the given expressive title.


In this syntactic structure, it can be seen that Prometheus has accepted the given expressive title as a grammatical construct and indicated the syntactic categories of constituent words.

Step 3: Construction of Noun Phrases

This step is fairly straightforward once Prometheus generates the syntactic structure of the title. The system identifies each noun phrase and adds it to internal database for further semantic processing. It should be noted that the system generates compound words with a `,` as connector. Thus the output of step 3 is:
Step 4: Identification of Elementary Categories

The noun phrases, thus generated by syntactic processing are then passed on to the frame based knowledge representation system based on the deep structure of subject indexing languages. Of the noun phrases generated, the system first picks up the word determination, and then searches for this word among the facts of the knowledge base until it encounters the fact:

\[ \text{value(determination, use, measurement).} \]

This fact indicates that the word determination is a non-standard term and should be replaced by measurement. The next step is to identify the category of the word measurement, for which the system encounters the following fact:

\[ \text{value(measurement, category, action-on-action).} \]

which says that the word measurement belongs to the category action-on-action. Then the next step is to identify the indicator for action-on-action, for which the fact

\[ \text{value(action-on-action, indicator, 8.2.9.9).} \]

informs the system that the indicator is 8.2.9.9. Thus the input word determination becomes 8.2.9.9 measurement. This fact will be added to a B-tree so that the system produces subject heading, and places measurement depending on the ordinal value of 8.2.9.9.

The same steps are followed in producing the relevant substring for the phrase depth-dose, except that a connector [of] is added before the depth-dose, as the knowledge base contains the following fact:

\[ \text{value(depth-dose, connector, [of]).} \]

The phrase roentgen rotation therapy, is factored into roentgen and rotation therapy. The word roentgen is replaced by X-ray using the fact

\[ \text{value(roentgen, use, X-ray).} \]

and rotation therapy is replaced by rotation technique, using the fact

\[ \text{value(rotation therapy, use, rotation technique).} \]

Thus the system generates,

8.2.9.5 [using] X-ray and
8.2.9.5 [using] rotation technique

To summarize, the following strings are added to the B-tree index

8.2.9.9, measurement
8.2.9.9.5, [of], depth-dose
8.2.9.5, [using], X-ray
8.2.9.5, [using], rotation technique
8.2.9, treatment
8.2.9.9.6, [using], ionization pocket chamber

Step 5: Modulation

The system attempts to find the broader terms, if any, for each and every phrase. The broader term, narrower term relation is expressed using kind-of or part-of relation. Thus, the term X-ray picks up a broader term radiation using the fact

\[ \text{value(X-ray, kind-of, radiation).} \]

from the knowledge base. Similar logic applies to the term
ionization"pocket"chamber
which encounters the following fact in the knowledge base
value/ionization"pocket"chamber, kind-of, ionization"chamber).
Each and every broader term along with its indicators and connectors are then added to the
B-tree.

**Step 6: Generation of Subject Entries**
The purpose of storing each term along with its indicators is to get the sorting order. The
final step involves simply printing the strings that are stored in the B-tree one after another.
Thus the expressive title *determination of depth dose in roentgen rotation therapy using
ionization pocket chamber* becomes

'1 medicine 8 human"body 8.2 disease 8.2.9 treatment 8.2.9.5 [using] radiation 8.2.9.5
[using] rotation"technique 8.2.9.9 measurement 8.2.9.9.5 [of] depth"dose 8.2.9.9.6 [using]
ionization"chamber 8.2.9.9.6 [using] ionization"pocket"chamber.'

6. Conclusion
Prometheus is not a fully automatic indexing system. It is an attempt towards building
such systems. Nevertheless, it substitutes human intelligence in all the required steps in subject
indexing except one -- preparation of the expressive title after carefully examining the
document title, abstract and text. Fully automatic indexing may take more time until the
natural language processing research offers definite results. However, Prometheus attempts
to generate subject strings from expressive titles presented in natural language form.

One of the significant aspects of Prometheus is its semantic representations. It attempts
to demonstrate that the deep structure of subject indexing languages could be used as a semantic
structure in subject indexing.

Prometheus is not without its limitations. The parser of Prometheus performs shallow
syntactic analysis. Though it uses feature based grammar, it does not include semantic features
of terms. It does only contain the syntactic features of terms. Most of the semantic task is
passed on to the frame based representation of knowledge. The question is, whether it is
advantageous to perform semantic processing along with syntactic processing or semantic
processing should follow syntactic processing, is highly controversial. However, an attempt
can be made in future to compare the results of both the approaches.

References
Dharwad: Karnatak University.
to computational linguistics. New York: Addison-Wesley.
Prasad, A.R.D. (1994). Application of computer based natural language processing tools and
University.
House.
Massachusetts: Addison-Wesley.
Toshiyuki Matsuo  
Advanced Technology Group, The Japan Research Institute Limited  

Toyoaki Nishida  
Graduate School of Information Science, Nara Institute of Science and Technology  

Intelligent Support for Construction and Exploration of Advanced Technological Information Space  

Abstract: This paper presents a practical method of extracting, structuring, summarizing and integrating technological information in metallurgy. We have implemented a system called METIS (METallurgy papers Intelligent Surveyor) which interprets technical papers in metallurgy written in a mark-up language and produces a variety of summaries and surveys such as structured technical summary, visualization of similarities, differences of relevant papers, and cause-effect relations. The heart of the method is a packet of domain specific knowledge called KP (Knowledge Piece) in which procedures for extracting and structuring technological information from technical papers are embedded. METIS extracts technological information by naive natural language process specified by keywords or key phrases embedded in KPs, and produces summarized information specified by KPs. We have undertaken qualitative and quantitative evaluation of METIS against 186 technical papers. METIS has reliable results especially with typical papers. In order to capture more precise information, we may have to double the number of KPs.

I. Introduction

According to rapid progress of multimedia and network technology, technical information is spreading more widely around researchers and engineers. Growing demands exists for systems which can classify, summarize, and produce technological information effectively from natural-language texts. Much more effort is needed to understand those papers which can be retrieved from conventional data base system in science, because those papers are left unclassified or unsummarized. Retrieving by keyword matching with full text or with an abstract often generates too many papers which make impossible to extract technological information.

In cooperative work with an expert in super alloy, which is a specific domain in metallurgy, at National Research Institute for Metals in Japan, we have been carrying out research on automatic extraction of technological information in metallurgy. We have three policies in our research as follows: (1) content processing, (2) reliability and robustness, and (3) application of present artificial intelligence technology. We have studied information structure of ten technical papers in metallurgy. The analysis has led us to a conclusion that there is a generalized schema in papers, and that we get a bright prospect to extract and structure reliable information with a schema automatically.

This paper presents a practical method of extracting, structuring, summarizing and integrating technological information in metallurgy. We have implemented a system called METIS(METallurgy papers Intelligent Surveyor) which interprets technical papers in metallurgy written in a mark-up language and produces a variety of summaries and surveys such as structured technical summary, visualization of similarities, differences of relevant papers, and cause-effect relations.

The heart of the method is a packet of domain specific knowledge called KP (Knowledge Piece) in which procedures for extracting and structuring technological information from technical papers are embedded. Based on the analysis of contents in metallurgy papers, we have concluded that about a hundred KPs would be sufficient for capturing important information in technical papers. METIS extracts technological information by naive natural language process specified by keywords or key phrases embedded in KPs, and produces summarized information specified by KPs. We have undertaken qualitative and quantitative evaluation of METIS against 186 technical papers. METIS has reliable results especially with typical papers. In order to capture more precise information, we may have to double the number of KPs. In section 2, we present our approach extracting, structuring, summarizing and integrating technological information from technical papers. In section 3, we describe a system called METIS which produces a variety of summaries and surveys. In section 4, we present qualitative and quantitative evaluation of METIS against 186 technical papers. Finally, we present a comparison with related work.

2. Our Approach

We have analyzed information structure of ten papers in metallurgy. The results of the analysis suggests that simple natural language process will do for capturing critical technical information. We have studied information structure of papers in metallurgy. The analysis of papers' structure has led us to the following conclusions. (1) Many papers in super alloy consists of a title, author, experimental procedure, experimental results and conclusions. (2) Researchers are interested in experimental procedure and experimental result. A more detail analysis leads us to a conclusion that papers should be divided into detail sections such as material for experiments, chemical composition, preparation before experiments, methods of experiments, experimental results and discussions. (3) Figures and tables, which show microstructure analysis by a microscope and summarize experimental results, produce the most important technological information in a paper. But it does not seem that there is a practical technology that extracts information from figures and tables automatically.

The results of the analysis make it clear that we can implement a system with three policies as follows. (1) We represent a paper's knowledge piece in metallurgy as a schema which has a standardized structure. (2) We represent papers' contents as a combination of a hundred schema. (3) We represent technological information in metallurgy papers with schema and naive natural language process, which is mainly pattern matching with technical terms and domain specific expressions.

We show a knowledge representation method which is a packet of domain specific knowledge called KP (Knowledge Piece) in which procedures for extracting and structuring technological information from technical papers in metallurgy are embedded. KPs have a parent-child relationship between relevant KPs. KP has four facilities as follows:

1. To select sentences
   \[ S := \text{a set of sentences}; \quad C := \text{conditions to extract sentences}; \quad E := \text{conditions to exclude sentences are given.} \]
   \[ S' := \text{a set of selected sentences is generated.} \]
   \[ S \times C \times E \rightarrow S' \subseteq S \]
2. To extract features
   \( S := \text{a set of sentences}; \ X_f := \text{conditions to extract features}; \ Cf := \text{a set of extracted concepts}; \ Af := \text{a set of co-occurrent words with} \ Cf \ \text{are given.} \)
   \( V_f := \text{a summary of features is generated.} \)
   \( S \times X_f \rightarrow V_f, \ \text{provided that} \ X_f = \{ <c_f, a_f> \mid c_f \in Cf, \ a_f \in Af \} \)

3. To structure information
   A \( KP \) is structured recursively by merging child of the \( KP \).
   \( K := \text{a} \ KP \); \( F := \text{a feature.} \)
   \( 2^{PK} \rightarrow K \)

4. To intersect relevant \( KPs \)
   A \( KP \) is structured by intersecting relevant \( KPs \).
   \( S := \text{a set of sentences} \)
   \( K_1 := \{ K_{1_1}, K_{1_2}, \ldots, K_{1_{i_1}}, \ldots, K_{1_m} \}; \ K_2 := \{ K_{2_1}, K_{2_2}, \ldots, K_{2_{j_1}}, \ldots, K_{2_n} \} \ \text{are given, provided that both} \ K_1 \ \text{and} \ K_2 \ \text{are a set of relevant} \ KPs. \)
   \( V := \{ \text{a set of sentences} \} \ \text{is generated.} \)
   \( K_1 \odot K_2 \rightarrow V \)
   \( V := \{ V_{ij} \mid 1 \leq i \leq m, \ 1 \leq j \leq n \}, \ \text{provided that} \ V_{ij} \ \text{is a set of sentences that satisfy extracting conditions of both} \ K_{1_i} \ \text{and} \ K_{2_j} \)

A \( KP \) is defined by items as follows. A name of \( KP \), conditions to extract information, conditions to exclude information, a set of child \( KP \) name, a set of relevant \( KP \) name, method of information process, a set of co-occurrent words, storage area for extracted data, storage area for processed data, storage area for merged data and storage area for intersected data.

3. Overview of \textit{METIS} 

We have implemented a system called \textit{METIS} (METallurgy papers Intelligent Surveyor) which interprets technical papers in metallurgy written in a mark-up language and produces a variety of summaries and surveys such as structured technical summary, visualization of similarities and differences of relevant papers, and cause-effect relations. Table 1 summarizes functions of \textit{METIS}.

The architecture of \textit{METIS} is applicable to other domains, because definitions of \( KP \) and an algorithm for natural language process based on \( KPs \) are independent.

3.1 Summary of a Paper in Metallurgy

Since we define \( KPs \) for summaries, \textit{METIS} produces two types of summaries, i.e., itemized expression and table expression, through \textit{information structuring} process and \textit{intersecting relevant} \( KPs \) process.

First, \textit{METIS} applies \( KPs \) to papers, and produces a \( KP \) instance which represents structured technical summaries as itemized expressions.

\textit{Structured technical summary.} \textit{METIS} produces a \( KP \) instance which represents structured technical summary through information structuring process of \( KPs \). There are five \( KPs \), i.e., for metal materials, for experimental results, for discussion, for tables and for figures (see Figure 1). \textit{METIS} produces a \( KP \) instance which represents experimental result through information structuring process of \( KP \) for creep properties, fatigue properties, corrosion properties and other properties. \textit{METIS} produces a \( KP \) instance which represents metal
material through extracting features process.

**Structured summary of experimental methods.** METIS produces a KP instance which represents structured summary of experimental methods through information structuring process of KP for morphology, heat treatments and so on. METIS produces a KP instance which represents morphology through information structuring process of KP for orientation and preparation method. METIS produces a KP instance which represents orientation through extracting features process.

Second, METIS produces **information about cause-effect relations** in a table form. Cause-effect relations are the most important information in a paper. METIS selects sentences including cause-effect relations. Cause-effect relations are classified into four patterns, (1.) experimental process → metal microstructure, (2) metal microstructure → metal properties, (3) metal microstructure → metal microstructure and (4) experimental process → metal properties. METIS produces three expressions of cause-effect relations.

1. Causes to give an influence on metal property
   METIS expresses cause-effect relations from a viewpoint that a experimental process or microstructure have some influence on some metal properties. METIS summarizes cause-effect relations in a table form through selecting sentences indicating cause-effect relations and intersecting relevant KPs.

2. Contents of cause-effect relations (see Figure 2)
   METIS expresses cause-effect relations by extracting features about relations, experimental methods, structure and metal properties per a sentence, and summarizes them in a table form.

3. Directional graph of cause-effect relations (see Figure 3)
   METIS expresses cause-effect relations by directional graph that has a cause as starting point, an effect as ending point, and a relation as a link. METIS locates extracted features on three dimensional space. The features are categorized into three, i.e., metal property, experimental process and microstructure. The location of the features are predetermined by METIS. METIS draws a line as a relation between causes and effects.

### Table 1: Functions of METIS

<table>
<thead>
<tr>
<th>function</th>
<th>(1) automatic</th>
<th>(2) semi-automatic</th>
<th>(3) automatic</th>
<th>(4) automatic</th>
<th>(5) automatic</th>
<th>(6) automatic</th>
<th>(7) automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>summary</td>
<td>(1) unstructured technical summary</td>
<td>(2) structured summary of experimental methods</td>
<td>(3) causes to effect metal property</td>
<td>(4) directional graph of cause-effect relations</td>
<td>(5) contents of cause-effect relations</td>
<td>(6) sentences including cause-effect relations</td>
<td>(7) information of figures and tables</td>
</tr>
<tr>
<td>survey</td>
<td>(1) plots of similarities between papers on 2-dimensional space</td>
<td>(2) plots of similarities between papers on 3-dimensional space</td>
<td>(3) similarities of relevant papers</td>
<td>(4) distribution of papers from a viewpoint of cause-effect relations</td>
<td>(5) merged cause-effect relations</td>
<td>(1) automatic</td>
<td>(2) automatic</td>
</tr>
</tbody>
</table>
Improvement of Creep Strength in a Nickel-base Single-crystal Superalloy by Heat Treatment

1. Introduction
2. Methodology
3. Results
4. Discussion
5. Conclusion
6. References

Fig. 1: Structured technical summary
Fig. 2: Contents of cause-effect relations

This figure indicates that some relations exist between causes and effects per sentence. When you click a numeral that indicates ID of a sentence including cause-effect relations, METIS shows a relevant sentence.

3.2 Surveys of Papers in Metallurgy

Plots of similarities between papers into three dimensional space. METIS calculates similarities between papers in metallurgy from structured summaries (see Figure 1). A KP for structured summary consists of five KPs, i.e., metal material, experimental results, discussion, table information and figure information. METIS calculates similarities between relevant KPs, and sums up them into a similarity between papers by the following equations.

\[
\text{similarity between paper } A \text{ and paper } B = \sum_i \text{ (weight) } \times \text{ (similarity between } KP_i) \\
\text{similarity between } KP_i = (1 - \frac{\text{NUM}(A \cap B)}{\text{NUM}(A)}) \times \frac{\text{NUM}(A \cap B)}{\text{NUM}(B)}^2
\]

\[
\text{NUM}(X) = \text{(the number of items extracted in } KP_i \text{ of paper } X)
\]

Matrix \( S = (s_{ij}) \) is defined as a similarity matrix between \( n \) papers, where an \( s_{ij} \) is a similarity between \( i \)-th paper and \( j \)-th paper. \( s_{ij} \) satisfies the conditions \( \{ s_{ij} \geq 0, s_{ij} = s_{ji}, s_{ii} = 0 \} \). An \( s_{ij} = 0 \) means that \( i \)-th paper and \( j \)-th paper are the same in summarized information level. METIS applies MDA (Multi Dimensional Analysis) method to similarity matrix \( S \), and get a
matrix $A = (a_p)$. This matrix means a location of each paper in a $n$-dimensional space. MDA calculates this matrix so that the distance between papers in this space of matrix $A = (a_p)$ indicates the similarity between papers. METIS calculates matrix $A$ and plots them on two or three dimensional space (see Figure 4). We usually select the first and the second axes, or the first, the second and the third axes in order to display similarities, because it is impossible for us to grasp a spatial relation in more than four dimensional space. The reliability of similarity by k-dimension plots is measured by a ratio between sum of the first k eigenvalues and sum of all eigenvalues. When we have to know the similarities of the papers located relatively nearly on three dimensional space, METIS extracts common features per KPs and produces similarities of relevant papers.

Distribution of papers from a viewpoint of cause-effect relations. METIS sums up a matrix which shows causes to effect metal property, and produces a distribution of cause-effect relations (see Figure 5).
Fig. 4: Plots similarities between papers on three dimensional space

This figure shows a similarity between papers by a length between spheres on three-dimensional space. Each ball in the virtual 3-D space indicates a paper.

4. Evaluation

METIS summarizes 87 papers published at 1992 national conference in super alloy, calculates similarities and plots them on three dimensional space. The results indicate usefulness of our method with KPs. We have a bright prospect of automatic papers' classification into some clusters.

Next, we have undertaken quantitative evaluation of METIS against 186 papers so far (see Table 2). The expert evaluates the technical summaries produced by METIS to five ranks, i.e., from A, B, C, D and E. Although METIS produces reliable results with typical papers, the expert in metallurgy points out that those summaries produced by METIS lack important keywords which must be essentially extracted. There are two features of papers in metallurgy which do not give rise to a satisfactory evaluation, (1) to discuss relation between production process and properties, (2) to discuss very general properties (e.g. tensile property). Lacks of important keywords are caused by lack of appropriate KPs. Although it seems that an addition of more relevant KPs would improve evaluation of METIS, we have to apply other artificial intelligence technology (e.g., mechanical learning technology) in order to (semi-)automatically construct appropriate KPs.
Table 2: Evaluated results

<table>
<thead>
<tr>
<th>A</th>
<th>A-</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>C+</td>
</tr>
</tbody>
</table>

Fig. 5: Distribution of papers from a viewpoint of cause-effect relations

This figure shows how many papers discuss cause-effect relations. This figure shows that there are many papers discussing cause-effect relations between cause (e.g., γ phase, single crystal, heat treatment) and effect (e.g., creep strength).

5. Comparison with Related Work

This work is considered to be an effort of acquiring knowledge from texts and representing knowledge by naive natural language process.

Previous and frontier work based on script theory in this direction involves: SAM (Script Applier Mechanism; Cullingford, 1978) and FRUMP (Fast Reading Understanding and Memory Program; Dejong, 1979). Claire Cardie (Cardie, 1994) shows current work based on case-based approach. SAM and FRUMP could acquire knowledge from texts about short stories and newspapers, and there were not detail evaluations of them from users. The system of Cardie (1994) is limited to summarize domain-specific sentences. Niki and Tanaka (1995) shows work on automatic classification of papers using neural networks. This system produces functions of dynamic generation of schema, browser of contents with a variety of
abstraction and vague information retrieval.

The evaluation of METIS demonstrates reliability and robustness, differently from SAM and FRUMP. METIS integrates technological information, that is not achieved by the system of Cardie (1994). While the system of Niki and Tanaka (1995) does not use domain-specific knowledge to extract information, METIS uses domain-specific knowledge which enables it to investigate details of contents. Therefore METIS can perform structural extraction of information which is embedded in structure of contents.

6. Summaries

We have applied artificial intelligent technique to extract, structure, summarize and integrate information from papers in metallurgy. The heart of our method is a packet of domain specific knowledge called KP. We have undertaken qualitative and quantitative evaluation of METIS. The evaluation demonstrates reliability and robustness of our method. An interesting open problem is to acquire and represent deep knowledge, so that METIS can produce the most important information from summarized results and to research on an intelligent integration of extracted information.

References


Abstract: Natural language processing systems use various modules in order to identify terms or concept names and the logico-semantic relations they entertain. The approaches involved in corpus analysis are either based on morpho-syntactic analysis, statistical analysis, semantic analysis, recent connexionist models or any combination of two or more of these approaches. This paper will examine the capacity of natural language processing systems to create databases from extensive textual data. We are endeavouring to evaluate the contribution of these systems, their advantages and their shortcomings.

1. Introduction

Setting up the conceptual design of a database (object-oriented or relational) is not a simple task for the designer. He or she has to examine a new field of knowledge i.e. analyzing an extensive textual data (e.g. analyzing users' needs and profiles, interviews with field specialists, term glossaries, technical documentation, etc.). The designer may not be acquainted with the representation of the field, its structures and the articulations between its objects.

To make the designer's task easier, natural language processing systems can be of help particularly those dedicated to the identification of terms or concepts names related to a specific field of knowledge (construction of a reference terminology) and the logico-semantic relations they entertain. These systems can be applied to the modelling and designing of the following types of systems:

- The modelling of an object-oriented database design (static aspects : i.e. describing the structure),
- Knowledge-based systems : modelling the hierarchies between classes and the relations between the objects concerned by a set of rules,
- Modelling the conceptual design of a relational database (domains, relations, coherence maintenance);
- Thesaurus construction (documentary databases, ...),
- Terminological database construction, etc.

This paper will examine the capacity of natural language processing (NLP) systems to help the designers in creating databases from extensive textual data. The scope of our study will be limited to terminological databases.

NLP systems we are evaluating use various modules in order to identify terms or concept names and the logico-semantic relations they entertain. Most of these systems need, in addition, a general language dictionary, a glossary of technical terms covering the relevant field, etc. The identification of terms is in fact an extraction of noun phrases corresponding to
the concepts representing the field of knowledge. In their current state, these systems are mostly semi-automatic processing tools.

We are presenting in part two a state of the art of terminology building tools and a brief account of the model we are proposing. Then we review the systems proposed for the evaluation from a linguistic point of view i.e. the underlying linguistic models. The comparison is continued with the typology of systems from the output point of view. At this point we discuss the possibility of integrating different models in order to have better systems output i.e. ideally terms and semantic relations they entertain with each other. We compare the capacities of the tested systems and SEEK and we show in what respect SEEK can be a complementary module to syntactically and statistically-based systems. In section three we propose a paradigm of systems evaluation based upon the following two criteria: digital evaluation and qualitative evaluation.

2. Scientific Background

2.1 State of the Art

We can distinguish several areas in which terminology develops. The most traditional field of application is that of translation. The second area is that of monolingual activities which are generally prescriptive. (Wijnands, 1993). The third field of terminology is documentation and information science (respectively thesaurus descriptors whose role is to describe documents, knowledge acquisition tools i.e. terms representing a conceptual field etc.). This field of application reveals the importance of terminology as a means for information retrieval i.e. domain knowledge can be acquired by repeated information retrieval (Kukulska, 1993). Ideally a terminological database or terms bank should be a multi-functional tool answering the needs of a wide range of users: information systems designers, technical writers, knowledge-based systems designers, specialized translators technical and scientific dictionaries authors. The question of potential users of these tools is discussed in (Sager, 1990, 169).

If we focus on the tools devoted to terminology building we can distinguish two types: conventional Term Banks and Terminological Database Management Systems. The questions a specialist can address to this type of data model (conventional term bank) can be on the gender, the spelling, the equivalent, definition, subject label, synonym, contextual usage of a term, etc. They are in principle the same ones that a specialist would address to a general language dictionary. In addition to the type of the questions we listed above, some retrieval capabilities are now available in some term banks: compilation of all terms used in a domain, definitions and conceptual links, subordinate/partitive terms, the name of the terminologist who entered a group of terms over a period of time (Sager, 1990, 168).

In addition to providing terms to describe concepts in specialized domains terminology is a means of representing all the knowledge relating to a particular term (linguistic and conceptual information). Moreover, computing techniques such as concordancing can be used to explicit information from texts such as word frequencies, collocations, etc. thereby enabling terminologists to assess the extent to which different levels of language are needed. Collocations in terminology leads to phrases and phraseology: it is only recently that most advanced banks such as TEAM and DANTERM (see Thomas, 1993) have begun to include these elements.

The new trends are represented by term banks such as COGNITERM (University of Ottawa) a hybrid between a conventional and a knowledge-base system (in the sense known in
artificial intelligence. COGNITERM is a concept-oriented and multidimensional handling tool (see Meyer, 1993; Meyer and Bowker, 1992).

2.2. Terminological Database Design Proposed

Assuming the theoretical bases of terminology (see Dahlberg, 1978, 1981; Jouis and Mustafa, 1995) our approach to terminology is a concept-oriented one and it will be taking into consideration the state-of-the-art in terminological databases. The database structure we are proposing consists of records. In each record we describe a term related to a certain field of knowledge. In each description hypertext links point towards textual areas in the text as well as towards labelled relations. The database layout is more precisely shown in fig. 2. It will consist of basic entries: term, definition(s), contextual definition(s), conceptual relations underlying terms. The latter will be presented in a semantic-network fashion (see below), hypertext links pointing to contextual definitions, a set of pointers oriented towards other records of the database pinpointing to semantic relations (see fig. 2).

The typology of the semantic relations we identified in our model are based on the Applicative and Cognitive Grammar (AGC3). ACG articulates several levels of representations especially a cognitive level in which the meanings of linguistic units may be analyzed under the form of layouts (semantic-cognitive representations) in order to constitute the knowledge representations associated to a given text. ACG proposes a set of semantic concepts which defines an organized system of meanings. We distinguish the semantic types of linguistic units, fundamental static relations and dynamic relations (movement, change of state, conservation of a movement, iteration, intensity, variation, constraints, causes ...).

2.3 Evaluated Systems Typology within the Context of Linguistic Perspectives

We can distinguish three categories:

(1) The first category consists of syntactically-based systems, to which statistical modules are sometimes added. These systems use a set of grammatical words and punctuation symbols in order to isolate terms in the textual data. The following systems belong to this category:
(a) **ACABIT** is a prototype geared to automatic terminological bank construction. Its task is to help an expert in a certain field of knowledge by proposing a set of potential candidate terms in a pre-defined morpho-syntactic format. The statistical module role is to rank the produced terms according to their relevance to the field (see Daille 1995).

(b) **KES** (Knowledge Extracting and Structuring) is a set of tools developed within the framework of a European research program (GRAAL) a consortium including three shareholders: GSI-ERLI (a French language engineering firm), Aérospatiale (French Aircraft Company) and EDF-DER (French Electric Company). KES's job is to extract and structure knowledge of a given corpus of text thereby providing a structured data from a non-structured one (see Ogononowski 1994).

(c) **LEXTER** is a system used in extracting potential candidate terms related to a field of knowledge. The system is based on an endogenous learning procedure enabling it to automatically acquire syntactic information of sub-categorization. Each term is grammatically analyzed into head and modifier in order to build a lexico-syntactic network. A further step is knowledge modelling which can be undertaken by a terminologist or a knowledge engineer. The latter can exploit this network in order to reconstruct the underlying conceptual network (see Bourigault 1992).

(2) The second category consists of pure statistical and/or connexionist models:

(a) **ANA** identifies single terms and characteristic expressions of a field of knowledge and their organization in a non-labelled network. The system is based on the following postulate "Frequent co-occurring events are significant" (Enguehard 1993).

(b) **CONTERM** is based on a connexionist model which applies a neuronal network to textual chunks in order to build classes of networks of terms linked to a major one. By so doing CONTERM tries to identify distinct semantic networks for each identified term.

(3) The third category is the one including semantically based systems such as **SEEK**, **IOTA** and **NOEMIC**. These systems focus particularly on semantic relation.

(a) **SEEK**, for instance, uses a Contextual Exploration Approach, a method based on the linguistic knowledge of a reader when he explores a text in a given language. The system is capable of identifying concepts and concept relations (see Jouis 1994; 1995). SEEK can be considered as the most atypical system for the following reasons: (i) Most natural language systems use parsers, general language dictionaries, etc.); (ii) Most of the systems we are testing use intensive resources compared to a contextual exploration-based system like SEEK which needs only a limited number of linguistic markers to do the same job.

If we focus on these systems in order to see how they exploit the linguistic resources we can say that syntactically and statistically based systems use a set of grammatical words and punctuation symbols in order to isolate terms in the textual data. Contextual exploration approach, unlike these systems, considers the contents of this set (i.e. grammatical words and punctuation) as meaningful elements hence their correspondence to the linguistic knowledge independently of a specific subject-field. Their role is to convey meaning and allow the designer of the database to locate concepts and the relations they hold between them in an interactive fashion. The system does not need either a general language dictionary, a parser or a glossary of terms. This is why we consider SEEK as a fundamentally atypical system. This system is capable of identifying concepts and concept relations.

We assume that contextual exploration is a strategy that an analyst uses while examining textual data: he or she builds up a data model by observing certain linguistic indicators or markers. A semantic analysis using contextual exploration is presented as a knowledge-based system. The model based on linguistic knowledge is chosen for its capabilities in analyzing concepts and/or terms.
IOTA is a system based on a Sowa conceptual graph model. Its purpose is identifying noun phrases in order to index documents from a textual data. It also dynamically builds an automatic thesaurus used for information retrieval (see Bruandet 1980).

As for NOEMIC, it is based on a componential analysis model which accounts for a variety of typed semantic relations (thesaurus relations). NOEMIC does not need a 'vertical' linguistic analysis (i.e. morphological, syntactic). For further information (see Penel 1994).

2.4. Systems Typology from the Output Point of View

The evaluation of these systems raises a series of questions that we should answer: The first question we will try to answer is: to what extent do natural language systems manage to adequately and efficiently identify automatically or semi-automatically terms/concepts and concept relations that can be used to build information retrieval tools? As input the systems will get textual data from specific fields of knowledge (manuals, specialized articles, newspaper articles, dispatches, etc.). The output produced by these systems can be divides into two categories: (i) lists of candidate terms, eventually ranked by frequency of occurrence, (ii) semantic relations. We can already observe that morpho-syntactic methods combined with statistical modules produce lists of terms without significant semantic relations while semantically-based systems tend to focus on semantic relations resulting in rather limited terms productivity. The idea is to tend towards the integration of systems (a hybridization of models or cross-breeding) in order to enhance the capability of these systems. We are putting forward the idea of integrating two or more systems within the framework of this research: integrating SEEK, for instance, to a syntactically and statistically-based system.

2.5. Corpus

The second question deals with the criteria of choosing adequate corpuses. The elaboration of terminology includes the analysis of documents of different types and varying sizes. (textbooks, technical manuals and specialized dictionaries, transcripts of interviews with field experts, etc.). The "ideal" corpuses to efficiently identify terms and relations which can be considered as good descriptors providing efficient access to specialized information: automatic thesaurus construction, broadening of existing thesaurus, etc. should be representative of the field of knowledge considered. Three types of texts are needed (see Ahmed, 1993): instructional (textbooks, technical manuals, encyclopedic texts); informative (learned papers, advanced treatises, interview transcripts of experts, patent documents); and imaginative (popular science material, public information material i.e. advertisement about the goods and services of the subject domain). This variety is needed because the terms in each type of corpus serve different purposes. This typology is used by a team at the University of Surrey to build term banks in about 10 different subject fields (see Ahmed 1993).

In addition to the mentioned elements we are trying to define a set of criteria for evaluating these various systems according to the following areas of performance: (i) information retrieval (recall and precision); (ii) information display (output: lists of terms, semantic networks) or any combination of two or more of these paradigms, etc.).

3. Evaluation Problem

Automatic term construction raises a series of questions: what type of terminology to create? Applied terminology (used for a certain purpose such as translation, indexing, etc.) or a
'general' terminology (used for all possible applications). The major difference between applied and general terminology is that a general terminology tends rather towards the complete coverage of a field. In other words, the whole concept structuring of a certain field of knowledge is supposed to be provided by such a terminology. It is a sort of "scholarly terminology" placing fidelity to the coherence of the relevant field above all else.

But the end-user is an important element to take into consideration when conceiving a terminology. It is possible to create a tool for specialists of a certain field (knowledge representation device), a general purpose tool, or a terminological database specially tailored for translation and technical writing purposes. In both cases the users have to master the terms used in the relevant fields and the relations they entertain. The following diagram sums up the question of systems comparison:

Two different systems S1 and S2 applied to the same corpus C1 can yield two comparable results. But if we take into consideration the purposes of S1 & S2, the nature of linguistic elements (E1 and E2) they use and the conditions of their use (end users) the comparison between the two systems would be rather difficult. For this reason it is important to organize the system in such a way that their output can be comparable. Another problem has to do with the nature of corpus provided as input (type of documents, fields of knowledge). These elements are not necessarily the same for S1 and S2. Given that, it is important to provide the systems with different types of corpus (see above).

4. Evaluation Paradigms

Comparing terms extraction tools and the logico-semantic relations they entertain with each other is not as easy task if we compare it to the work done within the framework of conventional algorithms comparisons (such as sorting algorithms, etc.). For those we normally have as input a set of totally formalized and structured data (normally digital) at our disposal. Furthermore the expected results are usually defined beforehand (sorted value lists, etc.). The purpose of 'conventional' evaluation process is not to compare results but rather to rank algorithms according to quantifiable criteria: processing time, required processing resources (usually measurement in required memory space).

In our approach the concept of evaluation is different even if we take into consideration required processing time, required systems resources (i.e. required linguistic tools, electronic dictionaries, etc.). The only really relevant evaluation criteria is the quality of the resulting terminology. So we are going to be focusing mainly on the relevance of the terminology obtained. Within this perspective we are going to put forward two different evaluation methods.
4.1 Digital Evaluation

The first way is to use digital quantification based on methods of statistical correlation i.e. correlating terms and the semantic relations they hold with each other. But this raises other theoretical questions linked with the interpretation of the resulting measurements:

(a) What is the relevance of terms and semantic relations provided by the systems being tested? Does the terminology satisfy minimum requirements? Do we need to define a minimum level of terms production?

(b) Are discrepancies meaningful? For example, it could be that most of the systems being tested are having qualitatively poor outputs, while only one or two produce worthwhile results.

(c) Finally to undertake statistical analysis of the results it is necessary to have access to a reference terminology dealing with the relevant field of knowledge. But such a terminology has to be evaluated beforehand and only a specialist of that field can assess its relevance and exhaustivity of coverage. We have then to quantify the variations between the various results and the terminology of reference to define the degree of relevance enabling us to rank the systems being evaluated.

4.2 Qualitative Evaluation

We are putting forward a strategy based on qualitative evaluation. In other words we would rather submit the results to specialist (we are distinguishing here between specialists, terminologists and knowledge engineers). This approach supposes that manual results obtained by field specialists can represent a reference terminology. We have to admit that for the same corpus we can get different results (i.e. two different terminologies produced by two different teams of specialists). We are assuming that manual results represent a coherent class. To undertake this evaluation we distinguish two types of human expertise (a) Those who undertake evaluation, i.e. specialists who will rank systems according to the relevance of terms to a certain field;

(b) Those who will establish a reference terminology from the same corpus processed by the tested systems.

To guarantee the accuracy of the test and its neutrality the external aspect of the results (those obtained by automatic processing and the ones obtained by manual) processing should be the same.

4. Concluding Remarks

Morpho-syntactic methods combined with statistical modules produce lists of terms without significant semantic relations while semantically-based systems tend to focus on semantic relations resulting in rather limited terms productivity. The idea is to tend towards the integration of systems in order to enhance their capacity.

As far as evaluation is concerned the most relevant criteria to take into consideration is the quality of the resulting terminology. Within this perspective we are putting forward a strategy based on qualitative evaluation. The idea is to submit the results to specialist (i.e. field specialists, terminologists and/or knowledge engineers). This approach supposes that manual results obtained by field specialists can represent a reference terminology.
Notes

1. We are conducting this research within the framework of a program sponsored by the "Association des Universités Entièrement ou Partiellement de Langue Française" (AUPELF) - an international association whose mission is to promote the dissemination of French as a scientific medium. This research aims to evaluate software capabilities in automatic terminology building from corpuses in French. Softwares submitted to this evaluation are conceived by French and Canadian research institutions (National Scientific Research Centre and Universities) and/or companies: EDF (French Electric Company) among others.

2. The question we are raising which terms are best to build a terminology of reference (absolutely all the terms representing the field of knowledge) and which ones are best used for information retrieval.

3. For a complete and thorough description of ACG the reader may refer to (Descles, 1990 or Jouis, 1994) In this model we distinguish a certain number of elementary types which enable us to categorise the entities of a certain field: individualizable (i.e. separately accountable) entities, Boolean entities (expressions denoting a value of truth), massive entities, collective classes, distributive classes, places, etc.

4. Static relations permit the description of some states (static situations) related to an area of knowledge. Static situations remain stable during a certain temporal interval where neither the beginning nor the end are taken into account. We distinguish more than twenty relations, particularly: identifications (or equivalence), incompatibility between two entities, dimensions (measures, etc.), cardinality, comparisons between two values, inclusion among distributive classes, belonging of one individualizable entity to another distributive class (broader/narrower terms), -relations part/whole among collective classes, localizations of one entity in one place.

5. Formerly TAIGA (see Penel 1994).

References


Abstract: We can observe terminologies changing across time, languages, cultures, subject fields, users and usage. It is not only terminology per se that is changing, but so is its structure. In the present paper we want to postulate the following: since knowledge is closely connected with conceptual continuum which is structured with the help of system of notions represented by terminological system, then the cultural-linguistic divergences (and changes occurring here) affect the knowledge organization through terminology structure. The tool for controlling this process is Comparative Terminology Science.

1. Introduction

In the broad scope of the problems concerning terminology organization and change, four most important issues will be highlighted:

1. Terminology management as a basis for knowledge organization and knowledge engineering
2. Levels and immediate reasons of terminological changes
3. Synchronous and diachronous dynamics of terminology
4. Deep reasons for terminological fluctuations and tools for managing them.

It is our intention to aware the reader of the ways terminology and knowledge organization are determined by changes in subject fields, cultures, languages, and usage. By change we understand both major and minor terminology changes, be it dramatic change in terminology structure due to technological revolutions and complete change of paradigm, or just dynamic variation in terminology usage, due either to divergences in scientific schools approach, or to translation, or to different language structures and cultural traditions. Analysis of terminology organization and change is based on Russian, Ukrainian, and English terminologies of Mathematics, Logics, Computer Science, Linguistics, and Applied Semiotics.

2. Terminology Management and Change

As different from Terminology Science proper and terminology teaching, terminology management is a practical activity of any purposeful manipulation of terminological information (Budin, 1995, 41). Terminology management is most essential with respect to:

1. Computer-assisted terminology processing with the help of specially designed systems for translators and technical writers who actually "manage" multilingual terminology (i.e., database operations of data entry, retrieval, etc., in Translators' Workbenches and similar systems)
2. Basic work on contents of terminological information used for a broader variety of purposes, like industrial and information management, the work involving term selection, concepts description, checking and updating multilingual definitions, source information additions, etc.

To hold the edge, terminology management for any of the aforementioned purposes should identify the kind and degree of changes/divergences occurring in this or that subject field. For instance, in multilingual dictionaries on subjects other than natural sciences, the socially and culturally determined knowledge structures may differ greatly (Picht, 1995, 24); the terminology manager
must add language-pair-related information to such a dictionary, in order to make clear the kind and degree of the divergences revealed. So, we agree with Picht (1995, 29) that any TBD worthy the name must of necessity contain extensive semantic information, i.e., explanation/definition, non-verbal means, plus information about subject field, through the position of the concept in a system of concepts; the latter permits the concept to be viewed in the context of the system of concepts, allowing the TBD user to make a proper decision. This is easier said than done, for the simple reason that most terminology managers do not have enough background logico-linguistic information to do the job; so, more special research is necessary.

Recent investigations rightly point out the increasing fusion of Terminology and Knowledge Engineering, because all central areas of knowledge engineering are heavily based on terminology work of the following kinds:

• Terminology work relevant for special subject fields
• Content analysis of documents
• Extraction of knowledge from multiple sources, etc.

These are the common areas of interest of Terminology and Knowledge Organization/Engineering. Most importantly, terminological system is just the instrument that turns a conceptual continuum into a system of concepts—by making it structured, fixed—and, consequently, preparing a basis for knowledge organization. Ordering of terminology is, in a way, ordering of knowledge. The system of notions/concepts is subject to changing in time, as our knowledge about the world develops; it also varies from language to language—and so does the system of knowledge. Thus, it is clear that knowledge organization leans heavily on terminology organization, though neither of them is a part of another.

It is almost a truism in linguistics that human beings cannot exercise any significant influence upon language as such in terms of change, for language is an entity independent of us, not created by man. This is the area where we can never be in perfect control; we are rather observers than managers. Still, we do speak of terminology management, and we know very well by now that terminology and knowledge can be organized, at least certain things can be done to our best benefit. How to considerably improve our level of terminology management under conditions of constant change of this open system (and terminological system, as any other language system, is open by definition)—that is the question. To adequately answer it, we should know more about the nature of terminological changes, of how these changes occur, and when, and where, and why. It is impossible to cover this broad range of questions in the frames of the present article, so, of necessity, we only outline the research to be carried out in this direction, and present some of our own results.

3. Levels of Terminological Change

Let us consider changes in terminology (and knowledge) organization, along with immediate reasons underlying them. The changes take place on the following levels:

• Individual level. Changes are inevitable here as specialists of one and the same subject field do not necessarily think in the same way. Individual perception, individual vision, individual thinking are highly characteristic to scientific world. Thus, Linguistics is a centuries-old subject, still the term Perfect tenses remains controversial for years. In Sciences, especially in Mathematics, where definitions of basic terms are strictly given at the beginning, individual level changes/variations are rare; in Humanities, on the contrary, they are almost legitimate, and should therefore be given special attention.

• Scientific school level. Changes/differences here are determined by the approach adopted within the same scientific school or group, and concern both lexical and semantic levels of term
coining and term usage. Different approaches often give rise to new ideas generation and development, which eventually leads to progress in science, but "on the way" to this progress, terminological conflicts and differences present a stumbling stone for terminologists and knowledge engineers. Thus, striking semantic changes could be seen in definition of the economic term *surplus value* when we consider the definitions made in Economics in the former Soviet Union and in the West. Viewing the term through the prism of communist ideology made a difference, adding a sema of *instrument of class exploitation* to the definition of otherwise equivalent terms *прибавочная стоимость* (Russ.) and *додаткова вартість* (Ukr.).

A less dramatic example can be drawn from Applied Semiotics. Russian and American scientists came in touch in 1995 on Semiotic/Situation Control issues, and, since for many years the national schools worked in isolation from each other, often under security clearances, both parties immediately faced misunderstanding their basic terminology. Thus, even the terms *Semiotics*, *Semiotic Control*, representing the notions central for emerging Applied Semiotics, are interpreted differently.

**Subject field level.** It is interdisciplinary interaction that triggers this kind of change. In other words, subject fields often borrow the terms from one another, modifying their meanings according to their own purposes.

This is what happened to *predicate*, originally a linguistic term, when it was used by Aristotle in Logics where it acquired a new meaning. In Mathematical Logics *predicate* was further developed to acquire a meaning "function defined on a set of objects, and taking values of truth or false", or, more exactly, "a predicate or sentential function is a formula which contains some variables and describes a property or a relation. The predicate \( x > 0 \) describes the property of being a positive number. The predicate \( x=y \) describes the identity relations" (Davis and Weyuker, 1983, p. 4). This case can also provide us an example with *predicate* as a linguistic term landing in different terminological systems in different languages; thus, in Russian/Ukrainian we have two terms, a primordial and a borrowed one, to denote what is represented by sole *predicate* in English. In Russian/Ukrainian normative grammars, it is a primordial term *сказуемое* (standing for *predicate*), while in Russian/Ukrainian Structural Linguistics, Mathematical Logics, and Logics we mostly find *предикат*, a borrowed term, though in certain cases a primordial term is used as its absolute synonym. As a result, we can find expressions like *предикат* или *логическое сказуемое*/*предикат* або *логічний присудок*, a phrase completely untranslatable into English because of appearing tautology "predicate or logical predicate".

**Language level.** Due to divergences in the accepted terminological norms or semaseological structures of different languages, establishing direct correspondences of terms and adequate translation often present a problem.

A typical example of different terminological norms is the one where in one language we have only a general term, while in the other—two terms for derivative notions only. Thus, for Computer Science and Applied Semiotics Russian and Ukrainian have only one term: *сжатие* (Russ.) and *успільнення* (Ukr.); in English, on the other hand, there is no general term like that; instead, there are two derivatives: *compression of information* (the one without information losses) and *aggregation of information* (the one with losses). This phenomenon accounts for divergences in the terminological systems and calls forth different visions of reality, which results in misunderstanding between scientists.

Semantic divergences on language level are often there when we deal with the terms coined independently. Semantic divergence may concern not the structure, as in the previous example, but imaging, or motivation of terms; thus, this is the case with the Computer Science term *разъем* (Russ.), and its English equivalent *connector*. The matter is that *разъем*, a primordial
term, verbally means "disjunctor", "disconnector", so the Russian phrase "соединить с помощью разъема" — "to connect with the help of disconnector" sounds controversial—but it is still perfectly Computer-Science correct.

Some cases of this kind can be improved by harmonization; the important point here is that what we should tackle is not harmonization of isolated pairs of terms in the languages concerned, but the harmonization of terminological SYSTEMS as wholes).

- **Cultural level.** Cultural differences may sometimes bring about divergences in terminology. The appropriate example here will be methodology методология, as it is understood in Russia, in a more general sense, implying adherence to Marxist-Leninist ideology and presupposing materialistic thinking Marxist-style; as a result, for scientists the term bears a negative implication of showing one's political orientation. A phrase methodologically correct—методологически верно — meant only one thing: that of implementing Marxist materialistic principles in one's investigation. Coming now to the West, Russians cannot immediately accept this term in a different, very simple meaning, that of a particular procedure or set of procedures, or the analysis of the principles or procedures of inquiry.

Another typical example is Sciences, and different, totally culturally-determined interpretations this term has in English-American and Russian-Ukrainian environments. While in Russian and Ukrainian the term Sciences generally stands for anything that can be studied or learned as systematized knowledge (i.e., covering Natural Sciences and Humanities), in the West it is usually used with reference to exact or natural sciences, the ones that use "exact methods" of investigation, as opposed to Humanities. Hence, in Russian/Ukrainian there is the general term NAUKI and its derivatives: естественные науки (Russ.), природные науки (Ukr.) and гуманитарные науки (Russ.), гуманитарні науки (Ukr.), meaning Natural Sciences and Humanities; in English, however, we have no general term like this at all. This is nothing else but a difference in terminology structure due to cultural traditions.

4. **Types of Terminology Dynamics**

On each of the aforementioned levels, there is a certain dynamics. We will refer to the lexico-semantic dynamics, which may be interpreted both in the aspect of synchronous variance across specific communicative situations within a ten to thirty years span of time, as well as in the aspect of diachronous change across stages of subject fields development that takes place over many years. Thus, synchronous variance and diachronous change can be regarded as two types of terminology dynamics.

Things are relatively simple with the diachronous changes of terminology, inevitable with the development of knowledge in time, when people are redesigning some subject fields, introducing the new terms to denote the new concepts, and modifying structures of conceptual continua. Studies of terminological systems over the years, diachronically, prove that the only constant here is change, which is certainly natural.

Synchronous variance of terminology presents more difficulties to perception due to its more subtle character. However, this is what terminology managers should first and foremost take care of, because here we deal with actual usage, not registered in dictionaries yet, but strongly affecting the sort of terminology we will have in future. Let us consider a simple example of lexical internationalisms as they fluctuate in the Russian texts of Mathematical Logics. "Terminology dynamics" here will mean historical changes of the quantitative aspect only. Three lexical classes have been investigated: pure internationalisms, hybrids, primordial terms.

The corpus of 2980 English-Russian pairs of terms has been selected from scholarly paper titles and their translations in the monthly Abstracts Journal of Mathematics (AJM) covering a
thirty-two-year-period. This analysis was meant to show dynamics in terminology usage with respect to internationalisms, while the "static" data have been taken from four classic monographs that use universally accepted, standardized terminology. The results, briefly, are as follows:

<table>
<thead>
<tr>
<th>Periods</th>
<th>Sources</th>
<th>Pure internationalisms</th>
<th>Hybrids</th>
<th>Primordial terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Monographs</td>
<td>29.68%</td>
<td>38.91%</td>
<td>29.41%</td>
</tr>
<tr>
<td>1</td>
<td>AJM, 1954-1963</td>
<td>37.15%</td>
<td>38.41%</td>
<td>24.44%</td>
</tr>
<tr>
<td>2</td>
<td>AJM, 1964-1973</td>
<td>36.66%</td>
<td>45.10%</td>
<td>18.24%</td>
</tr>
<tr>
<td>3</td>
<td>AJM, 1974-1986</td>
<td>34.60%</td>
<td>49.79%</td>
<td>15.61%</td>
</tr>
</tbody>
</table>

Table 1: Internationalism Dynamics in Mathematical Logics

We received experimental evidence that present data fall into three periods; already in the first period the ratio of the three aforementioned term classes differed from the static state (i.e., analogous distribution in monographs). Thus, in monographs these three classes are approximately equal, while in actual usage in 1954-1963, the primordial terms quantity considerably drops, and the quantity of internationalisms increases. Later, the quantity of primordial terms was constantly dropping, the quantity of hybrids—constantly increasing, while the quantity of pure internationalisms did not change significantly. The tendency to relative stabilization of this process indicates, in our opinion, that over time Russian terminology developed with a sufficient degree of independence. Nevertheless, the revealed excesses of internationalization—which cannot be substantiated in the narrow frames of this paper—make it clear that self-regulation in LSP is far from sufficient; so, ordering of terminology of the rapidly developing subject fields, like Mathematical Logics, should be well-grounded by comparative-terminological analysis; harmonization of national terminologies and renovation of bilingual dictionaries must be preceded by studying terminologies not only in statics, but in dynamics as well, because this is the only way to learn about real usage.

Both types of terminology dynamics should be consistently studied by terminology managers in order to introduce timely improvements into the terminology structure to date.

5. Hypothesis of Terminological Relativity

Knowledge engineering specialists, whether they want it or not, work with signs that represent concepts. These may be the signs of either formal or natural language, and, inevitably, the structure of signs they use for communication and knowledge representation imposes its structure on knowledge organization (or on knowledge base) that they build. This happens because the specialists intuitively tune in this knowledge base structure to the sign (language) structure inherent in them, i.e., they naturally build knowledge base according to the way they structure knowledge—in accordance with their native language and native terminology. More exactly, being positively ancestral to B. L. Whorf and the Sapir-Whorf linguistic relativity hypothesis, we may speak here of hypothesis of terminological relativity. Really, structures of terminological systems, reflecting the conceptual structures, are not constant and can change from language to language, from one subject field to another, so they are relative. Each of us views reality through the lattice/prism of our own terminology, representing a system of concepts in our native language. This is why interdisciplinary and international efforts in science often bring about success—because they allow to view reality from different angles, they allow our vision of our subject field (to a great extent predetermined by our terminology) to be complemented by
another vision, equally right, so that this double vision could prompt us to pay attention to different relations/properties of objects and phenomena under analysis.

As different from linguistic relativity, a situation here, in terminological relativity, is two dimensional. Thus, here we deal not only with system divergences resulting from language structure, but also with divergences determined by specific subject field vision of the object of study. Some subject fields are closer than the others, some even overlap, some just study the same objects/phenomena from different points of view, with different purposes, applying different methods. So, the properties revealed by them about same object under study may be essentially the same, but the concepts/notions land different networks, different systems of concepts that are pre-established in the given subject field; meanwhile, the new terms are placed into pre-established terminological systems. This is the added dimension of terminological relativity: different vision and frame of mind, when investigating same object/phenomena, bring the concept into different systems of concepts, and put the term into different terminological systems.

Thus, knowledge structure reflects terminological structure, as the structure of terminology in our native language imposes a structure of knowledge representation upon us. Consequently, if we want to create knowledge organization which could be universal enough, we cannot rely only upon analysis of terminology and knowledge structure in ONE LANGUAGE. Comparative Terminology Science is an indispensable tool for harmonization and for creating knowledge representation that would be language-independent.

6. Comparative Terminology Science and Terminology Harmonization

Comparative Terminology has recently evolved into a relatively autonomous trend, forming on the basis of General Terminology Science, Comparative Linguistics, and Translation Studies. In fact, we can speak of the new paradigm under way in the peripheries of these universally accepted linguistic giants and of a new trend emerging there, the trend that claims to be autonomous (for more details see (Citkina, 1994)). This autonomy claim is methodologically correct, based on the fact that Comparative Terminology Science (CTS) has a clearly distinguished object, subject, and methods of investigation, as well as theoretical and applied products, and goals of its own. Most importantly, it has its very own applications in terminological practice: analysis and establishment of interlanguage correspondences and correlations of terms and terminological systems on different language levels generate adequate terminological maintenance for international unification, standardization and harmonization of terminological systems, for bilingual lexicography, for computer-assisted translation and multilingual terminological banks. Though within this trend we can see no uniformity either in subjects and goals or in methods of investigation, a certain experience has been accumulated, and the trend's inner logic of development points to the necessity of consolidating on a uniform theoretical platform. To establish this platform, we should define specific object and subject of investigation in CTS, as well as specific system of methods and products. Let us introduce these definitions in succession.

The object of investigation in CTS are bilingual terminological pairs and terminological systems as the wholes. The subject of CTS are the regularities of the system of similarities and divergences in lexical, grammatical and semantic structures of terms and terminological systems of the languages compared, along with the principles of term translation.

With respect to methods of investigation, CTS, as many other emerging trends, makes use of a number of methods provided by neighboring branches of Linguistics. These methods include structural-comparative, comparative-typological, semaseological, contrastive-definitive, and contextual analyses, procedures of linguo-statistics and engineering linguistics, of informants poll,
of information theory; besides, a number of procedures of investigation appeared anew. Two
types and four kinds of CTS methods and procedures are described in (Citkina, 1993).

Theoretical products of CTS are the revealed types of correlations of confronted terms and
term systems on lexical, grammatical, and semantic levels. Applied products have already been
briefly mentioned above. Both theoretical and applied aspects of CTS have been clearly
distinguished to date, and the distinction will be enhanced as research progresses.

The present-day goal of CTS as a science in its own right is further deepening comprehen­sion of the subject of CTS, and further elaboration of its methods and procedures of investigation;
a more remote goal is creation of a comprehensive theory of comparing terminological systems,
which presupposes establishing general interlingual and interterminological regularities, based on
different language pairs and terminologies of different types. CTS is currently on the stage of
"sensual contemplation", making its first steps towards the stage of "abstract thinking". Therefore,
now the comparative-terminological analysis goal is not immediately reaching any universal
conclusions but just accumulating more particular comparative material, and working out
universal, generally accepted methods, critically necessary at present for obtaining the compar­able
results on the broad range of terminological systems and languages. This unheroic background
work will prepare solid ground for consolidating theoretical and applied works in CTS. Thus,
we believe that short-term tactics should prevail for now, though the long-term strategy should
be born in mind.

Comparability of results obtained by different linguists engaged in Comparative Terminology
research is what we are after right now; and this is a realistic objective in the frames of CTS.
Comparison here is made in four areas:

- Intrasystem comparison, i.e., investigation of correlation of different parameters, or
dynamics of one parameter
- Comparison of data within a terminological system of the same subject field in different
languages
- Comparison of data in different terminological systems within a fixed language pair
- Comparison of terminological systems of different types (for instance, sciences and
humanities types)

Consistent terminology work in these areas will reveal relevant changes and divergences
embedded in a multitude of languages for special purposes and provide us a sound basis for
terminology harmonization.

7. Conclusions

On considering some practical aspects of terminology organization with respect to change,
we can see two kinds of reasons for the changes, immediate and deep. While the immediate
reasons come out clear on different levels of change, each characterized by its own type of
dynamics, deep reasons are accounted for by terminologic relativity. Terminological relativity
phenomenon, however backstage and unseen, has a profound impact on terminology and
knowledge organization. Harmonious, universal terminology and knowledge organization—this
is what we are seeking. A great deal of work should be done to achieve this goal, even given the
best of circumstances. However, it is doable, and a powerful instrument to do it is Comparative
Terminology Science.
References


Preparation Terminological Definitions for Indexing and Retrieval Thesauri: A Model

Abstract: A model for standardizing existing definitions and/or writing new definitions for thesaurus descriptors has been developed, within the framework of a research project concerned with the usefulness of terminological definitions for indexers working with a thesaurus. The proposed model is an expansion of a model presented by Sager and L'Homme in 1994. Examples of its application in a thesaurus describing the field of Adult literacy programming and training are introduced.

1. Introduction

A critical function of the indexing and retrieval thesaurus is that of clarifying meaning, of revealing the relationship that exists between a concept and its expression in the controlled language of an information system. When the meaning of all descriptors is clear and unambiguous, indexers are expected to more consistently and accurately select the most precise and appropriate descriptor to express an indexable concept.

Contemporary thesauri, however, have become large and complex vocabulary control tools, and it is increasingly difficult for the indexer to assess the intended meaning of each one of the terms that constitute the entry vocabulary. Many descriptors appear to have very close if not overlapping meanings, and it is often not possible to distinguish among them and to identify correctly the term needed in a given indexing situation. The difficulty for indexers to elucidate meaning is felt even more in social science thesauri.

Traditional thesaural networks of semantic relationships and scope notes are considered as the main sources of descriptor definition. They are not always efficient and sufficient however, especially when the entry vocabulary in a thesaurus is large, and when semantic and terminological ambiguities exist in the field described.

The lack of proper defining information creates difficulties in using a thesaurus as indexing aid, and, more importantly, leads to inconsistencies in descriptor assignment, with assumed consequences on representational predictability, indexing correctness, and overall indexing quality.

The thesaurus is, and is likely to remain for years to come, an essential indexing tool, and it is important to consider at this time alternate ways of providing indexers with much needed semantic information about individual subject descriptors. Within the larger context of a study focusing on the potential usefulness of terminological definitions in the thesaurus used as indexing tool, this paper presents a model for the development of such semantic addition.

2. Definitions in Thesauri

The idea of providing definitions for descriptors has surfaced regularly for as long as thesauri have been in existence and use. Mooers (1963) believed that definitions were necessary to transform index terms into descriptors, giving them a very distinct meaning determined from context of use. Mooers excluded the possibility of using regular dictionary definitions.

Over the years, some real definitions, to be distinguished from scope notes, were provided for thesaurus descriptors. Definitions were either grouped in a separate section, or presented in...
the main section of the thesaurus, as part of the defined term record. These definitions, however, have been of questionable quality and usefulness. There were never any precise guidelines available to help in the process of writing and standardizing them.

Recognizing obvious links between terminology work and thesaurus design, interesting proposals for developing terminological thesauri came from researchers in the field of Terminology. Among others, Sager (1982 and 1990) envisioned a semantically rich tool which would be useful to a wide audience of specialists interested in information production and communication. In a terminological thesaurus, each descriptor record would include a standardized definition, as well as semantic relationships with other terms.

3. Terminological Definitions: Nature and Functions

Within the field of Terminology, the objective of describing accurately the meaning of a concept is normally pursued through the development of one or more conceptual or true definitions for each term in the language of a specialty (LSP). The terminological definition is domain dependent; it describes the intension of a concept and it determines its position in the specific system of concepts to which it belongs.

Although many types of acceptable terminological definitions have been identified and used, analytical definitions, which define a concept by identifying its genus proximus and its differentia specifica, are seen as the necessary foundation of most terminological systems, becoming an essential component of terminologies and termbanks.

4. Defining Rules and Models

The primary defining rules are well-known. The rules of adequacy, completeness, and substitution, as well as comments on tautology and circularity, refer to the semantic contents of the definiens. The rules of brevity, simplicity, clarity, and comments on the use of negative forms and synonyms, refer to the style of the defining text.

It appears that, even in termbanks, definitions often leave much to be desired in clarity and adequacy. The rules are not sufficient to ensure that the semantic content of definitions is sound, complete and appropriate.

Although the need for defining templates, which would call for and organize in a standard order of appearance the various elements of a well-formed analytical definition has long been recognized, few of the defining patterns that were presented were formal enough to be of much help as content and style regulators.

Sager and L'Homme have recently proposed a model for standardizing existing term definitions that "consists of a regularized form of the traditional analytical definition by categorizing and restricting the modes of description and thereby reducing the free-text element in the defining phrase" (1994, 352). Sager and Ndi-Kimbi are also searching for "more sensible rules and proper patterns for definitions that can be used for authoring packages for the writing of definitions" (1995, 72).

The model for rewriting existing definitions is an uncomplicated one. It divides the defining phrase into seven components, as shown in Fig. 1.

The model does not offer a solution to the problem of identifying the characteristics which are necessary and sufficient in a well-formed definition. The choice of characteristics to be included in (6) above will depend on what type of concept is being
defined, the concept system that it is part of, and the needs of the user. It has been suggested that one way to construct an effective *differentia* would be to answer one or more of the news reporter’s question: Who? What? When? Where? Why? and How? (Strehlow, 1983). Dahlberg (1989) has proposed a similar approach.

5. Description of a Corpus

Our test thesaurus is made of 361 descriptors extracted from an existing and widely used thesaurus in the field of Adult literacy. Of these descriptors representing core concepts in the field, a large majority (275) refer to entities/objects. 259 descriptors are Level 2 terms (i.e., a precoordination of two concepts which may or may not be part of the thesaurus).

The conceptual structure and terminology of the field of Adult literacy illustrate most of the semantic and linguistic problems inherent to the social sciences. Most terms used by Literacy specialists are borrowed from the neighbouring fields of Education and Adult education; the rest comes from a large number of disciplines (Linguistics, Psychology, Sociology, etc.) As is common practice in the social sciences, the meaning of most borrowed terms has been slightly modified, and often more than once, to reflect new general points of view or even personal opinions. Through a process of unregulated precoordination, terms have been created on demand, and lack of standardization now leads to communication of information difficulties among specialists, which a thesaurus in the area is expected to fix. But since *series* of terms with obviously overlapping meanings have made their way into the thesaurus (e.g., literacy consultants, literacy coordinators, literacy facilitators, literacy instructors, literacy practitioners, literacy tutors, literacy workers), it looks as if nothing short of a verbal definition will be of any help to the indexer of literacy-related information in the process of index term selection.

6. Expanding Sager and L’Homme’s model

Using Sager and L’Homme’s defining model as a tool for transforming existing definitions, or for writing new definitions for all descriptors in our corpus, this project develops section (6) of the model with a view to standardizing the nature and order of conceptual characteristics that will be emphasized and considered sufficient for the completeness of definitions within our context.

The expanded model now being tested is illustrated in Fig. 2. The template shows how
Section (6) has been subdivided further, using seven facets (eC1 to eC7) which we consider as being of primary importance in establishing clear distinctions among concepts. The model allows for great specificity in the description of the various types of concepts (entities, activities, properties) that are represented in our thesaurus.

Term:

"Source definition(s)"

(1) Domain:
(2) Concept class:
(3) Genus:
(4) Genus class:
(5) C/G:

(6) eC1: [nature: being..., having..., doing....]
(6) eC2: [purpose/function]
(6) eC3: [means/instrumentation]
(6) eC4: [origin]
(6) eC5: [destination]
(6) eC6: [place]
(6) eC7: [time]

(7) iC1:

Fig. 2: Proposed defining template for an entity/object (type of concept)

7. Application of the Expanded Model

The proposed model is used to add semantic information in a thesaurus which describes a limited domain and which we consider as a logical conceptual system. In this thesaurus, all definitions will be interdependent. The use of the templates requires prior categorization of all concepts represented, as well as a thorough search for existing definitions of all concepts and/or terms.

In preparing definitions for all 361 descriptors in our corpus, three situations have been encountered:

a) one or more appropriate definitions already exist for a concept, and all that needs to be done is to standardize their content and style. A standardized definition for library literacy programs is provided as an example of this situation in Fig. 3. Note that the genus proximus, literacy programs, is itself defined in the same corpus and will be marked as an entailed term in the definitions of more specific concepts.

b) one or more inadequate definitions exist, and they can be used as a basis for the creation of an adequate standardized definition. A standardized definition for literacy specialists is provided as an example of this case in Fig. 4. Note common defining mistakes in the source definitions: the first definition shows an inappropriately restrictive definiens, while the second definition defines the term by itself.

c) no definition has been obtained for a specific concept; a standardized definition is
Term: library literacy programs

Source definitions: = "programs designed, administered, and staffed by a library"

(1) Domain: Literacy
(2) Concept class: AE
(3) Genus: literacy programs
(4) Genus class: AE
(5) C/G: a type of
(6) eC4: designed, administered, and staffed by libraries
(7) iC1: usually conducted in a library setting

library literacy programs = literacy programs designed, administered, and staffed by libraries, and usually conducted in a library setting.

Fig. 3: Standardization of an existing adequate definition

developed using elements of adequate existing definitions of the more general constituting concepts, or using elements of adequate definitions of specific concepts with which the definiendum shares essential characteristics. A standardized definition for the descriptor literacy centres is provided in Fig. 5 as an example of such a situation. In this example, the specification of the genus proximus is made necessary by the fact that the natural genus, Centre, has no meaning attached to it.

Term: literacy specialists

Source definitions: = "individuals involved in literacy research"
"person who is specialized in literacy"

(1) Domain: Literacy
(2) Concept class: MEA
(3) Genus: [Individuals]
(4) Genus class: MEA
(5) C/G: a subset of
(6) eC1: who have extensive theoretical and practical knowledge of issues pertaining to literacy assessment, literacy instruction, and literacy programming

literacy specialists = individuals who have extensive theoretical and practical knowledge of issues pertaining to literacy assessment, literacy instruction, and literacy programming.

Fig. 4: Modification and standardization of existing inadequate definitions
Concept: literacy centres

Source definitions: "centres = facilities serving as a focal point for activities or services"
"community centres = facilities at which social, educational, recreational and other activities are held for the benefit of the community"

(1) Domain: Literacy
(2) Concept class: MEI
(3) Genus: [Centres]
(4) Genus class: MEI
(5) C/G: a

(6) eC1: physical facilities
(6) eC2: at which a variety of literacy services are provided
(6) eC2: at which literacy related activities such as literacy classes and literacy workshops are held

literacy centres = physical facilities at which a variety of literacy services are provided, and at which literacy related activities such as literacy classes and literacy workshops are held.

Fig. 5: Creation of a standardized definition

Terminological definitions are not complete definitions. Their only function is to make clear what difference there is between two or more concepts in a particular domain or LSP, and only the required essential characteristics needed to do so will be used in the description of the concept. The standardized definitions given here as examples show that prior knowledge is required from their reader. The meaning of terms like "physical facilities", "individuals", or "libraries" is considered common knowledge; these terms, which do not belong to the same LSP as the term which they are used to define, are not themselves defined in the thesaurus.

8. Conclusion

Our application of the proposed expanded defining model has shown that, even in a field such as Adult literacy where the terminology is far from being standardized and has a tendency to vary significantly from one geographical area to another, and from one social environment to another, it remains possible to describe concisely through a verbal definition the semantic coverage of a descriptor, adding in this manner to the semantic information already available to the indexer to assist in term selection.

The same defining model could also be used to modify definitions when the meaning of a concept changes over the normal course of evolution of a field of knowledge, or under the influence of new points of view, new connotations, etc.
Notes

1. In the NASA Thesaurus, for example, the grouped definitions constitute a thesaurus dictionary, designed as a complement to the thesaurus rather than as one of its parts.

2. This has been, and still is, the most common practice in thesauri. Definitions appear as scope notes. Scope notes then play the dual role of describing the intension of concepts and providing rules for using the terms that represent them.

3. Analytical definitions are also known as true, generic, Aristotelian, and referent-oriented definitions.


5. The Canadian Literacy Thesaurus / Thésaurus canadien d'alphabétisation. Toronto, ON: Canadian Literacy Thesaurus Coalition, 1992.

References


Nancy J. Williamson
Faculty of Information Studies, University of Toronto

Deriving a Thesaurus from a Restructured UDC

Abstract: The derivation of a thesaurus from a new schedule for UDC Class 61 Medical Sciences which has been restructured into a faceted classification system using the framework provided by the Bliss Bibliographic Classification. The resulting thesaurus is intended to serve as a tool for indexing and searching but will also be the index to the 61 class itself. The background for the research is briefly described. The sources and methods used to select the descriptors and define their relationships are discussed. Problems are identified and some solutions proposed.

1. Introduction

This research is the third stage of "A Feasibility Study on the Restructuring of the Universal Decimal Classification into a Fully-Faceted Classification System" (McIlwaine and Williamson, 1994), described at the Third International ISKO Conference in 1994. The study is a pilot project in which the UDC Class 61 Medical Sciences is being reorganized and restructured using the facet framework of the Bliss Bibliographic Classification, second edition (BC2) as the basis for overall organization and content. Reports on progress are described in several articles (McIlwaine and Williamson, 1993; Williamson, 1994). The first stage, the reorganization of the topics, is nearing completion and portions of the restructured schedules were published in Extensions and Corrections to the UDC, 1995 (McIlwaine and Williamson, 1995). The second stage, the assignment of the notation, is in progress (Williamson, 1995) and the third stage, the derivation of a thesaurus from the restructured 61 schedules, is in its initial stages. The fourth and final stage will be the review and testing of the results.

2. Background and Previous Research

Normally, information retrieval systems should provide for two fundamental approaches to searching—direct access through the use of verbal descriptors and browsing capability through the use of systematic knowledge structures. Together they supplement and complement each other. Classification systems and their indexes are tools which combine these two approaches, offering the potential for optimal aid in searching information systems. Moreover, the development of a systematic knowledge structure as the first stage in the preparation of a thesaurus is a sound approach to the construction process. It helps to clarify the parameters on the subject domain and acts as a mapping device for organizing the contents of the system. It is the derivation of a thesaurus by this process, that is the subject of this paper. In this context, it is assumed that the thesaurus will be used not only as a tool for indexing and searching in systems using the new UDC Medical Sciences class, but that it will also serve as an index to the 61 class itself.

While all large classification systems require indexes for effective and efficient use, the type of index varies with the system. For example, the Dewey Decimal Classification is well known for its "relative index" and Ranganathan developed "chain indexing" to support the Colon Classification. Nor is the use of a thesaurus as an index to a classification scheme a new idea. Over the past 25 years a number of thesauri have been compiled by deriving terms from classification schemes. In some cases a new classification has been created and the thesaurus
derived from it, while in others existing schemes have been used to derive the thesaurus. Typical examples are *Thesaurofacet* (Aitchison, 1967), *BSI ROOT Thesaurus*, and the *International Thesaurus of Refugee Terminology* (Aitchison, 1989). In some cases the resulting thesaurus has been an alphabetical thesaurus alone, while in others the classification used to derive the thesaurus has been incorporated into the system, becoming an index to the classified structure. This kind of a thesaurus is best generated using a faceted (as opposed to an enumerative) classification. Indeed the title *Thesaurofacet* itself aptly describes the format of such a system. A variation on this type of system has been referred to as a "Classaurus".

Many of the systems, developed in this way are a result of research carried out by the British Classification Research Group and others. Of particular interest in this study is the research carried out by Jean Aitchison (1986), and Gerhard Riesthuis and Steffi Bleidung (1990, 1991). They represent two different approaches to a similar problem. Riesthuis and Bleidung experimented with Class 314 Demography in the existing semi-enumerative *UDC*. Building on the fundamental premise that alphabetical and systematic systems are complementary to each other, the authors created two thesaurus displays. Both contained preferred terms displayed with their scope notes, Broader Terms, Narrower Terms and Related Terms and equivalence relationships. An additional category designated as "Use for Combinations" (UFC) was included to take care of terms which are precoordinated terms in *UDC* but needed to be factored for the thesaurus because the component terms have their origins in different hierarchies. This took care of the use of colon combinations which are an important feature of the *UDC* system. Containing the same data, the two thesauri were arranged differently at the primary level. One was organized alphabetically and the other was a systematic thesaurus (as opposed to a classification scheme) arranged by *UDC* class number. Findings revealed the weaknesses of the *UDC* system, among which were out-of-date and missing terminology, and the complexity of the enumerative system which hinders the efforts to convert into a concept based system such as a thesaurus. Nevertheless the authors concluded that, given the time and effort to revise the schedules, *UDC* might be converted to a thesaurus with two complementary parts—alphabetical and systematic.

Aitchison, a member of the CRG and an expert in thesaurus construction, has developed a number of thesauri of this type, some by first constructing the classification and then deriving the thesaurus; other by drawing on an existing classification systems. Out of her experience comes her experimentation with and use of the *Bliss Bibliographic Classification* second edition as the basis. The *DS/HSS-Data Thesaurus* (Aitchison, 1993) provides controlled terms and Bliss class numbers for indexing and searching and the *ECOT Thesaurus* (Aitchison, 1984) uses software designed for the British Standards Institution *ROOT Thesaurus* to generate the alphabetical display from the systematic display derived from the Bliss schedules (Aitchison, 1986, 160). Normally, the pattern here has been to include both the classification scheme and the thesaurus. Not unexpectedly, Aitchison demonstrates that the best basis for such work is a well structured faceted system and opposed to a enumerative scheme. Also from her research Aitchison has identified problems and developed some rules and policies for conversion. Together with the ISO 2788 *Guidelines for the Establishment and Development of Monolingual Thesauri* (ISO, 1986) and *Thesaurus Construction: A Practical Manual* (Aitchison and Gilchrist, 1987), the work of these researchers forms the basis for the current project. The two pieces of research are particularly apropos to this project, since together they deal with the two classification systems being used.

3. The Restructured Class 61

In terms of its structure the "new" *UDC* Class 61 follows the *BC2 Class H* schedule for "Health Sciences. Medical Sciences". The result is a complete reorganization of *UDC*, as the *BC2
structure emphasizes collocation by systems and organs of the body (e.g., Cardiovascular System, Digestive System, etc.) as opposed to collocation by characteristics such as Anatomy, Physiology, Clinical Medicine, etc. BC2 facet indicators are part of the schedules and the captions are based on simple concepts. Of the two systems, BC2 is a much more in-depth classification and much richer and more up-to-date in terms of vocabulary than the existing UDC. As a result the present UDC 61 is subsumed under the BC2 umbrella with some characteristics of UDC being retained. For the time being, this restructured class must be used in conjunction with other schedules in the existing UDC system. Hence the same kind of notation is being used and the new 61 schedule will be "accommodated " to the existing system where necessary, to permit cross referencing and the use of general auxiliary subdivisions from the main UDC schedules. All relevant vocabulary from both systems is being retained. However, since in general, the goal for UDC is a Standard Edition approximately equal in size to the UDC International Medium Edition, English Text, some very precise topics will be grouped together as "including notes" and subsumed under the appropriate caption as exemplified in Figure 1.

Immunological deficiency syndromes, immunodeficiency

*Including* Antibody deficiency syndrome.
Aldrich syndrome. Wiskott-Aldrich syndrome.
Thymic alymphoplasia. Digeorge syndrome

Fig. 1: Including Note

Because the new 61 class is a faceted system, insofar as possible topics are given a primary location and the use of the UDC "colon combination" feature will be retained as a tool for number building. Using this device, concepts will be imported from their primary location to other parts of the system as needed. For example under "Nutritional disorders, dietary disorders, malnutrition" at "Food poisoning, (Higher plants) " the colon combination will be used to import the number for plants from class 58 to be added to the basic notation.

"Alternative" locations (with preferred location indicated) for numerous topics are a significant feature of BC2. Where appropriate some of these will be retained. Where this happens the thesaurus will need to accommodate both locations. The many auxiliary tables in UDC 61 are to be amalgamated, insofar as possible, Since most of the 61 auxiliary subdivision topics are located in the main part of the schedule in BC2, the number and size of the auxiliary tables in 61 can be much reduced.

4. Source of Terminology for the Thesaurus

Terminology for the thesaurus will be gathered primarily from the classification topics. Two problems must be addressed—the choice of preferred terms as descriptors and the determination of the form of those descriptors when removal from their context in the classification renders them meaningless or ambiguous. Frequently, multiple terms represent the same topic. Which term becomes the preferred term? BC2 tends to string terms together the captions separated by commas. A particular caption may contain multiple scientific terms or a mix of scientific and "lay" terms. Frequently, but not always, in BC2 the scientific term precedes the non-scientific term. While the choice preferred term should be based on a policy which will best suit users, for the sake of expediency, the first term in each string will become the preferred term in this thesaurus and "use" references will be created to accommodate the other terms. Figure 2 illustrates two instances of choice of preferred terms and the resulting references.
In general, synonymous terms will not be too difficult to deal with although they must be certain that the terms are indeed synonymous or quasi-synonymous in nature. Much more difficult and time consuming is the task of providing context for terms which require qualification in order to be correctly understood. These terms are of two types. As illustrated in Figure 3, some terms are "incomplete" as they appear in the classification schedules and require the context of superordinate terms if they are unambiguous when displayed thesaurus. To solve this problem some terms need to be "precoordinated" for inclusion in the thesaurus.

As can be seen from Figure 3, the resulting descriptors may, in some cases, become cumbersome. Moreover, some of them will violate the ISO Guidelines for compound terms and factoring. Nevertheless this kind of precoordination will be necessary if the descriptors are to be understandable out of context of the systematic display. Such problems occur more often at the upper levels of a particular topic than they do under more precise topics located at lower levels in the hierarchies.

A second terminological problem is one in which a general term appears in a location in the schedules which is not its primary location. For example, terms which describe the properties of a subject may come from completely different domains. This is well illustrated at the most general level in "anatomy and physiology" where in the facet (Properties, etc.) such terms as "ratio,"
"periodicity," "rhythm," "dimensions," and "shape," can be found. In the appropriate situation all of these terms have relevance in this domain and locations must be preserved for them. Nevertheless when removed from context these terms take on a completely different meaning. If the thesaurus is to be an index to the classification, these terms must be accommodated in the thesaurus. While at the time of writing it is not clear how these will be handled, two approaches seem possible. One solution would be to retain only the facet indicators in the schedule and depend on special auxiliary tables to provide for the topics themselves. An alternative to this is to "qualify" the terms in the thesaurus. For example: "Periodicity (anatomy and physiology)," "Shape (anatomy and physiology)." This would be compatible with the standards and would preserve the classified display.

5. Source of Thesaural Relationships

The facet framework being used for the classification provides for a well organized schedule and the division of topics by facet and subfacet labelled with facet indicators clarifies the characteristics of division and makes it fairly easy to recognize relationships between terms and to extract them for use in the thesaurus. The reconstructed 61 schedule will permit the extraction of the traditional thesaural relationships—equivalence (Use/UF), hierarchical (BT/NT) and associative (RT/RT). In addition, scope notes and definitions from the classification will be used in the thesaurus. As discussed above, the equivalence relationships will be easily identified from the captions in the classification schedules. Figure 4 illustrates this.

Classification: Visual aphasia, alexia, word blindness

Thesaurus: Alexia Use Visual aphasia

Visual aphasia
UF Alexia
Word blindness
Word blindness Use Visual aphasia

Fig. 4: Equivalence Relationship

Hierarchical relationships are of two types, generic relationships and whole-part hierarchical relationship. The facet indicators are of immense help in identifying this particular relationship as there are number of "(By type)" and similar facet indicators. An example appears in Figure 5.

The hierarchical whole-part relationship covers a limited number of situations where the name of the part implies the name of its possessing whole is any context. Medicine is an excellent example of this kind of relationship, because parts, systems and organs of the body qualify for this type of relationship. For example "Cardiovascular System (Parts special to the heart)" includes the pericardium, ventricles, septum valves, blood vessels, etc. These would all be related to the whole in a BT/NT relationship. Polyhierarchical relationships are seldom made clearly explicit in Class H in BC2, primarily because of the faceted nature of the system. There are a few cases where compound heading appear but these are at the very
Classification: Parasites, parasitism, medical parasitology (general)

(Types of parasites)
Ectoparasites
Endoparasites

Thesaurus:
Ectoparasites
  BT Parasites
Endoparasites
  BT Parasites
Medical parasitology (general)
  Use Parasites
Parasites
  UF Parasitism
  Medical parasitology (general)
  NT Ectoparasites
Endoparasites
Parasitism
  Use Parasites

Fig. 5: Generic Relationships

top of the class for example the multi concept "Anatomy and Physiology" is used but only to provide for work which may deal with the two topics together. The system leans toward one topic or the other.

Classification

Cells
(Physiology) Cytophysiology
(Biochemistry) Cell chemistry
(Special physiological processes)
... 
Motility
Aggregation
(Development and growth)
Change
Regeneration

Thesaurus

Cells
RT Cytophysiology
Cytophysiology
NT Cell aggregation
Cell change
Cell chemistry
Cell regeneration
RT Cells

Fig. 6: Associative and Hierarchical Relationships

Associative relationships are numerous in the new Class 61. Properties, processes, actions and their agents, actions and the products of the actions, among others are prevalent in the Medical Sciences. Typical examples are Respiration RT Absorption; Kidneys RT Transplantation; and Heart RT Open heart surgery; Determining the associative relationships is greatly facilitated by the presence of facet indicators such as (By action); (Surgery); (By chemical process): (Disorders by cause) etc.

The various thesaural relationships are present in the classification scheme and transferable to the thesaurus. However, the transfer cannot take place without careful thought at each step and in terms of each descriptor. This is particularly true of the BT/NT relationships. Some hierarchies are very long indeed and considerable patience is required to determine exactly which descriptors
are subordinate and superordinate to each other. This partly due to the physical nature of the schedules as well as to the length of the hierarchies.

6. Problems

Deriving a thesaurus from this structure is not without its problems. Some of these are a result of dealing with two quite different classification systems. Also, the fact that this project is focused on only one discipline which exists in the context of a larger system. As Aitchison (1986, 166) indicates, there are gaps in the coverage so that it is not always possible to make all of the necessary links with other descriptors. If this is true because all of the BC2 schedules have not yet been published, it is even more so with the new UDC 61. As a result of the great depth of detail incorporated into the new schedule through the use of the BC2 framework, there are many terms, particularly in other classes which may be present eventually in BC2 that are not present in other UDC classes. For example, there is detailed coverage of Biochemistry in Medical Sciences which draws heavily on Biochemistry in Science in BC2. However, this new schedule, while based on BC2 must be used in the context of the existing UDC. Repetition of concepts in many places may also add to the size and complexity of the thesaurus. However, alternate locations may not be a great problem because the enumeration of topics only takes place in the "recommended" location. The user of the classification scheme is left to develop the alternate. Another problem arises in dealing with the emphasis on parts, organs and systems of the body. Some topics are scattered in many places. A typical example is "Neoplasms" which go with the site of the disease. Thus the term neoplasms turns up in many places in the schedules all of which will need to be included in the thesaurus/index. In this context it is interesting to note that in the existing BC2 index there is only one entry for "neoplasms" which leads to the general section on "Diseases". There a general directive can be found leading to "parts, organs and systems of the body" without more specific directions.

to the precise location in the various sites.

7. Conclusion

This is a fascinating undertaking which, with time and patience, may be accomplished. However, it is a compromise at best. The results should give a more logical system which will be better suited to use in online systems than its predecessor, but the ideal situation would be a completely new classification system.

References


The Dewey Decimal Classification at 120: Edition 21 and Beyond

Abstract: The twenty-first edition of the Dewey Decimal Classification system will be published in summer 1996, 120 years after the publication of the first edition. This paper describes the major changes in Edition 21 and explores some of the social, cultural, political, disciplinary, and knowledge organization forces behind the changes. The print version of Edition 21 is one manifestation of a larger underlying database that has also responded to several forces to meet present and future needs. Developments in the database and future directions to meet new uses of the Dewey Decimal Classification are addressed.

1. Introduction

The twenty-first edition of the Dewey Decimal Classification system (DDC) will be published in summer 1996, 120 years after the publication of the first edition (Dewey, 1996). The basic outline of main classes and divisions has changed little since the first edition; below the hundred divisions, however, much has been transformed and developed to keep pace with knowledge and with social, cultural, and political changes. Compared with its most recent predecessor, Edition 20 (Dewey, 1989), Edition 21 contains substantial changes in the organization of several disciplines. It provides expansions for new areas of knowledge, accommodates many political and social changes, and includes updated terminology. Edition 21 also features several structural changes that reflect modern classification theory and make the Classification easier to apply. The print version of Edition 21 is derived from a larger underlying database that also contains several changes to provide for current and future needs.

2. Major Revisions: Public Administration, Education, Life Sciences

In the DDC, keeping pace with knowledge means accommodating new topics, as well as revising existing schedules. Schedules are revised to reflect new views of the field, to provide for new topics, and to correct existing problems in bias and structure. Edition 21 includes three major revisions: 350-354 Public administration, 370 Education, and 560-590 Life sciences. The changes in public administration and education have been underway for more than a decade; the changes in the life sciences for more than two. Each reflects basic changes in the respective discipline over time.

Public administration is completely revised. The schedule still occupies 351-354, but the subdivisions and citation order have been changed. The revision addresses a longstanding problem with U.S. bias in structure and terminology. For example, compare the outline in Edition 21 (emphasis on topic) versus Edition 20 (emphasis on jurisdiction, especially U.S.) in figure 1.

In the new public administration schedule, citation order is reversed from jurisdiction/topic to topic/jurisdiction. This was done to reflect the shift in the literature of the discipline away from jurisdiction to topic as the central emphasis. The citation order is retroactive; in general, the classifier uses a number later in the schedule as the base number,
Edition 21 Edition 20

350 Public administration and military science Public administration and military science
351 Public administration Administration of central governments
352 General consideration of public Administration of local governments
administration
353 Specific fields of public administration Administration of U.S. federal and state
governments
354 Public administration of economy and Administration of specific central
environment governments; international administration

Fig. 1: Comparison of 350-354 in Edition 21 vs. Edition 20

and then adds as instructed from numbers coming earlier in the sequence. For example:

Rural administration of irrigation projects in Australia: 354.3672170994
354.367 Irrigation projects
2 Facet indicator for General considerations (from internal table under
352-354)
17 Rural administration (from 352.17 Rural administration)
09 Facet indicator for Geographic treatment (Table 1)
94 Australia (Table 2)

In the example above, citation order is topic (irrigation projects) + jurisdiction (rural administra-
tion) + specific area. Each element after the base number is introduced by a facet indicator. The
digit "2" (from 352 General considerations) is used as a facet indicator for adding topics from 352
to base numbers later in the sequence (including other numbers in 352).

370 Education represents an extensive revision. This means that the main outline of the
schedule has remained the same, but some subdivisions have been reworked and expansions
provided for new topics. The most visible changes, the relocation of 376 Education of women and
377 Schools and religion, are responses to social and disciplinary changes. The revised schedule
reflects the current view in the discipline of education that each of these topics is an aspect of a
broader topic (i.e., kind of student and type of school, respectively) rather than a central division
of education. For example:

371.82 Specific kinds of students; schools for specific kinds of students
Add to base number 371.82 the numbers following --08 in notation
081-089 from Table 1, e.g., education of women 371.822 [formerly 376],
education of students by racial, ethnic, national origin 371.829 ...

In Edition 21, the number for the education of women (371.822) shares the same base number
(371.82) as the education of other kinds of persons, e.g., girls 371.823, Asian Americans
371.82995073.

Religious schools have moved from class 377 to within class 371 as an aspect of types of
schools. For example:
380

> 371.01-371.07 Specific kinds of schools
371.07 Religious schools
371.071 Christian religious schools
   Add to base number 371.071 the numbers following 28 in
   281-289, e.g., Catholic schools 371.0712
371.072-.079 Other religious schools
   Add to base number 371.07 the numbers following 29 in
   292-299, e.g., Islamic schools 371.077

The third major revision, 560-590 Life sciences, has been underway for over two decades. The 570 schedule is completely revised, along with 583 Dicotyledons. The rest of 560-590 is extensively revised. The 570 schedule features a reversal in citation order from organism/process to process/organism that in turn addresses a fundamental shift in the discipline away from a focus on organism to a focus on internal biological process. As in 351 Public administration, 570 features retroactive citation order in number building. The new schedule collocates microorganisms, fungi, and algae in 579; regularizes the use of notation 1 as a facet indicator for general topics in the taxonomic schedules 579-590; and provides more specific and shorter numbers for fishes and mammals. New (1996) explains the history of the development, the disciplinary forces, and the new arrangement in detail.

3. Changes in Religion, Areas, and Historical Periods

In addition to the three major revisions of disciplines described above, there have been numerous other changes to address cultural, social, and political issues. With Edition 21, we have initiated a multi-edition plan to further reduce Christian bias in 200 Religion. As a first step, we have relocated the standard subdivisions of Christianity in 201-209 to 230-270. The standard subdivisions of comparative religion are now integrated with the standard subdivisions of religion in 200.1-200.9. We have revised and expanded the schedules for two major religions: 296 Judaism and 297 Islam. A new optional arrangement for books of Tanakh (222-224) has been added to the note for 221 in the Manual.

The impact of political flux on Dewey is perhaps most evident in Table 2 and the 900s. There are numerous adjustments in Edition 21 to reflect political changes, e.g., the revision of the area table for Bulgaria, the major revision of the area table for the countries of the former Soviet Union, a new historical period for the administration of Nelson Mandela. We have worked with national libraries and other groups to prepare revisions of the area tables for Brazil, Colombia, Greece, New Zealand, Nigeria, and Norway. Even as we publish Edition 21, we are working on a revision of the South Africa area table to reflect recent changes in the political divisions of the country.

4. Accommodation of New Topics and Relationships

Many new topics that have gained literary warrant since the publication of Edition 20 are now mentioned in Edition 21, e.g., virtual reality, rap music, Internet, in-line skating, snowboarding. In addition to new topics, the Classification also accommodates new relationships within and across disciplines. These relationships are explained in the notational hierarchy; through notes in the schedules, tables, and Manual; and in entries in the Relative Index.

The Internet serves as a good example of a cross-disciplinary topic for which new links must be introduced in the Classification. Edition 21 includes a new number in computer science,
004.678, for the Internet. This number also serves as the interdisciplinary number for works on the Internet. An extensive Manual note on the Internet at "004.678 vs. 025.04, 384.33" explains the relationship of the computer science/interdisciplinary number with numbers in other schedules:

Class computer science works about the Internet that are wholly or predominantly about communications software in 005.713, e.g., works emphasizing software packages for connecting to the Internet or emphasizing the commands needed for electronic mail, FTP, and telnet. Class in 025.04 interdisciplinary works about the Internet that do not contain enough computer science material to be classified in 004.678, but do contain some information science material. Class in 025.04 information science works that emphasize search and retrieval, including use of front-end systems and interfaces such as Gopher and Netscape® to facilitate search and retrieval on the Internet. Also class in 025.04 works that describe information resources on the Internet. Class in 384.33 works on Internet access providers and works on economic and public policy issues concerning the Internet. . .

This Manual note makes explicit what is summarized in the schedules and Relative Index. In the schedules, there are references to the Manual from each of the three numbers in the entry for the Manual note. At 004.6 Interfacing and communications (a broader number in the hierarchy for 004.678), 004.6 is identified as the interdisciplinary numbers for computer communications, with references leading to 005.7 and 384.3. The Relative Index also summarizes the relationship among the numbers. It includes two entries for the Internet:

Internet 004.678
see also Computer communications
see Manual at 004.678 vs. 025.04, 384.33

Internet resources 025.04

The index entry for Computer communications, to which there is a see-also reference from the index entry for Internet, includes subentries for communications services (384.33) and programs (005.713).

5. Terminology

Terminology throughout the Classification has been updated to achieve currency, ensure sensitivity, and reflect international usage. Many of the changes in terminology have resulted from a concerted effort to provide descriptive representations in the schedules of persons, social groups, national groups, etc., using the terminology preferred by the group and understood by an international audience. To this end, some of the transformations in terminology that have taken place in the schedules and tables include:

Gypsies >> Romany people
Handicapped children >> Children with disabilities
Old persons >> Older persons
Sick and infirm >> Persons with illnesses and disabilities

Since the Classification is used by different kinds of libraries throughout the world, Edition 21 includes several changes to address international or special needs. For example, the Relative Index includes numerous entries for English-language terminology in use outside of the United
States (e.g., A level examination, School leavers). The Manual note for specific levels of education (372.24 and 373.23) explains the dividing lines between elementary and secondary education using the grade in school and four sample patterns. The aforementioned optional arrangement for books of Tanakh provides alternate notation for libraries preferring arrangement of the Old Testament as found in the Jewish Bible.

6. Structural Changes

Edition 21 also includes several structural changes. In recent editions, we have made a conscious move away from enumeration in the schedules. Where possible, special developments for standard subdivision concepts have been replaced by regular use of standard subdivisions found in Table 1. For example, the development at 370.7 Education, research, related topics now matches the development of --07 in Table 1. The special development for geographic distribution of temperature at the earth's surface has been moved from 551.5252 to regular use of --09 at 551.525.

In Edition 20, the development for 780 Music was highly faceted; 351 Public administration and 570 Life sciences continue this trend in Edition 21. The earlier example on rural administration of irrigation projects in Australia highlighted notational synthesis and the use of facet indicators in 351 Public administration. The 570 schedule also features notational synthesis and facet indicators. For example:

Cytopathology in Rattus: 571.93619352
571.93 Generalities of diseases
6 Cytopathology (from 571.6)
1 Facet indicator for Animals (from 571.1 Animals, in accordance with instructions under 571-572)
9352 Rattus (from 599.352)

Why is notational synthesis important? First, it makes Dewey more hospitable to the specification of emerging topics. Second, it promotes improved information retrieval. The use of facet indicators to identify meaningful components in a number and the use of uniform notation to express recurring aspects of topics within a schedule expand retrieval possibilities by providing access to information represented by parts of a number. Liu (1993) demonstrated the feasibility of "decomposing" Dewey numbers into their component parts; research remains to be done to explore the use of Dewey facets in information retrieval.

We have made a conscious effort in Edition 21 to reduce adjectival headings, prepositional phrases, and many vague headings. Why is this change important? Within the print edition, it is relatively easy to glance at the page header or up the hierarchy on a page to put a heading in context. When schedules are used in an electronic environment, the individual records have no context unless viewed in separate page displays or hierarchical displays (see Beall, 1996 for a description of displays for viewing hierarchies in Dewey for Windows). During the last two years, we have seen an increasing use of the Classification in other venues, e.g., the DDC summaries have been used by several sites to organize World-Wide Web resources. Such uses often employ the Dewey summaries without the accompanying note structure, and include captions with obscure meanings (e.g., 000 Generalities). I will return to this topic in section 8.

Other structural changes include simplification of the note structure and expansion of the Relative Index. The note structure in the schedules has been simplified by replacing "example" and "contains" notes with an existing note type, the "including" note. In addition, "standard-
subdivisions—are-added notes" have been introduced in many entries with multiterm captions to make clear which topics approximate the whole of the concept represented by the number. For example:

542.6 Filtering and dialysis
Standard subdivisions are added for either or both topics in heading

The Relative Index includes more terms and entries for selected built (synthesized) numbers, terms to provide entry vocabulary for international users, and index entries for Manual notes. We have continued the identification of interdisciplinary numbers in the schedules and Relative Index.

7. The DDC Database

Many of the structural changes in the Classification are motivated by present and future uses of the underlying database. The Dewey Decimal Classification database currently serves as the basis for the standard English-language print and electronic editions, translations, and various research projects. The electronic database of the DDC was first created by Inforonics in 1984 from the phototypesetting tapes used to produce Edition 19. Edition 20 was the first edition prepared using the Editorial Support System (ESS), a UNIX-based system with DDC-specific fields developed by Inforonics under contract with Forest Press. Before work began on Edition 21, several fields in the ESS were modified or added to accommodate elements in the USMARC Format for Classification Data (Library of Congress, 1991; for a full description of changes, see Beall, 1992). The motivation for the modifications was to permit the eventual import and export of data in the MARC format while retaining the DDC-specific record structure in the ESS.

One of the new fields in the Edition 21 database is the 685 history field. The 685 history field has been used in the Edition 21 database to document the history of many expanded, relocated, and discontinued numbers from Editions 19 to 20, and Editions 20 to 21. This information could be used in online systems to guide users to relevant information split between former and current numbers.

Another change in the database is the incorporation of selected Library of Congress subject headings. Library of Congress subject headings have always been a source for terminology in the Relative Index, but their inclusion depends upon literary warrant for the concept and compatibility with the rules for Relative Index entries. Electronic Dewey (1994), a DOS version of Edition 20, features frequently used LC subject headings derived from records in the OCLC Online Union Catalog (OLUC) and mapped to Dewey numbers. Dewey for Windows (1996) also includes statistically mapped LC subject headings. However, such mappings are not available for the major revisions in Edition 21. In preparation for the first Dewey for Windows database, selected LC subject headings have been entered into the DDC database. Experimental coding has been developed to describe the relationship between the heading and the number to which it is mapped:

(1) This heading points to this number exclusively
In most cases (~ 90% of the time), this subject heading will map to this number alone
Example: Education, Humanistic to 370.112 Humanistic (Liberal) education
This information may be incorporated later into a classifier's assistance tool for *Dewey for Windows* and other products.

8. Future Directions

We are moving towards a model of the Classification that views the DDC database as the central resource from which products may be derived and/or developed. The print edition is an example of a derived product; it includes much, but not all, of the information in the DDC database. The print version of the Relative Index is a subset without the additional terminology and mapped terms found in the electronic version. *Dewey for Windows* is an example of a product derived from the database and then further developed. Can the DDC database be used to develop representations of the Dewey structure in ways other than those found in the current print and electronic editions? Cochrane and Johnson (1996) suggest one approach using amplified captions without DDC notation. Vizine-Goetz (1996) is experimenting with revamped DDC summaries to serve as a browser for a database of Internet resources.

Over a decade ago, Svenonius (1983) highlighted the potential of classification as a switching language. Edition 20 has been translated into Italian, Spanish, and Turkish. An intermediate French edition based on Abridged Edition 12 with selected additions from Edition 20 has also been published. The first three translations are very close in structure and content to the English-language standard edition with minor cultural adaptations. A translation of Edition 21 by the Russian National Public Library for Science and Technology is already underway. As a first step towards Svenonius's vision, OCLC Research is developing a prototype multilingual browser based on the summaries of the English, French, Spanish, and Russian editions (Vizine-Goetz and Mitchell, 1996).

What is the core of Dewey? It is the hierarchical structure and defined relationships. The hierarchical structure gives context to any topic; the defined relationships allow one to move within and across disciplines. The derived products, new representations, and linkages do not change the core of Dewey; instead, they enable us to address users' needs in a variety of ways and complement our efforts within the database to keep pace with knowledge.

Notes

1. The Dewey home page (http://www.oclc.org/fp/) has links to Internet resources organized by DDC. Some examples include:
   "CyberDewey" (http://ivory.lm.com/~mundie/DDHC/DDH.html)
   "Gnosis" (http://www.slac.stanford.edu/~clancey/dewey.html)
   "The Morton Grove Online WEBrary"
   (http://www.nslsilus.org/mgkhome/orrs/webrary.html)
References
Revision and Stability in Dewey 21: The Life Sciences Catch Up

Abstract: The Dewey Decimal Classification (DDC) has faced the problem of innovation in knowledge organization since it first appeared 120 years ago. During its first 75 years, the DDC took two different approaches that in retrospect proved to be wrong ones. In the last 45 years, however, the system has developed an approach to innovation that can serve as a model of how knowledge organization should be revised. This approach is illustrated by a review of the three major revisions appearing this year in Edition 21. The specific problems found in revising the life sciences in 560-590 are reviewed in detail.

1. General Issues of Innovation versus Stability

When it comes to facing the issues of innovation versus stability in knowledge organization, the Dewey Decimal Classification has been on three sides of the question. That is, Dewey has blindly pursued stability at the sacrifice of all innovation other than straightforward expansion, Dewey has tried innovation in a manner that got it all wrong, and, judging by its present viability as the most widely used universal system for classifying knowledge, Dewey finally found the right approach to innovation.

Managing innovation of knowledge organization involves two simple principles. First we must organize according to enduring principles, then we must keep the general outlines and most of the landmarks of the system recognizable over time. If we do not change at all, we fail in the competition for new customers, and eventually lose our old ones. If we change too fast, we risk losing old customers in an effort to win new ones, and we may fail to win new ones if we look too reckless in our innovations.

Comaromi (1976) provides historical perspective on change in the DDC. When the first edition of Melvil Dewey’s classification appeared in 1876, it proved to be an instant success. It was widely adopted by libraries because it introduced two powerful enduring principles: shelf arrangement determined by simple numeric symbols representing subjects, and subjects arranged according to disciplines. That is, ordinary gardening is found in 635 under agriculture, but landscape gardening is found in 711 under fine arts. The original disciplines have endured, even if their contents have changed radically.

The second edition of the DDC published in 1885 was considerably revised as well as enlarged, and was adopted by an increasing number of libraries. Soon thereafter, however, Dewey became opposed to further changes in the meaning of numbers. His opposition was canonized under the principle of "integrity of numbers." Expansion was allowed, but not adjustment in the meaning of numbers. Use of the system grew rapidly for decades, but it became increasingly dated and it started losing ground to other classifications. The anti-innovation policy continued through 1942 when Edition 14 appeared.

During the 1940s the pressure for innovation became overwhelming. But when Edition 15 appeared in 1951, it simply changed too much. There was a major loss of old friends to the more stable Library of Congress classification. In reaction, however, people learned to handle...
innovation better. Under the able leadership of editor Benjamin Custer, the DDC was soon back on track.

Provisions of editions 14 and 15 were carefully reviewed. Changes that met the approval of the library profession were retained; the ones that required excessive reclassification or eliminated needed specificity were reversed in Edition 16, which appeared in 1958. Thereafter, particular attention was given to disciplines that were obsolete or unsatisfactory. One or two disciplines were selected for complete revision in each edition: psychology in Edition 17 (1965), law and mathematics in Edition 18 (1971), sociology in Edition 19 (1979), and computer science and music in Edition 20 (1989).

2. Lessons from the First 75 Years

Three important lessons may be drawn from the Dewey experience. First, no universal system for classifying knowledge can last even a generation without regular adjustments to accommodate the growing and changing body of knowledge produced by modern civilization.

Second, as soon as the classification becomes established in the hearts, minds, and institutions of people not controlled by its creators, the creators inevitably find themselves sharing control with the users. The first thing shared is the burden of innovation, because the system that is widely adopted has a dual manifestation: the body of the canon, and the body of materials classified under the system. The more the system gratifies the hopes of its developers by being widely adopted, the more chilling the effect of success upon innovation. Fortunately, Dewey has survived the chill, and has learned to innovate.

Third, successful innovation must address the problem of reclassification required in major collections of knowledge using the system. The problem remains even though many Dewey libraries avoid the worst of it. For example, well over half of these libraries have small collections that are in constant flux. New material is constantly added, old material is gradually worn out or is discarded when it becomes dated. Current national bibliographies and publishers catalogs arranged by the DDC also benefit without paying a price; they can immediately use the revised schedules in forthcoming issues. The material in the old numbers remains harmlessly in back issues. Electronic resources like the Internet and the World Wide Web, are also in constant flux; the sources tagged with Dewey numbers will probably change faster than the numbers do.

Nevertheless, the problem will not go away. Consider the large academic and research libraries that cannot avoid the problem created by old numbers with new meanings. These libraries keep the old material for research and archival purposes, but the bulk is so large in proportion to use that they do not even consider reclassifying it. Their needs and practices must be respected, even if we cannot stop the process of innovation out of deference to their needs.

Therefore, innovation must be measured and focused. A certain amount of pruning and spot adjustments to eliminate minor mistakes in development or obsolescence of distinctions among classes can be done on a continuous basis. Any drastic revision, however, should target the most obsolete or poorly developed disciplines, and must limit the change at any given time to what large and busy institutions using the system can accommodate at a reasonable pace. Then, even institutions that cannot reclassify will be able to understand both the immediate benefits of the occasional major revision for many users now, and the ultimate need of the knowledge professions for a classification that remains current. Over time, they may find some way to accommodate the innovation.
3. How Things Go Wrong

We speak glibly about obsolescence of knowledge, but more often the problem is that knowledge was not yet being explored at the time the system was developed. The literature available then often does not represent what appears later. Just a little bit of incautious development to fit the then current works creates a distorted frame of reference for future literature. Expansion built upon the faulty foundation compounds the problem instead of relieving it.

The three disciplines revised in Edition 21 of the DDC illustrate three different degrees of distortion compounded over time. In none has there been a real obsolescence of the material for which the original scheme was designed. Yet bad guesses by the original developers about how the literature should be organized have become increasingly apparent over time. The situation was worst in public administration, and most bearable (but still difficult) in education. It was intermediate in the life sciences.

The DDC devised three quite different remedies for the three situations. Let us review the first two briefly, because what we did is straightforward. Then let us explore the interesting juggling done for the life sciences.

4. The Revision of Education (370)

In the early 80s we considered a complete revision of the education schedule in 370. But there was neither consensus for any single principle of reorganization nor a willingness to accept wholesale changes if the principle was not changed. We dropped the idea of radical change.

Now, over a decade later, we bring out a revision that leaves the main outlines of the schedule basically the same, but reworks numerous subdivisions where the terminology was confusing. We also removed many irregularities that had crept in over time, especially ones involving our standard notation for concepts like educating teachers or using apparatus and equipment. Both the terminology problems and the irregularities invited errors in applying the schedule, and made the overall structure difficult to grasp.

In addition, we relocated several important topics to more predictable locations. For example, education of women and ethnic minorities now joins education of other kinds of students, religious schools are now next to other kinds of schools, and educational sociology joins sociology of other disciplines in 306.

All the changes taken together are not drastic. We list education as one of three major revisions of Edition 21, but we have mainly pruned the twigs rather than cut the branches. In fact, only six numbers are reused, and these are for minor topics. We warn users to expect big changes, but the changes will not require a dismaying burden of reclassification.

5. Revision of Public Administration (350-354)

In contrast to the old education schedule, the old public administration schedule in 350-354 had fundamental problems that could not be removed by small adjustments or half measures. The problems were a veritable catalog of what can go wrong when development is not based on an informed understanding of the needs of the profession using the discipline:

The old schedule required a meaningless distinction between works on central government administration of any given topic (classed in subdivisions of 351) and comprehensive works on public administration of the same topics (classed in corresponding subdivisions of 350). For administration of these same topics in specific jurisdictions, one had to start with different numbers, and follow number-building instructions that involved use of single, double, and triple zeros for different kinds of topics. The instructions, however, required a change in the number
of zeros in each case. The number building for central administration gave preference to country over topic, while most libraries wanted emphasis on topic. For local government, there was an arrangement of the topics in 352 quite different from the one used for central governments. In 352 preference was reversed to topic over jurisdiction. The number fell between the ones used for works on central governments in general (351) and works on specific central governments (353-354). Development overall was very unbalanced; many important topics required long numbers, while many minor ones had short numbers.

The whole schedule was maddening in its complexity and irregularity. One could hardly have a better case for a complete revision. Such a revision is offered under 351-354 in Edition 21. Not a single number in the new 351-354 has the same meaning that it had in Edition 20.

6. Innovation With Caution in the Life Sciences

The problems in the life sciences suggested a more cautious approach to innovation. These sciences were (and are) spread over four units:

560 Paleontology
570 Life sciences (the comprehensive number)
580 Botany
590 Zoology

There were two central problems: First, too much emphasis was given to specific kinds of organisms, and too little was given to biological processes that are the focus of most current research. Second, there was a terrible imbalance in the use of notation, typified by the fact that 80 percent of the material in 570 was classed in two subdivisions, much of it in numbers six or more digits long, while most three and four digit numbers were used for minor topics or were unused.

Even in the other three units there was little material in many subdivisions, especially in paleontology and botany. Other numbers were inordinately long without giving the necessary specificity, e.g., for popular mammals with heavy literature in 599. The scheme for dicotyledons in 583 was a mishmash of two obsolete classification systems. Microorganisms were scattered in three different places, and usually badly developed, while the fish schedule was very inadequate. There were many irregularities in the notation used to specify complex topics like behavior of birds, or ecology of insects.

The problems define the case for a measured and focused revision: Provide a complete revision only in 570, where biological processes can then be given preference over kind of organism, and in 583 where dicotyledons need a new arrangement under a currently recognized system. Bring microorganisms together somewhere with an up-to-date taxonomy, give mammals and fish the kind of revision that will provide greater specificity and shorter numbers, and clear out the irregularities in building numbers for complex subjects. Wherever possible, place new developments in numbers that either were vacant or were used for minor subjects, and avoid reusing old numbers that contain large bodies of material.

Edition 21 offers such a measured and focused revision of the life sciences. The story is a long one, but it shows the importance we attach to wide-ranging consultations, even when they extend over decades.
7. The First Round of Proposed Revision of 560-590

The story began with a series of memoranda from Benjamin Custer to the Decimal Classification Editorial Policy Committee (EPC) in 1972 (the year after Edition 18 appeared) concerning revision in the last of his four editions. In "Extent of revision in Edition 19" (EPC Exhibit 67-34, 23 August 1972), he evaluated nine possible major revisions, including three that were parts of the life sciences. Each proposal received one of four priorities, in descending order: urgent, important, desirable, and "not given priority." A phoenix of taxonomic botany and paleobotany (582-589 and 561) was important, an extensive revision of taxonomic zoology and paleozoology was desirable, and a substantial revision of general biology (570) was not given priority.

At that time, Mr. Custer was both editor of the DDC and chief of the Decimal Classification Division (the Division) which applies the Classification to bibliographic records produced by the Library of Congress. I had been a science and technology classifier for five years, and received a copy. I responded immediately, because I had become quite exasperated trying to apply the preference order of organism over process. Scientists publishing research in the biological processes always played up the processes, and gave so little emphasis to the research organisms that it was often difficult to determine what kinds of organisms they had been working with. In fact, I had become convinced that the emphasis on organism over biological processes was all wrong.

Mr. Custer was very open-minded when he received a memorandum the next day urging that, contrary to his recommendation, priority for complete revision should be given to 570, including a reversal of preference between organism and biological process, while taxonomic revision was much less important. He asked me to write a think piece elaborating upon my views, and extracted the substance of them in EPC Exhibit 67-37 on 24 September. EPC responded favorably, and by November 1972 I was drafting a revision for biological processes under the supervision of Margaret Warren, the assistant editor. The result of our work was submitted as EPC Exhibit 70-21 on 4 April 1974.

That draft was rejected largely on the advice of consultants from the United Kingdom led by Marjorie Jelinek, and work on biology revision stopped for Edition 19, yielding to a new sociology schedule. I was not sorry to see my first major editorial project adjourned until another day, because I had been overruled on one vital point. Mr. Custer and Ms. Warren had accepted the reversal of preference only to the kingdom level. In the 1974 draft, zoological processes were in 591, and botanical processes in 581.

8. The Second Round Fails to Make Edition 20

Work was resumed for the Edition 20 cycle. In 1981 Forest Press contracted with a group led by Marjorie Jelinek based in what became Lancashire Polytechnic to produce a draft representing the views of consultants in the U.K. It was submitted in May 1984. Meanwhile, I had gone back to work on the Division's draft. The new editor, John Comaromi, allowed me to bring biological processes of plants and animals into 570, and to develop several new ideas. The Division draft of April 1983 thus differed substantially from the 1974 draft.

There was a meeting in Washington, D.C., in July 1984 of most members of the Lancashire group with Division staff to attempt a reconciliation of the two drafts. The differences were too fundamental to be compromised. Two new drafts were produced, but two other major revisions were already moving smoothly toward inclusion in Edition 20, and the life sciences were held over for Edition 21.

The two issues that separated the drafts are ones that inevitably come up in any revision process. The first is extent of revision: does every subject need a new home, or only some
subjects? The second is preference order, e.g., should physiology of mammals be classed under physiology or under mammals? The second usually leads to side issues concerning exceptions to the order proposed because of the nature of the subject. For example, may we class physiology of mammals under physiology, but behavior of mammals under mammals? Biology abounds in issues that tempt different people to propose different exceptions.

9. Comparison of the Rival 1985 Drafts

Let us compare the drafts on the points just covered, the Lancashire draft of July 1985 and the Division draft of September 1985:

The Lancashire draft was a complete revision. Its most striking feature was that it relocated plants and animals from 580 and 590 to a single sequence in 570. This relocation enabled the draft to achieve a sequence of numbers fully consistent with its preference order. The relocation was simultaneously the draft's greatest strength because it allowed complete intellectual consistency, and its greatest weakness because it required reclassification even in sections where the current arrangement could hardly be faulted.

The draft provided that the digits following 57 for a specific group of organisms could be appended to notation 1 under any topic for works on that topic in relation to the group of organisms. For example, mammals were 579; therefore the notation for mammals in relation to a specific topic was 19. The number for ecology was 581. Combine the two numbers and we have 581.19 for ecology of mammals. This kind of number building was widely used in both drafts, and, in fact, is characteristic of the DDC. Its simplicity was particularly useful in the Lancashire draft because, with trivial exceptions, every topic in biology took preference over kind of organism.

In contrast, the Division draft proposed nothing in the taxonomic schedules beyond the scope of the measured and focused revision proposed above. It had a jerry-built substitute for notation 1 that could be added to specific topic to show that topic in relation to specific organisms. But more importantly, it had several exceptions to the reversal of preference order. For example, both ecology and behavior of mammals would stay under mammals.

The underlying principle was: if you can make a good case for the present location and development of a subject, do not relocate or redevelop it. For example, it does not make much sense to scatter the biochemistry of different organisms throughout the taxonomic schedule, but many biologists think that it does makes good sense to scatter behavior of different kinds of animals among the kinds of animals. Therefore, the preference order was reversed for biochemistry and other internal processes, but not for behavior, ecology, and description of external features such as flowers, leaves, horns, etc.

The Lancashire draft had an exception of another kind to accommodate the differences among the kingdoms, that is, plants, animals, and microorganisms. It devoted 580 to subjects where the processes in the kingdoms were held to be essentially similar. These consisted of ecology, behavior, genetics, evolution, cell biology, biophysics and biochemistry. In these topics, notation 1 was used for the full range of organisms.

Lancashire devoted 590 to physiological processes and systems in which the kingdoms were regarded as essentially different. There were four pairs of subdivisions, one for general and comparative physiology (591-592), and one each for the three kingdoms, e.g., physiology of microorganisms 593-594. In the first of these, there was no need for the notation for specific kind of organism. In the last three, one went first to kingdom, then to topic, and under each topic one used notation 1 for specific members of the kingdom.
In the Division draft the kingdoms were handled quite differently, largely as a result of reconceiving the problem. The radical differences in the biological processes of the kingdoms are an important feature in biology, but the inescapable differences pertain only to organs and tissues and physiological systems composed of them, e.g., circulatory and reproductive systems. Microorganisms do not have such organs, tissues, and systems. Their biology is entirely explained in terms of basic biological processes common to all organisms.

While there are some important differences between what is true of the physiology of all organisms and what is true only of animals, authors of specific books seldom labor the point. Except when we get down to specific physiological systems (circulation, reproduction, etc.) the uniqueness of animals does not require an exception to the preference order of process first, then organism. The base number is used for general works on the process; then the process in animals is distinguished by notation like that described above. But since the physiological systems do differ from what is found in the other kingdoms, they need unique numbers. In the 1985 draft as in Edition 21, the development for animal systems immediately follows the numbers for processes common to all organisms.

There are two ways to handle the plant kingdom. What ultimately prevailed in Edition 21 is treating plants the same way as animals. For all general topics not unique to plants, including biophysics, cell biology, and anatomy, notation for plants is appended to the general number for the topic. For the specific parts and systems that are distinctive in plants, there is a special development following that for physiological systems in animals.

The literature for internal processes in plants, however, is more distinctive from works on general biology than is the literature for animals. Authors never blur the distinction between the general biology of a process and the process in plants. Furthermore, the tissues and cells of the higher plants are unique; therefore, the biophysics and pathology of plants is often distinctive. Botanists have a much better case for a separate schedule for their kingdom than zoologists and microbiologists. In any event, a separate development for plants was built into the 1985 Division draft.

10. The Differences Are Resolved for Edition 21

As mentioned above, the two drafts were carried over for consideration as part of Edition 21. The Subject Analysis Committee (SAC) of the Cataloging and Classification Section of the Association for Library Collections and Technical Services in the American Library Association was asked to appoint a review committee for the life sciences, as it has done for all our recent important or difficult revisions.

The SAC Subcommittee to Review the DDC Life Sciences Schedule was appointed in May 1992. It was sent a thick packet of material including both 1985 drafts with related commentary and supporting information, such as literary warrant counts showing where material had been classed in the old schedule. The counts were particularly useful in identifying topics that needed expansion or shorter numbers, and heavily used numbers that might best not be reused if there were a complete revision.

The first full meeting was at the ALA midwinter meeting in January 1993. There it decided to give preference to the intent of the author, that is, to the focus of the book. When in doubt about the primary focus, one should use the following order:

1) Physiology [in a broad sense defined later]
2) Organism
3) Other biological process or topic
General physiology and physiology of plants and animals should be brought together in a single block of numbers in 570. Works that focus on the physiological features peculiar to plants or animals, however, should be kept separate from general physiology. On the last point, it rejected the Division draft on plants: rather than a separate unit on physiology of plants, only the works focusing on physiological features peculiar to plants should be kept separate.

The committee decisively rejected the relocation of plants and animals to 570 as being disruptive in the extreme, and likely to result in the rejection of the new schedule in most, if not all, American libraries. It also accepted a suggestion from the Division that 574 not be reused because it contained over 70% of all material classed in 570. All other numbers in 570 except 575 (evolution and genetics which contained about nine percent of the material), however, should be reused for a new schedule.

There was unanimity on bringing microorganisms together, but the question of where was left to the following meeting. The final decision in January was to reject a recommendation embodied in both the Division and the Lancashire drafts to relocate human anatomy and physiology from 611-612 in the medical schedule to the new physiology development in 570.

At its June 1993 meeting the committee agreed upon a tentative arrangement of subdivisions of 570 that largely determined the pattern of Edition 21. The most significant decision was to use 579 for a completely new development of microorganisms. It also fleshed out the list of specific topics in which preference should be given to process: physiology, anatomy, biophysics, biochemistry, cytology, and pathology.

The corresponding committee in the U.K. (The Library Association's Dewey Decimal Classification Committee) would have preferred to keep more of the Lancashire draft, but understood and accepted the SAC subcommittee's recommendations.

In June the SAC subcommittee had reiterated the recommendation to keep evolution and genetics in 575, but the Division later decided that it had to use 575 for specific parts and physiological systems in plants. By doing so, it kept all topics in which preference was given to organism together in one block.

The final arrangement came out like this:

560 Paleontology
570 Life sciences
571-575 Internal biological processes and structures
576-578 General and external biological phenomena
579-590 Natural history of specific kinds of organisms

"Natural history" is taken as symbolic of the topics in which preference would remain with organism, that is, the topics found in 576-578, including evolution and ecology. The specific organisms were subarranged as follows:

579 Microorganisms, fungi, algae
580 Plants
580 Animals

These three numbers are used for comprehensive works on microbiology, botany, and zoology, respectively.

The committee also agreed to the regular use of notation 1 throughout the taxonomic schedule (579-590) for general topics relating to specific kinds of organisms, although notation 04 was most commonly used in Edition 20. As a result, notation 13 is the symbol for evolution
and genetics of any organism, e.g., evolution and genetics of microorganisms 579.13; 17 is the symbol for ecology, e.g., ecology of plants 581.7; and 15 is the symbol for behavior, e.g., behavior of animals 591.5.

For specific taxonomic groups, it agreed that a complete revision of dicotyledons in 583 was needed, that the schedule for mammals in 599 should be revised to provide shorter and more specific numbers for the kinds that had heavy literature. In addition it made a recommendation not broached by the Division, to revise numbers in 597 for fishes and amphibians. This recommendation involved reducing the development for amphibians from three subdivisions to one (597.6-8 down to 597.8), and enlarging the development of fishes from four subdivisions to six (597.2-.5 up to 597.2-.7).

The committee accepted the principle of keeping reuse of numbers to a minimum. Thus, even though adopting a revision without reclassifying material in old numbers leads to illogical arrangement, doing so will result in limited mixing of old and new material in the same subdivisions.

Two of the relocations mentioned above illustrate the point. Both the old comprehensive number for amphibians (597.6) and the number for evolution and genetics (575) were reused. They were reused, however, as hook numbers, that is, the numbers themselves were not used, only the subdivisions. The subdivisions of these numbers that had previously been used for more than a handful of books did not need to be reused. The number 597.6 was reused for three superorders of fishes that are never written about collectively; the literature is on the specific orders and their families. The only subdivision previously used (597.65 for newts) could be left vacant because it was not needed for these orders.

Analysis of previously used subdivisions of 575 is even more revealing, because they had previously contained about nine percent of the total material in 570, the largest block outside 574. But the material was almost all concentrated in 575-575.2, numbers that were not needed in the new development for specific parts of and physiological systems in plants. Thus, in spite of the nominal reuse of 575, over 80% of the 570 old material in most large research collections will probably be found in a single short span, 574-575.2 that has no meaning in the new schedule. Thus, even in completely revised 570, reclassification can be limited to numbers containing less than 20% of the old material.

II. Summary

Whether we consider the three major revisions in Edition 21 as a whole, or focus only on the life sciences revision, we have a model of how revision in a widely used universal classification system should be carried out. In some cases (e.g., public administration), the faults in the old structure are so basic that the entire schedule must change. In other cases (e.g., education), judicious relocations and expansions may address the problems. Finally, in extensive schedules such as the life sciences, we can combine both approaches. We must keep the value of the enduring features of the schedule in mind and focus our attention on the parts that really need to change. We must consult our users, and heed their advice.

References
Dewey for Windows

Abstract: This paper describes how the features of Dewey for Windows (DFW) facilitate use of the Dewey Decimal Classification in CD-ROM form and suggests future enhancements to make DFW a more efficient tool. The strength of Dewey for Windows lies in the many different approaches that it provides to the classifier.

1. Introduction

Dewey for Windows provides a Windows interface for the Dewey Decimal Classification. The version described here is the first field test version. This version uses the same CD-ROM database as the earlier DOS version of Electronic Dewey, a CD-ROM that holds the 20th edition of the Dewey Decimal Classification (DDC 20). When Dewey for Windows goes into production, its CD-ROM will hold DDC 21. While the user interface of DFW is very different from that of the DOS version, the search engine (database access program) is similar. See Carpenter (1995) and Trotter (1995) for descriptions of the DOS version. This paper does not attempt to cover all the features of Dewey for Windows; it focuses on those features that seem most useful to the working classifier. See Vizine-Goetz and Bendig (1996) for a more comprehensive description of DFW.

2. Searching

Dewey for Windows offers multiple ways to search for a particular topic, but they all involve use of the Index window or the Search window. The most useful approach with the Index window is the Relative Index (Phrase) scan. With this approach, one scans and selects from a display of the Dewey Relative Index that is similar to the printed index. This approach is particularly useful for terms that appear in many different places in DDC, when the classifier wants to be guided to the main numbers for the topic, e.g., business, literature, programming, taxes. The built-in Scan view has a half-screen Index window (tiled vertically), a quarter-screen Search window, and a quarter-screen DDC Number window (see figure 1).

In DDC 21 additional nonprint Relative Index entries are being added to the CD-ROM version of DDC, entries that cannot fit into the crowded print index volume. The Dewey editors plan to add many more nonprint index entries over time.

The Search window has an input box used primarily for initiating keyword searches for terms or Dewey numbers, and it lists in DDC order the Dewey numbers and captions of the hits that result from searches. The Search view has a half-screen Search window (tiled vertically); that window is useful for viewing long hit lists. With the Search window one can do keyword searches on every word and number in the DDC database (except a very short list of stopwords), using Boolean combinations and truncation and mask symbols (asterisk [*] for zero or an unspecified number of characters at the end of a term, question mark [?] for zero or one character anywhere in a term). One can use index labels to limit a search. The most useful index label is dd: as in a search for dd:616.1-616.9; it

Fig. 1: Scan view

limits retrieval to the Dewey number as the entry number (or one of the entry numbers) in the record. A search for 616.1-616.9 without the index label retrieves every record with that span anywhere in any note, including the frequently appearing add note “Add as instructed under 616.1-616.9.”

Keyword searches are very efficient for distinctive terms that do not appear in many places in the DDC, e.g., batik. One can of course also scan for such terms in the Relative Index (Phrases), but unless one already has an Index window on the screen for another purpose, it is not efficient to use the Index window when a keyword search will suffice. Keyword searches are useful if one does not know the Dewey Relative Index well enough to predict the order in which terms will be presented. For example, there is only one Relative index entry under World War II, for the main number 940.53. To index all the relevant records under World War II would violate the Dewey editorial rule against recapitulating the schedule in the index; however, there are many Relative Index entries elsewhere in the alphabet, with World War II as a subfield, e.g., Refugees—World War II. A keyword search on world war ii and refugee? will retrieve the relevant record 940.53159. Keyword searches are useful as shortcuts for particular combinations that are typically not given in the Relative Index. For example, if one is interested in compensation by industry, one could search for compensation and industr* to retrieve 331.28. If one used the Relative Index approach, one would scan for compensation, look at the full record for 331.21, and follow the reference to 331.28. Keyword searches are also useful if one recalls a particular example in a Manual note or an add note that would enable one to retrieve the desired record, e.g., Seth Thomas clocks given as an example in the Manual note 300 vs. 600 Social sciences vs.
Restrictors are special indexes that allow a search to be limited to a specific main class (one significant digit, e.g., sl:300), division (two significant digits, e.g., s2:330), or section (three significant digits, e.g., s3:331) or to a specific auxiliary table (e.g., tn:11). For example, a search for waste? and s2:340 will retrieve waste(s) only in the 340 Law schedule, a search for germany and tn:2 will retrieve Germany only in Table 2 Geographic Areas. Classifiers who are familiar with a particular part of the DDC schedule often remember the first two or three digits of the schedule number they need or which table they need and are well prepared to make good use of restrictors. Even without that familiarity, classifiers can use the restrictors when they are following add instructions that specify a particular part of the schedule or a particular table. For example, a classifier at 155.413 Conscious mental processes and intelligence [in child psychology], following the instruction to add "the numbers following 153 in 153.1-153.9", can search for, e.g., learning and s3:153; a classifier at 808.819 Poetry displaying specific features, following the instruction to add "notation T3C—1-T3C—3 from Table 3-C", can search for, e.g., marriage and tn:13c.

Using restrictors has the advantage that the search goes much faster than if one used only truncation (waste? and dd:34* or germany and dd:12*), but using restrictors has the disadvantage that relevant Manual entries are not included. If one really needs to find Manual notes as well as table or schedule information, then one may wait for the results of searches like biograph* and dd:36* or biograph* and dd:338*. Both searches retrieve useful advice in the Manual about classifying biographies. Or one may do a simple search for biograph* and scan the 172-record hit list (arranged in DDC order) looking for relevant records. One hopes, however, that the restrictor indexes will be expanded in a future version so that they can be used to retrieve Manual entries. On the other hand, if one wants to exclude Manual notes, which are sometimes intrusive, the current disadvantage of restrictors may be seen as an advantage; there is no other way to exclude Manual notes as a category from a hit list. Ideally, a better way to exclude (or limit to) Manual notes as a category will be provided in a future version of DFW, possibly by means of a Manual index and index label.

Library of Congress Subject Headings have been statistically linked to DDC numbers to provide additional search terms for the Dewey numbers—the five headings most frequently appearing with that Dewey number in bibliographic records produced by the Library of Congress. Special efforts are being made to provide LCSH for completely or extensively revised DDC 21 schedules, where statistical matching with DDC 20 numbers in bibliographic records would not be helpful. Using the Index window, one can scan and select from the LCSH (Phrase) index. The list that one scans is not a list of LCSH authority records, rather just a list of headings stripped from bibliographic records. One can also do a keyword search for the LC Subject Headings associated with a DDC number, using the index label sut. Since the statistical matches for the most part have not been editorially reviewed, there are accidental bad matches; one must always check the full information in a record to be sure that one has found the correct Dewey number. The LCSH are useful as a searching aid when the editors have not indexed a particular sought term or synonym that fits a Dewey number, e.g., Parental leave in DDC 20. These terms will later be added to the Dewey Relative Index, though in some cases only to the nonprint index. The LCSH are also useful for terms that are excluded by policy from the Relative Index. For example, specific battles in a war are not named in the Relative Index unless they are named in the schedule. Thus one can find World War I's Verdun (940.4272) and World War II's Pearl Harbor (940.5426) by name in Dewey for Windows only by searching on the Library of Congress Subject Headings. Of the three hits in the LCSH index for Pearl Harbor, two are obviously bad matches (European theater; East and southeast Asian theater), but one (Pacific Ocean theater) is good. For a classifier.
weak in geography, there is another clue to which match is best: a click on the LCSH button for each record reveals that the bad matches were based on one bibliographic record each, while the good match was based on 27 bibliographic records.

One common reason for people to get no hits in the DDC database is that they are searching for a concept that would be represented by a built number. The pieces needed to build the number are in the database, but the built number is not. Educating classifiers is one answer to that problem, but adding more built numbers to the database is another. There is plenty of room on the CD-ROM, and there are plans for a continuing project of adding additional highly posted built numbers to Dewey for Windows, with their LCSH. Meanwhile, with the local annotation feature of Dewey for Windows, classifiers can add built numbers that they frequently use, with one numerical and one verbal index term searchable in a separate Notes index.

3. Selecting and Identifying Disciplinary and Interdisciplinary Numbers

Dewey for Windows offers several methods of finding a number in the correct discipline for a work and identifying the interdisciplinary number for a topic. With a display of the Relative Index (Phrases) in the Index window, the discipline of a number is shown by the subfields of the index entries, and the interdisciplinary number for a topic can be identified because it appears opposite the term without any disciplinary subfield. For example, the following Relative Index entries show that the social welfare number for pollution is also the interdisciplinary number, 363.73, and offer other numbers for the topic in the disciplines of law and toxicology.

Pollution 363.73
Pollution—international law 341.7623
Pollution—law 344.04632
Pollution—social welfare 363.73
Pollution—toxicology 615.902

If one of these numbers is found by some other approach, the index entries for the record can be seen in the full record display in the DDC Number window; that is another way to identify the discipline or subdiscipline, or to determine whether the number is the interdisciplinary number for the topic. One can also identify the interdisciplinary number for some topics by searching for notes in the schedules and tables, e.g., a search on death and interdisciplinary; however, many interdisciplinary numbers are identified as such only in the Relative Index.

Another way to find the numbers for a topic within a particular discipline or subdiscipline when the searcher knows the first one to three digits that specify the discipline is to use restrictors; for example, one could search for pollution and s2:340 to limit the search to the discipline of law, or pollution and s3:341 to limit the search to international law.

Another way to determine the discipline of a number is to view it in a Hierarchy window, and check its upward hierarchy. The Hierarchy window is described in detail below.

4. Moving Up and Down the Hierarchy

Dewey for Windows facilitates moving up and down the DDC hierarchy in several ways. The Hierarchy window presents the entire upward hierarchy for a number and the next step down in the downward hierarchy. For example, the hierarchy for 154.6 Sleep phenomena is
100 Philosophy, parapsychology and occultism, psychology
150 Psychology
154 Subconscious and altered states and processes
>154.6 Sleep phenomena
154.[.62] Sleep
154.63 Dreams
154.64 Somnambulism

The hierarchy for 612.821 Sleep phenomena is

600 Technology (Applied sciences)
610 Medical sciences Medicine
612 Human physiology
612.8 Nervous functions Sensory functions
612.82 Brain
>612.821 Sleep phenomena

Within a Hierarchy window, the user can highlight any number in the hierarchical display, click on Hierarchy, and see the upward and downward hierarchy for that number. To see coordinate numbers, the user need only highlight the next higher number and click on Hierarchy. This makes it easy to start at any point in a schedule or table, then move quickly and easily up or down as necessary to identify the most appropriate number for a work.

To clarify the difference between two or more apparently similar numbers, a user may wish to compare their hierarchies. It is possible to put several Hierarchy windows displaying hierarchies of different numbers on the screen at the same time. One convenient way to do that is to customize one of the four user views with a quarter-screen Hierarchy window and a quarter-screen Search window both on the right side of the screen, where they will not quickly be covered by additional Hierarchy windows, which appear in a cascade pattern starting in the upper left corner of the screen but can subsequently be dragged to other positions.

A user exploring up and down in the DDC may need to see more information than is found in the concise summaries presented in the Hierarchy windows. To view nearby numbers a user can drag and drop a number into a DDC Pages window to get a display much like that of the printed page. This is a good way to see centered entries, which are excluded from Hierarchy windows except when the record for which hierarchy information is sought is a centered entry. The DDC Pages window does not, however, offer a good way to see numbers that are distant from the starting number; paging up and down through many numbers is slow.

A good way to get information about the downward hierarchy of a number in a Search window hit list is to drag the number to the input box of the Search window. The index label \textit{dd}: will automatically appear to limit the search to instances of that number as the main number. The user can add an asterisk truncation symbol at the end of the number to get all the subdivisions of the number. To get information about coordinate numbers, the user can delete the last digit of a number before adding the asterisk. The resulting hit list will include all subdivisions of a number, not just the next step down; and it will also include centered entries, built numbers from the Relative Index, and Manual notes. Since built numbers from the Relative Index and Manual records are excluded from the DDC Page display, this approach gives the fullest listing of subordinate entries for a number. Since only the number and the caption are shown in the Search window hit list, many entries can be scanned quickly, especially if the Search window is maximized.
If the result of using the asterisk truncation symbol is too many hits for easy viewing even when the Search window has been maximized, a user can substitute for the asterisk two question-mark truncation symbols (each question mark specifying one character), or whatever number of question marks is appropriate. Thus the user can create tailor-made summaries of the downward hierarchy for a number.

One particular reason why classifiers need to move up in the DDC hierarchy is to find citation-order and precedence information. For example, a user is trying to classify a work on processing meat to be used for pet food and has identified two relevant numbers, 664.66 Food for animals and 664.9 Meats and allied foods. With the DDC Pages window, the user can easily browse from 664.66 to the centered entry 664.62-664.66 to learn that the number for meat should be given precedence. That approach often works; but sometimes the precedence information is too far away for simple paging upward to be efficient. For example, a user is trying to classify a work on wages of immigrant workers in the garment-making industry and has identified two relevant numbers: 331.28 Compensation in specific industries and occupations and 331.62 Immigrants and aliens [as workers]. The number in common is 331. The classifier can use the Hierarchy display to get to 331 or can simply search for dd:331 and display the record to find the table of precedence indicating that 331.62 is to be preferred. If neither of these approaches works, the citation-order or precedence information may be hiding in an elusive centered entry. Since every centered entry contains a note specifying where to class comprehensive works, one can search for comprehensive and limit the search to the relevant portion of the schedule; for example, to get help in choosing between 331.52 Veterans [as workers] and 331.69 Indigenous ethnic groups [as workers] by consulting the centered entry 331.3-331.6, one can search for comprehensive and s3; 331. Only some of the hits will be centered entries, but those entries are easily spotted because they are marked with a "C" at the left in the hit list. Since centered entries are often crucial for various reasons, one would like to see the system provide a simpler and more obvious way to find all the relevant centered entries in the upward hierarchy for any given number. An alternative hierarchy display could be offered that would include centered entries. It would sometimes look awkward and cluttered, but it would be very useful.

5. Displaying Schedule, Table, and Manual Entries and Full Records for Built Numbers from the Relative Index

Dewey for Windows offers two basic ways to display schedule and table entries, and one way to display Manual entries and full information about built numbers in the Relative Index. Classifiers will select the display method depending upon their individual preferences and upon what information they need in a particular situation. The DDC Number window displays the fullest information about a single DDC number—the number and caption plus published notes, Relative Index entries for the number, and the LCSH linked to the number. The DDC Pages display shows multiple adjacent schedule and table records much as they appear in the printed DDC, with numbers, captions, and published notes. The user can page upward or downward in the schedule or table. The DDC Number window has a Hierarchy button to activate the Hierarchy window; the DDC Pages window gives no information about upward or downward hierarchy beyond what appears in adjacent entries displayed in the window. The DDC Number window shows slashes that indicate logical places to truncate DDC numbers; the DDC Pages window does not. The DDC Number window displays Manual records; the DDC Pages window does not. The DDC Number window displays built numbers from the Relative Index with their slashes, all the Relative Index terms and LCSH associated with them, plus the nearest upward nonbuilt number and its caption. The DDC Pages window does not display built numbers from the Relative Index;
but a quick way to find the add note used to build a number is to drag the number to a DDC Pages window and drop it.

The DDC Pages window displays local notes in blue before the published notes in black; the DDC Number window does not display local notes as part of the record, but has an icon that can be double-clicked to activate a separate window that displays local notes. Classifiers who make much use of local notes will favor the Pages display for notes that they want to see every time they view a record and that are short enough not to crowd much other information off the screen; they will favor the DDC Number window approach for long Manual-like notes that they need to consult only occasionally.

Both the DDC Pages and the DDC Number windows work well with short and medium-length records; neither works particularly well with long records, especially records with long add tables. The DDC Pages display is more compact than the DDC Number display, and thus offers some advantage for records with long add tables. One wishes a comparable compact display were available for long Manual notes. In the DDC Number display but not the DDC Pages display, the word or number used in the search query is displayed in bold; a classifier can then scan the long record looking for the bolded term. For example, a classifier at 670 Manufacturing, following the reference "See Manual at 338 vs. 060, 381, 382, 670.294, 910, T1—025, T1—0294, T1—0296," might search on 670.294 so that 670.294 would be bolded as an aid to finding relevant portions of the long note. What would really help, though, is a string-searching feature to allow the classifier who has retrieved a long record to search for words or numbers within that record.

A classifier can easily take advantage of both the DDC Number and DDC Pages windows using the Browse view, which has a quarter-screen Search window, a quarter-screen DDC Number window, and a half-screen DDC Pages window (tiled vertically) (see figure 2). The quarter-screen DDC Number window rarely shows everything in the record, but it is easy to scroll down or to maximize the window for a quick check of information. For example, in the lower left corner of figure 2, the record for 341.754 has been scrolled down to show the DDC Index Terms, which identify the number as an international law number. A classifier who strongly favors the DDC Number window will prefer the Search view, which has two windows tiled vertically, Search and DDC Number. With a half-screen, vertically tiled DDC Number window, one rarely needs to scroll down or maximize.

For complicated number building, a classifier may customize one of the four user views to serve as a number-building view. An approach that works well is a tiled view with one Search window and either three DDC Number windows, three DDC Pages windows, or some combination of DDC Number and DDC Pages windows. The key is that the classifier should be able easily to display three records at once. For example (see figure 3), when building a number for a critical study of costume in fiction, a classifier might display the add note at 809.1-.7 in one DDC Pages window (upper right), the add note at 808.839 in another DDC Pages window (lower right), and the Table 3C record —355 in a DDC Number window (lower left). Dewey for Windows provides a Work Area at the bottom of the screen where numbers can be built, then copied into a local annotation note or copied to the Windows clipboard and thence into a bibliographic record. After the first part of a number has been built, the relevant add notes technically no longer need to be on the screen; however, if they go off screen and then the classifier wants to review the number, checking to be sure that it has been built properly, then the Past feature must be used to bring them back.

A classifier can easily use different views in succession while classifying one work. For example, for a work that treats both the public finance and macroeconomic aspects of taxes, a classifier can begin with a scan of the Relative Index (Phrases) (see figure 1 again). A useful strategy is to select one entry from the Relative Index (Phrase) display (e.g., Taxes 336.2,
Fig. 2: Browse view

Fig. 3: Customized view being used to build number 809.39355
which the Relative Index presents as both the interdisciplinary and public finances number for
taxes), drag the entry to the DDC Number window and drop it to see the full record; then select
another entry (e.g., Taxes—macroeconomic policy 339.525) and either click on the Search button
in the Index window or drag the entry to the Search window and drop it there. The classifier then
switches to another of the views that contain a Search window and a DDC Number window (all
the built-in views contain both windows). The information in the Search and DDC Number
windows is carried over to the corresponding windows in the other view, but the Index window
no longer fills half the screen. With the tax example, if the classifier tentatively plans to use 336.2
but would still like to check the full record for 339.525, one approach is to switch to the Browse
view, drag 336.2 into the DDC Pages window to see it in the context of neighboring records, then
drag 339.525 to the DDC Number window to see the full record.

6. Following References

In Dewey for Windows the primary device for following references is drag and drop. In an
Index window with the Relative Index (Phrases) index, one highlights the line with SA (e.g.,
"Melanoma SA Cancer"), drags it to the Scan button, and drops it to go to the referred-to portion
of the Relative Index. In a DDC Number or DDC Pages window, one highlights the referred-to
Dewey number and drags the resulting suitcase icon to the place that will produce the appropriate
response. If the reference is an add note referring to a record with a long add table (e.g., "Add
as instructed under 930-990"), the classifier may want to drop the icon in a DDC Pages window
to get the most compact display of the long add note. For other references to a schedule or table
number, a classifier may display the record in a DDC Number or DDC Pages window, may drag
the number first to a hierarchy window to see its hierarchical context, or may drag the number to
the input box of the Search window and add an asterisk to create a list showing all the number's
subdivisions. In short, the drag-and-drop approach to following references makes many choices
available to the classifier.

The drag-and-drop approach, however, generally requires more effort than clicking on a
hypertext link. When a classifier simply wants to look at the other number with the least effort,
a hypertext link would be preferable. It is hoped that hypertext-like links can be added for those
references where links are most likely to be desired, without obstructing any of the flexibility made
possible by the drag-and-drop approach.

7. Conclusion

The strength of Dewey for Windows is that it offers many different ways to approach the
problem of classifying works. Classifiers can select approaches depending on the nature of the
topic, on their own familiarity or unfamiliarity with a particular part of the schedule, and on their
own individual preferences.

References
Classification Quarterly, 19 (1), 134-137.
Dewey for Windows, version 1 field test, CD-ROM and 5 computer disks. Dublin, OH:
OCLC, 1996.
Vizine-Goetz, Diane, and Mark Bendig. (1996). Dewey for Windows: accessing the Dewey decimal classification from the technical services workstation. To be included in a work tentatively titled *Planning and implementing technical service workstations* that will be published by the American Library Association.
The DDC, The Universe of Knowledge, And The Post-Modern Library

Extended Abstract

Discussions of the origins of the DDC are ultimately not very satisfying because they do not answer the more trenchant question of why Dewey developed this form of access tool in the first place. That question is especially important because now at more than a century's remove from the beginnings of the system many have concluded that the DDC is essentially outmoded, and that little more than this need be said of it. Yet, this conclusion is far too simplistic an answer to a serious question about the nature of the system and its origins. The intent here is to begin a more comprehensive answer to the question of the DDC's origins, first, by summarizing developments in the long history of the DDC; second, by addressing the relationship of the DDC to the nineteenth century classification of knowledge and the sciences movement which formed a significant broader context in which it was created; third, by addressing its relationship to the rise of library classification theory in the present century; and last, by commenting on the position of the DDC in relationship to the new post-modern library which has begun to appear.

1. DDC Developments

The history of the DDC can be usefully divided into three eras: the period of its establishment (Editions 1-6, 1876-1899); a middle period of conflict and struggle (Editions 7-15, 1911-1951), and a more recent period of expansion (Editions 16-21).

1.1 Establishment (Editions 1-6, 1876-1899)

The first six editions of the DDC extend from its founding to its wide acceptance in the library community during the 1890s. Editions one through three (1876, 1885, and 1888) represent the DDC's period of growth into a full-blown library classification system. The comparatively tiny first edition is surprising for it is actually much larger than is at first apparent, possibly larger in extent than any other contemporary system except Brunet's. The first edition also saw the beginnings of what now would be called faceted structures for form, genre, language, and geographic divisions, all of which Dewey called mnemonic aids. The second (1885) and third (1888) editions served chiefly to expand the DDC. Whereas the first edition contained all of the system in a bare 42 pages, the second edition was expanded to 314 pages and the third to 416 pages (although it did not contain the 48 page introduction of the second edition). The chief features of these two editions were:

1. Filling out all of the previously unused second hierarchical level sections of the scheme as well as a number of third level sections, incorporating as notational subdivisions topics previously listed only in the index, and in general expanding all sections of the scheme often to fourth and fifth levels of hierarchy.

2. Expanding the 800s by enumerating literary authors and by placing their names in the index.

3. Revising the order and extent of the general form divisions, supplying a special alphabetical index to form division terms by the 3rd edition, and enumerating form divisions in the schedules at many second hierarchical level locations and even at one third hierarchical level location.

4. Expanding philological subcategories to fifty-five by the 3rd edition, placing them in a special table with instructions as to how they should be synthesized with the notations for individual languages, and supplying a special index of those categories which by the 3rd edition had 159 entries.

5. Supplying a special alphabetical list of languages and their notations to be used in both the 400s (Philology) and 800s (Literature) which, together with the philological subcategories, amounted to a "mini" faceted structure within the system.

6. Expanding geographical "divide-like" provisions from one place in the 1st edition to 112 locations in the third, with a special alphabetical table of those locations in the schedules, the latter including cross-references.

7. Including at least one instance of a global "divide-like"—i.e., "divide like 100-999"—important as a harbinger of the kind of subject collocation which would become possible in the UDC and in the DC in later editions.

Dewey also relocated a significant number of subjects in the second edition and this led him to promise thereafter to observe a policy of "integrity of numbers" in the system by not changing the meanings of numbers unless it was critical to do so.

The fourth through sixth editions of the DDC (1891, 1894, and 1899) were edited by May Seymour and represented a plateau for the system because little new material was added to it during this decade. The decade of the 1890s also saw the system widely adopted in American libraries. This created a new kind of discussion regarding the system, one in which those who had adopted the system criticized it as users of the system, and others evaluated the system on the basis of the logic of its class order or on the basis of its usefulness for scientific and specialist subject areas. Two other significant events for the DDC were Dewey's permission for Paul Otlet and Henri LaFontaine to use and expand the system for the work of the Institut Internationale de Bibliographie (IIB) and his refusal to let the system be used and modified for the Library of Congress.

1.2 Period of Conflict (Editions 7-15, 1911-1951)

The second period in DDC development was marked by conflict related to the purposes and objectives of the system. The participants in the conflict included members of the DDC editorial and administrative staff, librarians of diverse backgrounds, the American Library Association, the Lake Placid Club Educational Foundation, and those responsible for the Classification Decimale Universelle (UDC) of the IIB. The conflict partly turned on personal issues of authority and loyalty, but it also arose from a variety of more general interrelated issues: the size of libraries that the system should serve (medium or large), the subject orientation that the system should serve (general or special), the bibliographical purpose of the system (shelf classification or catalog access), and the degree of theoretical or scientific accuracy in knowledge category arrangement that the system should reflect.

Editions 7 through 11 (1911 to 1921) were edited by May Seymour, although she died before the 11th edition was actually published and her place was taken by Dorcas Fellows. The seventh edition was published in 1911, a full twelve years after the 6th edition. However, if one
counts from the third edition in 1888 when major amounts of new material were last added to the scheme, the seventh edition represents the first time in nearly a quarter of a century that the system was expanded in a major way. Two groups of persons exerted pressure to expand the system—those whose library collections were becoming larger than the typical general library, and others, such as Otlet and LaFontaine of the IIB, who wanted to serve the information access needs of specialists. The latter group not only wanted the system to be expanded, but also to be changed where it was not considered up-to-date or scientific in its arrangement. Both of these groups were in conflict with those who wished for the system to remain relatively small so as better to serve the needs of the moderately sized general library. Further those who used the system primarily for shelf classification did not want it to violate Dewey's policy of integrity of numbers.

Expansions were made through the 12th (1927) and 13th (1932) editions, but in a very uneven way both in extent and timeliness. Some of the expansions were extensive and were made by specialists who occasionally also worked on the UDC. Others came from the DDC staff but these tended to be much more modest in scope. In all such changes, the DDC staff tended to approach expansions without a clear-cut policy in mind. In 1924 Paul Otlet and Godfrey Dewey convinced Melvil Dewey that concordance between the DDC and the UDC was possible and that should it be achieved it could become the basis for creating three related editions of a single decimal system—an Abridged edition for very small libraries, a Standard edition for the shelf classification needs of moderately sized general libraries, and a Bibliographic edition for use in large libraries and for classified catalogs. However, conflict among the participants in this task through 1931 caused the effort to fail.

The 14th and 15th editions of the DDC (1943 and 1951) were driven by the ideal of creating a Standard edition of the system. The 14th edition actually expanded in size over the 13th edition but also included the first efforts to even out levels of specification in various parts of the system. The 15th edition constituted a compromise. The system was radically pruned in size. But, it was also extensively modernized in terms of relocating subjects, in its terminology, and in the way the system was visually arranged. This edition changed the system so fundamentally (in fact, nearly fatally) that for all practical purposes it severed the system from the historic tradition which had accumulated up to that point.

1.3 Period of Expansion (Editions 16-21, 1958-1996)

The DDC has developed in a significantly different direction since the nearly fatal 15th edition. The 16th edition (1958) edited at first by David Haykin but afterward by Benjamin A. Custer, functioned as something of a holding action, restoring some of the system's size. But, it also kept the modernization basic to the 15th edition. Subsequent work on the system under Custer (17th through 19th editions, 1965-1979), John Comaromi (20th edition, 1989) and both Comaromi and Joan Mitchell (21st edition, 1996) has been marked by two general themes. First, some balance has been gained between the two extremes in the conflict of the previous period. For example, relocating and reordering subjects have been accommodated, but for the most part in a very judicious manner. Second, the system has rigorously and increasingly adopted new classification technologies, especially in adopting synthesizing devices and formal faceting structures.

2. The Relationship of the DDC to the Larger Realm of the Classification of Knowledge

One obvious factor in DDC development is that the work of those responsible for it has been related in some way to the broader realm of the classification of knowledge. At first glance
this seems to be entirely understandable because the system, along with other library classification systems of the late nineteenth century, appear to have arisen out of the seedbed of the classification of knowledge and of the sciences movement which had begun with Francis Bacon in the 17th century and had continued through the nineteenth century. That more philosophical movement is given two different names here (the classification of knowledge and the classification of the sciences) as a way to reflect its changing emphasis from the time of Bacon to the 1890s.

While the relationship between the classification of knowledge and of the sciences movement and library classification might seem to be a firm one, evidence exists to suggest that there was no essential connection between the two approaches to classification. Library classificationists simply did not approach the classification of knowledge and of the sciences with the concerns that their more philosophical counterparts had. Classificationists of knowledge and of the sciences were essentially engaged in creating apologias for the role of the sciences in society by enumerating the sciences and portraying their unity as both a cause and a result of the social evolution and progress of human civilization. Library classificationists had a no less exalted task—to make knowledge accessible to a broad range of people—but this task did not require them to argue for the legitimacy of the role of the sciences in society. Library classificationists appear simply to have accepted the legitimacy of what those classifying the sciences were attempting to do and merely proceeded from there to the practical problems of their own work. Their most significant borrowing from the more philosophical movement was to accept without question the utility of the method that the classificationists of knowledge and of the sciences had used to portray knowledge—hierarchical classificatory structures of knowledge categories that proceeded from general to specific, from the theoretical to the practical, and from the abstract to the concrete.

The very practical attitude of library classificationists that began by assuming the legitimacy of the classificationists of knowledge efforts and which borrowed without questioning it the idea of a hierarchical classificatory structure of categories is nowhere better illustrated than in the work of Melvil Dewey. Dewey was the king of practicality among all the modern library founders. He did not pursue practicality merely for some broader social educational purposes, however, but also because of his business sense. The DDC in this context was an enormously successful invention (not unlike creating a successful piece of software for the Internet in our own day) for solving a very real problem—that of organizing library collections of books for easy access. That it was brilliant solution for the purpose for which it was invented goes without saying and in the end it made the man. It was essentially user friendly and practically sold itself. He had developed the original prototype of the system (i.e., its beta version) at Amherst College, and then obtained what amounted to free labor for further research and development of the tool at Columbia College. Thereafter he marketed the system with great skill, his most significant difficulties being how to estimate how many copies to print and how to keep standing inventories down. His sound business orientation regarding it is no better illustrated than in the idea of integrity of numbers. In the latter, he needed a system in which version 10.0 worked with version 8.0, version 6.0 and version 3.0, and so on. Only in that way could he insure the continuity and therefore the marketability of his invention. Looking at the DDC this way helps greatly to explain how the struggle of the middle years of the system developed. Dewey himself appears to have been intrigued with the possibility of expanding the system's influence and use by taking on a wider range of purposes, the latter including serving larger libraries and being adopted in classified catalogs for specialists. Ultimately, however, he balked at following this course of action because to do so would have necessitated redesigning the system, a prospect for which he had neither the energy nor the personnel.
3. The Relationship of the DDC to the Development of Library Classification Theory in the 20th Century

The classification of knowledge and of the sciences movement all but died out after the start of the present century. Classification remains vital as a methodology used by the sciences to organize their objects of study but its use as a motif for arranging knowledge categories in general schema has become lost in the backwaters of contemporary philosophical thought. R. G. A. Dolby surmises that this occurred not only because of the increasing artificiality of the arguments of many in the movement, but also because the sciences no longer needed philosophical explanations to justify their role in society. In the 20th century the sciences have resorted chiefly to showing the potential value of their work to society in order to justify their value and role. In contrast to the dying out of the philosophical classification of knowledge and of the sciences movement, librarianship has not only continued to classify knowledge but has built a sizable literature of classification theory and technique during the present century, not to mention several new general classification schemes. This raises the question of why this is the case?

Two basic factors help to explain the expansion rather than the contraction of an interest in the classification of knowledge in librarianship. To begin with, rapid growth in information production by and information access needs of specialists and scholars beginning in the 1880s and 1890s and continuing unabated ever since has fueled the continuation of an interest in the classification of knowledge in librarianship. However, this factor alone does not account for the development of a theoretical aspect to library classification. Any close look at the work of Otlet and LaFontaine, for example, those who most exemplified an interest in meeting the information access needs of scholars and specialists, or at the work of Samuel Bradford, A. F. C. Pollard, and Fritz Donker Duyvis, the principal persons who worked on the UDC from the 1920s to the 1940s, reveals no serious intent and effort to develop a philosophical or theoretical approach to the classification of knowledge. Here, just as among earlier classificationists, the most important problems were practical in nature.

It was only as a second important factor arose, a shift in the focus of the classification of knowledge from the classification of the sciences per se to the pursuit of scientific classification that a theoretical cast began to take shape in library classification. Approaching the classification of knowledge as a matter of being scientific first appeared in the writings of E. C. Richardson who proposed that for a classification of knowledge to have a trustworthy base, it must reflect the actual relationships of the real corresponding objects to which knowledge categories refer. Having concluded this, however, even Richardson backed away from pursuing it and concluded that any actual classification scheme must ultimately give place to issues of practicality. H. E. Bliss continued Richardson's theme that classification would take on a truly scientific cast only as it approached a scientific consensus about the objects which corresponded to knowledge categories. But, he also expanded greatly the realm of what might be called sound "scientific" technique that must accompany a truly scientific approach to classification.

Since Richardson and Bliss, these two themes—the scientific basis of knowledge categories and the scientific basis of classificatory techniques—have become the fundamental elements of library classification theory in the twentieth century. The first was continued in the work of the Classification Research Group in Great Britain as it struggled to found classificatory categories on the phenomena of material and human existence and in S. R. Ranganathan's appeal to a mathematical analogy for the universe of knowledge. The second was continued superbly in S. R. Ranganathan's development of an entire nomenclature of ideas and techniques related to faceted classification structure. These two themes also underlie the changes that have overtaken the DDC since its 15th edition, for they are fundamental to attempts to change categorical order in the DDC so as to become more modern and they also underlie the incorporation of modern
classificatory techniques in the DDC.

In all of this accumulation of theory, however, three basic assumptions have never been directly questioned—first, that classificatory structures of knowledge must necessarily be hierarchical in nature, second, that there is the possibility of discovering the one true or at least most accurate portrayal of such a knowledge structure, and third, that library classification is useful chiefly for document retrieval.

4. The DDC and the Post-Modern Library

The DDC as well as all other library classification work has now entered an age in which a post-modern library is emerging. The idea of a post-modern library specifically refers to a new kind of technologically based library which differs from the "modern library" invented towards the end of the 19th century. This new kind of library has the potential of creating electronic "personal-space" libraries rather than simply "public-space" institutions. This new kind of a library also has the potential for solving some of the information access problems of the modern library of the past century by having a more cohesive user target or focus rather than the heterogeneous user group focus necessarily served by the public-space modern library. One way to identify the impact of this new form of a library is to speak of it in terms of two additional laws to be added to Ranganathan's five laws of library science—"Every person his or her own library." "Every library its own user."

Another aspect of the post-modern library is the appropriateness for it of at least some of the ideas related to other themes in contemporary thought. First, post-modernism as discussed by literary scholars and others in the human sciences suggests strongly that there really is no single true classification structure of knowledge, but rather that there are any number of such schema which have arisen as kinds of discourse about the structural relationships of knowledge categories. This appears to be appropriate for the post-modern library for which the ideal appears not to be the imposition of classification schemes which serve a "public-space" in some economical and practical way, but rather the creation of individually tailored classificatory schema which serve individual personal space libraries via methods most appropriate for each. Second, other modern studies suggest that classificatory knowledge structures do not require formal or logical hierarchical structures to reflect the way people think about knowledge categories and their relationships. In a sense this has already been recognized in part within library classification, because the effect of Ranganathan's faceted techniques has been to flatten hierarchical knowledge structures in favor of what might better be called relational structures. Third, the use of classificatory structures in the field of librarianship has over the course of the twentieth century become strongly associated with the retrieval of documents and of information. The fact is, however, that a much broader base of purposes of classification suggests uses for classifying knowledge other than retrieval, not the least of which are the exploration of idea relationships for educational and teaching purposes, the creation of new ideas, and for analyzing relationships among information-bearing entities.

The potential uses for the DDC in the age of the post-modern library are enormous. However, those who work on the scheme must be prepared to deal with the scheme in terms of these different uses and this will require in turn that modern techniques for structuring and developing the scheme be explored. Possibilities for the latter include the systematic application of data-mining and knowledge discovery techniques to the scheme as a database of categories, to the application of the scheme in actual library collections, and to the transactions of actual use of the scheme in retrieval, in order to discover previously hidden patterns of relationships among categories. It might also include exploring the possible reformatting of the entire system as a
purely object relational database management system of categories. Finally, it will most likely include creating a system so supple that end users can use it to create their own structures of knowledge categories for their own personal space libraries and for their own discourse about the relationships that pertain among knowledge categories.
List of Contributors

Ingetraut Dahlberg
Woolstr. 36a
D-60431 Frankfurt, GERMANY
+49-69 523690 (telephone)
+49-69 520566 (fax)

Giovanni Adamo
Consiglio Nazionale delle Ricerche
Lessico Intellutuale Europeo
Via Nomentana, 118
1-00161 Rome, ITALY
+396 86320527 (phone)
+396 49917215 (fax)

Julianne Beall
Decimal Classification Division
Library of Congress
Washington, D.C. 20540-4330, USA
jbea@loc.gov
+1 202 707 5715 (phone)
+1 202 707 0279 (fax)

Carol Bean
Center for Medical Informatics
Columbia University
Columbia-Presbyterian Medical Center
161 Fort Washington - AP 1310
New York, NY 10032, USA
beancar@cucis.cis.columbia.edu
+1 212 928 4610 (phone)
+1 212 305 3302 (fax)

Patrick Braekevelt
CeDAR Centre for Database Access Research
School of Computing & Mathematics
The University of Huddersfield
Queensgate, Huddersfield HD1 3DH, U.K.
cedar@hud.ac.uk
+44 1484 472147 or 472248 (phone)
+44 1484 421106 (fax)

Faina Citkina
2453 Kennedy Blvd, #2
Jersey City, NJ 07304, USA
102044.275, @compuserve.com
+1 201 435 2123 (phone)
+1 201 435 2123 (fax)

Pauline A. Cochrane
Graduate School of Libr.& Inform. Science
University of Illinois at Urbana-Champaign
501 E. Daniel
Champaign, IL 61820-6212, USA
cochrane@alexia.lis.uiuc.edu
+1 217 244 1403 (phone)
+1 217 244 3302 (fax)

Ron Davies
Bibliometrics Inc.
48 200 Owl Drive
Ottawa, K1V 9P7, CANADA
rdavies@bibliom.synapse.net
+1 613 523 7981 (telephone)
+1 613 523 4417 (fax)

Karen M. Drabenstott
School of Information
University of Michigan
550 East University Avenue
Ann Arbor, MI 48109; USA
karen.drabenstott@umich.edu
+1 313 763 3581 (telephone),
+1 313 764 2475 (fax)

Geoffrey P Ellis
CeDAR Centre for Database Access Research
School of Computing & Mathematics,
The University of Huddersfield
Queensgate, Huddersfield HD1 3DH; U. K.
cedar@hud.ac.uk
+44 (0)1484 472147 or 472248 (phone)
+44 (0)1484 421106 (fax)

J.P.J.M. Essers
Brasmus University Rotterdam
P.O. Box 1738
3000 DR Rotterdam, NETHERLANDS
jessers@staff.fbk.eur.nl

Janet E. Finlay
CeDAR Centre for Database Access Research
School of Computing & Mathematics
The University of Huddersfield
Queensgate, Huddersfield HD1 3DH, U.K.
cedar@hud.ac.uk
+44 (0)1484 472147 or 472248 (phone)
+44 (0)1484 421106 (fax)

Ana Flavia Fonseca
Information Operations Unit
Inform. & Technology Services Dept.
The World Bank
Washington, DC 20043, USA
afonseca@worldbank.org
+1 202 473 3176 (phone)
Widad Mustafa Ëlhadi  
Université Charles de Gaulle Lille 3  
UFR IDIST/CREDO, BP 149  
F-59653 Villeneuve d'Ascq Cedex  
FRANCE  
mustafa@univ_lille3.fr  
+33 20 41 62 30 (phone)  
+33 20 41 63 79 (fax)

Gigliola Negrini  
Consiglio Nazionale delle Ricerche  
Ist. di Studi sulla Ricerca e Doc.Scientifica  
Via Cesare De Lollis, 12  
I-00185 Rome, ITALY  
+39 6 448791 (phone)  
+39 6 4463836 (fax)

Gregory R. New  
Decimal Classification Division  
Library of Congress  
101 Independence Avenue, S.E.  
Washington, D.C. 20543 3300, USA  
gnew@mail.loc.gov  
+1-202 707 6983 (phone)  
+202 707 0279 (fax)

Toyoaki Nishida  
Graduate School of Information Science  
Nara Institute of Science and Technology  
8916-5 Takayama-cho, Ikoma-cho  
Nara 630-01 JAPAN  
nishida@is.aist-nara.ac.jp  
+81 7437 9 9211 ext. 5316 (phone)  
+81 7437 2 5269 (fax)

Marc Nodell  
Messaging and Collaborative Tools Unit  
Information & Technology Services Dept.  
The World Bank  
Washington, DC 20043, USAAA  
mnodell@worldbank.org  
+1 202 473 2860 (phone)

Hope A. Olson  
School of Library and Inform. Studies  
University of Alberta  
3-20, Rutherford South  
Edmonton, Alberta T6G 2J4 CANADA  
holp@ualberta.ca  
+1 403 492 4578 (phone)  
+1 403 492 2430 (fax)

Roberto Poli  
Department of Sociology and Social Research  
University of Trento, 26, Verdi Street  
I- 38100, Trento ITALY  
poli@riscl.gelso.unitn.it  
+39 461 881 403 (phone)  
+39 461 881 440 (fax)

Steven A. Pollitt  
CeDAR Centre for Database Access Research  
School of Computing & Mathematics  
The University of Huddersfield  
Queensgate, Huddersfield HD1 3DH, U.K.  
cedar@hud.ac.uk  
+44 1484 472147 or 472248 (phone)  
+44 1484 421106 (fax)

A. R. D. Prasad  
Documentation Research and Training Centre  
Indian Statistical Institute  
8th Mile, Mysore Road, R. V. College Post  
Bangalore 560 059 INDIA  
ard@isibang.ernet.in  
+91 080 8430975 (phone)  
+91 080 8430265 (fax)

Massimo Ragucci  
Ministero delle Risorse Agricole  
Alimentari e Forestali  
Via Carducci, 5  
I-00187 Roma, ITALIA  
+39 6 46655040 (phone)  
+39 6 4881707 (fax)

J.F. Schreinemakers  
Erasmus University Rotterdam  
PO Box 1738  
NL-3000 DR Rotterdam, NETHERLANDS  
J.Schreinemakers@fac.fbk.eur.nl

Debora Shaw  
School of Library and Information Science  
Indiana University  
Bloomington, IN 47405-1801, USA  
shawd@indiana.edu  
+1 812 855 3261 (phone)  
+1 812 855 6166 (fax)

Snunith Shoham  
Bar-Ilan University  
Dept. of Inform. Studies & Librarianship  
Ramat Gan 52900, ISRAEL  
+972 2 865626 (fax)

Martin P. Smith  
CeDAR Centre for Database Access Research  
School of Computing & Mathematics,  
The University of Huddersfield  
Queensgate, Huddersfield HD1 3DH, U.K.  
cedar@hud.ac.uk  
+44 1484 472147 or 472248 (phone)  
+44 1484 421106 (fax)
Dagobert Soergel  
Hombake Bldg. (So. Wing), Room 4105  
College of Library and Information Services  
University of Maryland  
College Park, MD 20742, USA  
ds52@umail.umd.edu  
soergel@umail.umd.edu  
+1 301 405 2037 (phone)  
+1 301 314 9145 (fax)  

Harold C. Steyer, Jr.  
Information Engineering Unit  
Inform. & Technology Services Dept  
The World Bank  
Washington, DC 20043, USA  
hsteyer@worldbank.org  
+1 202 473 2267 (phone)  

Eduard R. Sukiasyan  
Deputy Director, Russian State Library  
3 Vozdvizhenka Str., Moscow, 101000 RUSSIA  
irgb@glas.apc.org  
+095 200 22 55 (fax)  

Hideaki Takeda  
Graduate School of Information Science  
Nara Institute of Science and Technology  
8916-5 Takayama-cho, Ikoma-shi, Nara 630-01 JAPAN  
+81 7437 9 9211 ext. 5316 (phone)  
+81 7437 2 5269 (fax)  

Irene L. Travis  
6003 Highland Hills Drive  
Austin, TX 78731, USA  
itavis@aol.com  

Mark Treglown  
CeDAR Centre for Database Access Research  
School of Computing & Mathematics  
The University of Huddersfield  
Queensgate, Huddersfield HD1 3DH, U.K.  
cedar@hud.ac.uk  
+44 (0)1484 472147 or 472248 (phone)  
+44 (0)1484 421106 (fax)  

Diane Vizine-Goetz  
Office of Research, OCLC  
6563 Frantz Road  
Dublin, Ohio 43017, USA  
vizine@oclc.org  
+1 614 764 6084 (phone)  
+1 614 764 2344 (fax)  

Steven J. Wade  
CeDAR Centre for Database Access Research  
School of Computing & Mathematics  
The University of Huddersfield  
Queensgate, Huddersfield HD1 3DH, U.K.  
cedar@hud.ac.uk  
+44 1484 472147 or 472248 (phone)  
+44 1484 421106 (fax)  

William S. Wahl  
Information Engineering Unit  
Inform. & Technology Services Dept.  
The World Bank  
Washington, DC 20043, USA  
wwahl@worldbank.org  
+1 202 473 2801 (telephone) (fax)  

Thomas D. Walker  
University of Wisconsin-Milwaukee  
School of Library and Information Science  
P.O. Box 413  
Milwaukee, WI 53201, USA  
walker@csd.uwm.edu  
+1 414 229 5397, +1 414 447 7589 (phone)  
+1 414 229 4808 (fax)  

Nancy Williamson  
Faculty of Information Studies  
University of Toronto  
140 St. George Street  
Toronto, Ontario, M5S 3G6 CANADA  
william@fis.utoronto.ca  
+1 416 978 7079 (phone)  
+1 416 971 1399 (fax)  

Moshe Vlizhaki  
Bar-Ilan University  
Dept. of Inform. Studics & Librarianship  
Ramat Gan 52900 ISRAEL  
+972 2 865626 (fax)  

Marcia Lei Zeng  
School of Library and Information Science  
Kent State University  
Kent, OH 44242 0001, USA  
mzeng@kentvm.kent.edu  
+1 330 672 2782 (phone)  
+1 330 672 7965 (fax)
Name and Subject Index

A
Abercrombie, N. et al 224, ref 230
Absolute Syntax, searched for by Ranganathan 157
ACABIT
for automatic terminological bank construction 349
Adamo, G. paper 275-283, ref. 282
Add notes, similar to Standard Citation Order 123
Addison, E.R. 59, ref, 63
Afolabi, M. 303, ref. 311
Agosti, M . 90, ref, 93
Agriculture, integration of general information systems 284-292
AGROVOC 285
Ahlsweide, T. ref, 173
Ahmad, K. 350, ref, 353
Aho, A.V. ref, 353
Aitchison, J. 371, ref, 376
Aityceu, W.O. ref, 301
AIDS, Acquired Immunodeficiency Syndrome, Clearinghouse Educational Materials Database Thesaurus 295 Information Network (AIN) 295
Akerely, O. 58, ref. 58
Al-Quaimari, G. ref, 240
Alexandria Library - Pinakes of Callimachus 43
Allen, R.B. 95, ref, 104
Allerton Conferences 95
Alta Vista Web search engine of Digital 11
to replace indexers 211
ALA American Library Association 26, 407
Ames, A. ref. 260
Anatomie entities terms 84
Anatomical terminology relationship studies 81-86
Anderson, J.D. 59, ref. 63
Angel, Richard S. 36, ref 41
ANA a statistical and/or connectionist model 349
ANSI Z39.50 Search and Retrieval Protocol 137
ANSI/NISO Z39.50 ref. 143
Apple Macintosh's graphical user interface, balloon help 113
Applicative and Cognitive Grammar, applied for semantic relations 348
Archie 273
Area Table joint, for DDC and UDC 126
Amheim, R. ref. 107, 181-2, 188-9, 197
Art & Architecture Thesaurus (AAT) 65, 184-, 191- as example 142 ref. 71
Ashurbanipal library, archive 43
Atherton, P. ref. 105
Attributes of images describing and sorting 190-
Austin, D. 122, 174, ref. 122, 128, 181
Authorization formalism of broader terms 70
Automatic indexing system, Prometheus 329-335
AUDACIOUS, first online classification system 97
AUPELF Assoc.Universités Ent.Ou Part.de Langue Francaise 353
Avrin, L. 43-, ref. 49

B
Bacon, F. 156, Bacon, F. 409 Bainbridge, L. ref. 222
Baker, S.L. 196, ref. 197
Bakewell, E. 189, ref. 197
Barnwell, K. ref. 63
Bates, M.J. 67, ref. 71
BBC Bliss Bibliographic Classification 155, 163, 371 structure used for revision of a UDC class 127
BC2 Class H schedule model for UDC Class 61, medicine 371-
Beall, J. 382-3, paper 396-405 ref. 385, 395
Bean, C.A. paper 80-86, ref. 79, 86
Beamam, D. 21, 199, ref. 23, 205
Becket, R. 19, ref. 23
Beghtol, C. 196, ref. 197
Benbasat, I. 176, ref. 181
Bendig, M. ref. 405
Bennett, D. 210
Bernstein, R.J. 214, ref. 215
Besser, H. 182, ref. 188
Bevan, N. ref. 240
Bhattacharyya, G. 330, ref. 334
Bias, R.G. ref. 240 et al 239
Bible, uniform entry for 45
Bierbaum, E.G. ref. 300 et al 294, 297
Bioski, K. ref. 181
Bird, P.R. ref. 49
Bjerg, K, ref. 23
Björklund, L. 90, 95, ref. 93, 104
 Blair, D.C. 155, ref. 163
Bliedung, S. ref. 150
Bliedung, St 371, ref. 377
Bliss Bibliographic Classification see BBC
Bliss, H.E. 410
Blunt, A. ref. 311, et al 303
Bodley, Thomas 45
Boling, E. ref. 181 et al 177
Bolognesi, D. ref. 300
Bookhouse virtual library interface, uses icons 179
Boolean combinations via DDC Search window 396
Borgeest, M. ref. 173
Borghes, J.L. 323, ref. 327
Borgman, Ch.L. ref, 259, et al 259
Botafogo, R.A. ref. 93, et al 90
Bottle, R.T. ref. 49, et al 47
Bourigault, D. 349, ref. 353
Bowen, S.P. ref. 300
Bowker, L. ref. 353
Bradford, S. 410
Braekevelt, P.A. paper 231-240
Bremes, D. ref. 180
British Classification Research Group (CRG) 157, 371, 410
Britten, W.A. 254-, ref. 259
Broad System of Ordering see RSO
Broder, S. ref. 300
Categories

identification of elementary
333
and relationships reflect
semantic structure of
knowledge domain 85

Category of whole

as the top level of being 317

Cause-effect relations

METIS produces info about
339-341

Cawkell, A.E. 182, ref. 188

CAB-thesaurus 284

and AGROVOC in compara-
ison on Spikes, plant
breeding, etc. 288-291

Centered entries

in a DDC search, look for
comprehensive 401

Central University Library of
Bucharest 144

CERETHIS

facets of 285-287

thesaurus for cereals,
construction of 284-292

Chadwick, D. ref. 205

Chafe, W. 76, ref. 79

Chambers, S. et al ref. 180

Chan, L.M. 34-, 95
ref. 32, 41, 104

Chang, Sh.-j. ref. 260,
et al 257-

Change

terminology organization
and 355-361

in DDC, strategies for 310

of CS internal and external
reasons for 119-

oriented knowledge organi-
zation, phenomenology

as bases for 21

Change phenomena

prognostication of 120

of CS, processing of 120-

Changeability of a CS 119

Chemical agents terms in

MeSH 85

Chen, H. ref. 106

Chen, L.J.-j. et al 19, ref. 23

Chemoyatin, V. 169, ref. 173

China Academy of Medical
Science 52

China National Science
Foundation 51

Chinese herbal medicine
indexing practice in
MEDLINE 54

Chinese Libraries
Classification 52-

Chinese medicine 51

Chinese names of herbs
in titles and abstracts 55

Chinese Thesaurus for Medical
Science 52-

Christen, F.G. ref. 181

Chronological awareness in
group systems 19

Churcher, P.R. 89, ref. 93

CHESHIRE system

managed high-posted
searches by classification
108

Cikina, F. 360-361,
paper 355-361, ref. 362

Clack, D.H. 302, ref. 311

Clair, C. 44-, ref. 49

Clark, G. ref. 23

Class K, (Law), LCC

perimeter of classes,
adjustments, revisions
39-

Classes to describe tasks
comparison of, fig. 190

Classification

and graphic symbol systems
174-180

at LC, Minaret program for
29
categories for images,
selection of 184
culture clash 186
data, improvements needed
for online access 249

experts, opportunities for,
on design of 211

for slides and photographs
184

importance of 209

labels, graphic, used in
libraries 179

of discrete elements,
necessary to computerize
information 209

of graphic images by Gittins
178

of knowledge, relationship
of the DDC to 408

online, potential uses for
26-

Plus, a CD-ROM product of
CDS (LC) 31

records, new MARC record
distribution planned by
LC 28

scheme for user types 217-218

maintenance of 122-136

system, completely new - an
ideal situation 376

system development,
change as a problem of 119-121
system stability 119
systems for image attributes 189-197
demand for construction of adaptive 211
theory, sizable literature of in 20th century 410
to increase retrievals 114-used in low-posted searches 108
useful for relevance feedback 108
window for displays proposed 104
Classificatory structures
symbolic languages as 178
Classifier, Dewey for Windows provides diff.approaches for 396
Classifiers as users, resistant to change 124
Cleverdon, C. 231
CLASTHES software for hierarchies of terms and categories 287
Coates, E. 160, ref. 128, 163
Cochrane, P.A. 107, 236, paper 95-106
ref. 104, 106, 117, 240, 385 et al 95-, 103, 384
Cognitive Walkthrough, a usability inspection method 235
Cohen, P.T. ref. 300
Cole, J.Y. 34-, ref. 41
Coles, Anne P. ref. 41
Collocations in terminology lead to phrases and phrasology 347
Colon combination
a UDC feature, a tool for number building 372
Colon Classification see RCC
Comaromi, J.Ph. 194, 305, 386, 390, 408
ref. 197, 311, 395
Common Auxiliaries in the UDC and revision work 125
Communication required for devising automated systems 123
Communications of the ACM employs two forms of classification in WWW 210
Compaine, B. ref. 327
Comparative terminology science (CTS) 360
Compatibility matrix to correlate CS 161
Comprehensive, a search command in DDC 401
Concept identifiers fig. 170
Concept systems
dynamism of 278
evolution of 275-283
proliferation and integration 281-2
Concept-based and word-based measures, comparison between 59
Concepts and terminology
SemWeb, system for integrated access to 165-173
Conceptual Information Value (CIV) of titles 60-
Conflicts about DDC 407-
Conformity, needed in traditional schemes 123
Conklin, J. ref. 93
Conlon, S.P.N. ref. 173
Consistency
in number application, necessary for users 250
Construction and exploration of advanced technological information space 336-345
Context
and user needs 233
Analysis questionnaire and report SANe, DRUM, and SUMI as MUSIC tools 232-
Context-free Phrase Structure Grammar with Definite Clause Grammars and PATR 330
Contextual Exploration Approach used by SEEK 349
Core structure, common across several indexing languages 127
Cornell, D. 303, 310, ref. 311
Corporate Knowledge Management (CKM) new field of activity 213
Council on Library Resources sponsor of subject access conferences 95
Cover, R. 21, ref. 24
COGNETERM 347
CONTERM based on connectionist model 349
Crampton, G. ref. 311
Crawford, W. ref. 32, 260
CRG see British Classification Research Group
CS, abbr. for Classification System
Cullingford, R.E. 344, ref. 345
Custer, B. 387, 408, revision of DDC 560-590 390-
Customized view to build a number by Dewey for Windows, fig. 403
Cutter, distinction between title and subject approach 46 number 37
Cutter's Expansive Classification 36
Cutter, Ch.A. ref. 49
Cyberstacks(sm) uses LCC, established at Iowa State Univ.Library 254-
a selection practice for document-like objects 255
example for access of collections in the Internet 249

D

Dahlberg, I. 17, 275, 284, 348, 365
paper 155-164
ref. 24, 163, 282, 283, 291-292, 353, 369
Daille, B. 349, ref. 354
Daily, J.E. 96, ref. 105
Daniel, R. ref. 188
Danowski, J.A. 107, ref. 118
Dascal, M. ref. 327
Database access and analysis, improved by knowledge structure views, etc. 231-240
Daugherty, R.A. 107, ref. 118
Dauphin, E. ref. 354
Davies, R. paper 137-143
Davis, M. ref. 362, et al 357
DANTERM 347
DCTNet 16
DDC 155-, 184, 191-004.678 number for Internet 380
21 new topics and relationships 380-
structural changes 382
350-354, comparison between Ed 21 vs Ed 20, fig. 379
350-354, Public Administration revision of 388
370, Education revision of 388
innovation with caution 389-
candidate for a standard for
cataloguing 126
database 388-
episodes 1-21 history of 406-
enhancements sought for
online classifying 249
epistemic stance of 302-312
history of editions 386-
in a hypertext browser
for the library user 95-106
major revisions in ed.21 378
numbers, frequency of in
NetFirst Database, table
251
Number window, fullest info
on one class 401
Online Project 27
possible reformattting of
entire system 411
potential uses in the post-
modem library 411
Pages display shows multi-
ple adjacent schedule
and table records 401
Project of Markey and
DeMeyert 97
search for alcoholism, pro-
posed display (Fig.) 102
terminology, updated in
DDC21 381
used to decrease retrieval
items 109.
Universe of Knowledge and
the Post-Modern Library
406-412
Windows interface 396
De Vere, R. ref. 117
De Vinne, T.L. 44, ref. 49
December, J. ref. 24
Decimal notation
principle of 158
Deconstruction of the resource
to computerize knowledge
resources 208
Defining
rules and models 363
template for an entity/object
(type of concept),
fig. 366
Definition
inadequate, modification of,
fig. 367
standardization of an
existing, adequate, fig. 367
DeJong, G.F. 344, ref. 345
Demas, S.G. ref. 260,
et al 255
DeMeyer, A. ref. 32, 105
Dependence relation
among levels of being 315
Dependency, theory of
necessary in constructing a
general ontology 318-
319
Desceles, J.-P. 353, ref. 354
Description
levels of 244
role of in retrieval model 242
Descriptive Terms for Graphic
Materials 191-
Descriptor equivalence
kinds of 147
Devices for seeking, gathering
in probing in the OPAC 96
De Vita, V.T. ref. 300
Dewey 21: Revision and
Stability in 386-395
Dewey and DDC
epistemic stance of 302-312
Dewey Decimal Classification
see DDC
at 120: Edition 21 and
Beyond 378-385
Committee of the British
Library Association 393
Dewey for Windows 396-405
alcoholism displayed in 98
includes statistically mapped
LCSH 383
ref. 385, ref. 404
Dewey home page
links to Internet resources
organized by DDC 384
Dewey Online Catalog (DOC)
108
Dewey’s
business orientation 409
orthography, examples 306-
307
Dewey, M. 35, 156
ref. 311, 385, 395
Diaper, D. 235, ref. 240
Dichotomies handled in DDC
308
Dien, R. 60, ref. 63
Dietrich, H. ref. 23
Digital Equipment Corporation
used classification for
corporate-wide KB 210
Digital Libraries Initiative
of the University of
Michigan 187
Diodato, V. 47, ref. 49
Disciplinary DDC numbers,
selection and identification
of 399
Displays for end users
need special editing 98
Disruption to function terms in
MeSH 85
Divide-like rule of Dewey
used by Otlet and LaFontaine
for the UDC 123
DIALOG mentions of Internet
11
DLOs, document-like objects
on the Internet 244
Doan, K. ref. 240
Document, def. 199
components with added
metadata for classification
209
structure, formalization of 20
Documentary linguistics
contributions of to hypertext
89
Dolby, R.G.A. 410
Domanovsky, A. 46, ref. 49
Drabenstott, K.M. 108, 111,
189, paper 107-118
ref. 117, 197, s.a. Markey, K
Drag-and-drop approach
in Dewey for Windows 404
DragNet 16
DRUM, Diagnostic Recorder
for Usability Measurement
233
DS/HSS-Data Thesaurus 371
Dublin Core
small set of descriptive
elements 244
Metadata initiative 187
Dublin Metadata Core Element
Set, a record format for data
elements in the Internet
258
Duffy, T.M. ref. 240
Duke University online DDC
catalog, used for subject
heading searches 108
Duyvis, F.D. 410
E
Earth and space science
image databases strategy,
model for 187
Easterby, R.S. 177, 179,
ref. 181
ECOT Thesaurus 371
Edigo, C. ref. 180
Edinburgh FID Conference
160
Editorial Support System (ESS)
a UNIX-based system of
DDC 383
Editors of Classification
Schemes,
cooperation between 127
EDF-DER 349
Eisenstein, E.L. 320,- ref. 49, 327
Electronic Dewey 18
a Dos version of ed 20 383
CD-ROM database in DOS
version 396
designed for classifiers 97
ref. 385
User Guide ref. 105
Electronic environment,
description in 242-248
Ellis, G.P, paper 231-240,
ref. 240
Emmis sa ry 16
Encyclopaedia Britannica
online 325
Encyclopaedia Society
to be formed acc.to Wells
326
Enguehard, C. 349, ref, 354
Enset, P.G.B. 189, ref, 197
Entity types and identifiers 169
Epistemic stance,
a view of or belief rel.Lo
ture of knowledge 304
EPOQUE
European Parliament Online
Query System 232
Equivalence of descriptors
tions of 147
Issers, J. paper 213-215
Esteban Navarro, M.A.
ref. 24, 94
Evans, Luther 37, 41
Evens, M. ref. 173
Excite search engine in WWW
242
Expediency
Dewey’s criterion of a
universal system 305
Expert systems approach
proved difficult to apply in
other areas 231
Expressive notation
of value in online
environment 124
Extensions and corrections to
the UDC 124, 370

F
Facet structure
best basis for thesaurus
derivation 371
Faceted classification scheme
for user profiles and views
220-221
Faceted structures already in
first DDC edition 406
Fehrle, L. ref. 327, et al 320-
Fedorowicz, J. ref. 274
Feibleman, J.K. 157, 162
Feinberg, H. 47, ref. 49
Felley, M. ref. 64
Fellows, Dorcas 407
Ferguson, D. et al 95, ref. 105
Fidel, R. ref. 74, 79, 224, 230
Filaro, T.W. ref. 300
Fillmore’s case grammar for
basic relational structures 76
Finin, T. 19, ref. 24
Finlay, J.E. paper 231-240
FID,
Féd.International d’Inform.
et de Documentation 156
Florance, V. ref. 300,
et al 293, 296
Focal terms
derived from head terms 82
Fogg Classification Schemes
184
Fonseca, A.F. paper 198-205
Formal ontologies 17,-, 313
Fomney, Ch.D. 294, ref. 300
Foskett, A.C. 302, ref. 311
Foskett, D. 157
Foss, M.M. paper 216-222
Fowler, H.W. 170
Fowler, R.H. 21, ref. 24
Francis, W. 79, et al 74
Francu, V. paper 144-154
Freeman, R.R. ref. 105,
et al 95
Freextext-based searching
complexity of 55
Frost, C. 47, ref. 49
paper 182-188
FRUMP, Fast Reading
Understanding and Memory
Program 344
FTL,
see Full Text Lexicographer
Fugmann, R. 89, 146, 155,
ref. 94, 150, 163
Full Text Lexicographer
(FTL) 183
Function terms in MeSH 84
Furuta, R. ref. 94
G
Gaines, B.R. ref. 23
Galdes, D. ref. 106
Gatinski, Ch. ref. 173
Gallagher, A.L. ref. 259
Ganeshsundaram, P.C. 59,
et 63
Garcia Marco, F.J. ref. 24, 94,
et al 17, 93, paper 87-94
Gardin, J.C. 70, ref. 71
Garnier, F. 193, ref. 197
Gazdar, G. ref. 334, et al 330
GABRIEL 155
GAteway to and BRidge
between European
Libraries ref. 164
Geiselman, R.E. ref. 181,
et al 178
General concepts
in universal systems,
problems of 373-4
General Record Syntax GRS-I
accomodates diff.types of
thesaurus structures 139
Gesamtkatalog der
Wiegendrucke 321
Gesellschaft für Klassifikation
e.V., 20th Annual Conf. 21
Gesmer, Konrad 325
Ghosh, J.S. ref. 50
Gibb, F. 90, ref. 94
Gignac, P. ref. 274
Gilchrist, A. 371, ref. 376
Gillaspy, M.L. 294,
paper 293-301, ref. 300
Girill, T. 93, ref. 94
Gittins, D. 175-176
created classification of
graphic images 178
ref. 181
Giurelli, M. ref. 292
Glaister, G.A. 43, ref. 50
Godby, J. ref. 188
Godman, A. ref. 63, et al 59
Goldberg, J. 36, 38-9,
paper 33-42, ref. 41
Gormann, M. ref. 260
Gostin, L.O. 299, ref. 300
Gowin, D.B. ref. 94
Gönpel, R. 155, ref. 164
GrabNet 16
Gradenigo, G. ref. 93
Graphical Boolean Expressions
and Dynamic Queries
parallels view-based
searching 232
Graphical User Interfaces
(GUI) 174
for classification schemes on
CD-ROM 97
Green, R. 80, paper 72-79,
242-248, ref. 79, 86
Greenhalgh, M. ref. 41
Griffin, T. ref. 94
Griffith, B.C. ref. 300
Gruber, T.R. 17, 262,
ref. 24, 267
GRAAL 349
GSI-ERLI 349
Guarino, N. ref. 319,
Classification, see ICC
Information Management
universal personal 16
Management Architecture
of the World Bank 198-205

Informational Value (IV) of
titles 60-
Informativeness measure, two
types of 60
Informativity of titles, increase
in 20th century 46
Infoseek
popular Internet search
service 249
search engine in WWW 242
Ingarden, R. 319, ref. 319
Innovation vs stability, general
issues 386
Instantiation formalism to link
individual concepts 70
Institut Internationale de
Bibliographie (III) 407
Institute for Scientific
Information 20, ref, 24-25
Institutional information
services,
model for by World
Bank’s IMA 198-205
Int. Soc. for Knowledge
Organization, see ISKO
Integrated access to knowledge
SemWeb, system for 165-173
Integrated knowledge base,
usefulness of 166
Integrated Level Theory 157
Integrity of numbers
no change in meaning of
(DDC) numbers 386
Intelligent Information,
Collector and Analyzer
IICA 261-267
Inter-Concept Links ,
standardization and usage
65-71
InterCat Project of OCLC 244
Interdisciplinary DDC numbers
selection and identification
of 399
Interfaces, browsable
needed for users 257
International Thesaurus of
Refugee Terminology 370
Internet
and the nature of knowledge
320-327
changing, amorphous
catalyst 11
definitions of 12-13
has BURL Bulletin Board,
uses UDC 123
Library catalogs in the 155-164
knowledge structures may
play roles in searching
habits 324
Intranets 16
INDEKS Verlag 163
INSPEC Thesaurus 100
IOTA, based on a Sowa
categorical graph model
349-
ISKO 206, 208,
International Society for
Knowledge Organization
162
ref. 58
Seminar on Compatibility
and Order Systems,
Warsaw, Sept.1995 165
ISO 10162 describes inter­
action between client and
server 137
ISO 10162/63
ref. 143
ISO 2788 147, 371, ref 377
ISO CD 12620.2 ref. 173
ISO DIS 12200 ref. 173
Italian (CNR) Experimental
Institute on Cereals
Cultivation 284
Italian Literature, autonomous
field, structure of 277-281
Ivari, J. 17, ref. 24
Iwayama, M. ref. 79, et al 75
Iwazume, M. 261-2,
paper 261-267, ref. 267
Iwuji, H.O.M. 303, ref. 311
J
Jacob, E.K. paper 174-180
Jaenecke, F. ref. 222
Jefferson, Thomas
classification scheme 34
Jelinek, M. 390
Jewett, Ch.C. ref. 50
John, B.E. ref. 240, et al 236
Johnson, E. 99, ref. 105,
ref. 385
Johnson, E.D. ref. 50,
et al 43-
Johnson, E.H. paper 95-106
Johnson, L. ref. 181
Johnson, M. 76, ref. 79
Johnson, R. ref. 106
Jones, R. ref. 94
Jouis, C. 349, 353, ref. 354,
et al 348, paper 346-353
Jörgensen, C. 174, 196,
paper 189-197, ref. 181,
197
Judge, A.J.N. 21, 158,
ref. 158, 164
K
Kacmar, C.J. ref. 180
Kaliyan, S. ref. 63, et al 59
KB, stands for Knowledge
Base
KBSI ref. 24
KBSI Ontology modeling 18
Keister, L.H. 189, ref. 197
Kempf, A. 20, ref. 24
Keyword searches
efficient for rare terms in the
DDC 397
KES, Knowledge Extracting
and Structuring 349
Kiel, E. ref. 222
Kilgo, A.C. ref. 240
Kirby, D.G. ref. 173
Kirkley, S. ref. 181
Knight, M.A. ref. 49
Knowledge
base, functions of on
concepts and terminology
166-167
Engineering, terminology
and 356
field, identification and
definition of 276
intangibility of 213
justification, critically seen
by Kuhn and Habermas
214
Management, critical notes
on use of knowledge in
213-215
nature of, conceptions of
320
perceptions of and Web
dialectics 322
Knowledge Organization
and business productivity
206-212
challenges of change for 10-25
ethics of 22
hidden assumptions of 14
implications for 15–
in a business context 199-
matter for everyone 16-
ontology for 313-319
re-defining systems for?17-
re-orientation of aims of
systems for 18-
Knowledge Pieces (KP)
def. 338 336-
Knowledge representation method with procedures for extracting and structuring 337
model, frame-based, suited for nounphrases 329
Knowledge resources reverse engineering of 208 tacit, of members of an organization 213
Knowledge structure views improve database access and analysis 231-240
Knowledge structures incompatibility for lack of rules for links 67
Knowledge Sharing Effort 17 public library ref. 24
Knowledge Transfer International 206-212
Knowledge-based business paradigm to challenge information and libr.sci.people 212
Kohl, Herbert 170, ref. 173
Kong, L. ref. 64
Korfhage, R. 21, ref. 24
Korienek, G. ref. 181
Koujitan, K. ref. 267
Kovacs, J. 59, ref. 63
KO, abbreviation for Knowledge Organization
Krull, R. 176, 177, ref. 181
Ku, K.M. 59, ref. 64
Kucera, H. ref. 79
Kuhn, Th.S. 213, ref. 215
Kukulska-Hume, A. 347, ref. 354
Kumari, L. ref. 301
Kume, M. ref. 63
Kunze, J. 139, ref. 143
KULTURARW3 downloads W3 documents in Sweden 22
Kwasnik, B.H. 182, ref. 188
KWIC and KWOC indexes 47
LaFontaine, H. 407, 410
Lake Placid Club Educational Foundation 407
Lakoff, G. 26, ref. 240
Lancaster, F.W. ref. 150, ref. 300
Landee, B.M. ref. 181
Large, J.A. 59, ref. 64
Larson, M.L. 59, ref. 64
Larson, R.R. 107-8, ref. 117
Law Classification, LCC history of 38-
policies and principles 38-
Lawrence, G. ref. 260
LBC Library Bibliographical Classification (BBK) 155-163
LCC Library of Congress Classification 26-32 155-, 184, 302, ref. 385
Additions and Changes 27 conversion to electronic format 26-32, 40
Conversion into machine-readable form 26-32
Dewey refused the DDC to be used and modified for 407
enhancements sought for online classifying 249
for organization of knowledge fields 33-42
historical development 33-
problem areas 30
schedules linked to LCSH 31
terminology needs improvement of wording of captions 113
to organise WWW and other Internet resources 256
used to decrease retrieval items 110-
LCSH Library of Congress Subject Headings, see LCSH
MACK, R.L. ref. 240
Macleod, M. ref. 240
et al 232-3
MacLeish 37
Maegaard, B. ref. 354
Main Classes sequence according to Integr.Level Theory 157
Maitra, R. ref. 86
Mann, Th. 99, ref. 105
Mao, C. ref. 64
Mapping of MeSH and Chinese Medical SH Fig. 53
Marchetti, P.G. ref, 93
Marco, see Garcia Marco, F.J.
Mariner 16
Markey, K. 95-, ref, 32, 105, 117, 253
Martel, Charles 35
Martin, G. 126
Martin, H.-J. ref, 327
Massicotte, M. 107, ref, 117
Matheson, N.W, ref, 300
Matsuo, T. paper 336-345
Matthews, W.E. ref, 395
Maunsell, A. 45
Mayhew, D.J. ref, 240
MARBI, A.L.A.'s Machine-
Readable Bibliogr. Inform. Committee 26
MARC authority formats promote interoperability 139
McAnthur, T. ref, 172
McBryan, O. 261, ref, 267
McConnell, F. 303, ref, 311
McCray, A.T. 81, 84, ref, 86
McDonald, P. ref, 260
McGiH, M.J. ref, 267
McIlwaine, I.C. 370, paper 122-136
ref, 127, 150, 377
McInnis, R. 295, ref, 300
McKieman, G. paper 254-260, ref, 260
et al 256
Meadows, A.J. ref, 49
Meaning Representation Language (MRL) for a natural language
processing system 329
Medical information transfer, concept vs. word-based
methods, Chinese 59-64
Language, Unified in diverse cultural
environment 51-58
Science, MeSH-based Chinese thesaurus of 52-
Subject Headings (MeSH), not sufficient for HIV/
AIDS 294-5
titles, Chinese-English translation of 59-64
Thesaurus, Military, in Chinese 52-
Medicine Chinese, subject headings of 52-
highly nominalized sublanguage 84
Mehlenbacher, B. 236,
ref, 240
Melby, A.K. ref, 173
Mellish, C. ref, 334
Menex 66, 87
MenUSE Menu-based User Search Engine 231
Merigan, T.C. ref, 300
MeSH-based Chinese Thesaurus 52-
MeSH logical model for arrangement of HIV/
AIDS vocabulary 296
Metadata Manager of the World Bank's Inform. Admin.Service 203-
Metathesaurus requirements of an interdisciplinary 273-274
Metropolitan Museum of Art Classification System 184
Meyer, I. ref, 353, et al 347
MEDLINE database, accessed by MenUSE 231
indexing practice on Chinese herbal medicine 54
METIS functions of, fig. 339
METallurgy papers Intelligent Surveyor 336-
Michal-Johnson, P. ref, 300, et al 294
Miksa, F. paper 406-412
Military Medical Thesaurus in Chinese 52-
Miller, E. ref, 188
Miller, W.L. 47, ref, 50
Milstead, J. 302, ref, 311
Minaret PC-based software, used to create MARC records 29
software for LCC, plans to upgrade 31
Mitchell, J. 252, 408
DDC new uses 249
ref, 253, 385, 395
paper 378-385
Mittermeyer, D. ref, 222
MIRLYN retrievals 110
Mnemonic aids, introduced by Dewey 406
Mnemonics of ICC 160
Model schedule development in common law 38
Modeling, formal of the resource to computerize knowledge resources 208
Mokros, H. ref, 230
Molholt, P. 80, paper 65-71, ref, 86, et al 82
Montgomery, C. ref, 50, et al 47
Mooers, C.N. 363, ref, 369
Mosey, D.L. ref, 86
Morgan Stanley Trust Company, Case study on who needs classification 209
Morton Grove Online WEBrary (DDC), example of access to collections in the Internet 249
Mosaic's graphical user interface, extended to online catalog interface 113
Mowery, R.L. 302, ref, 312
Multi-dimensional analysis applied by METIS 341
Murray, P.C. 21, paper 206-212
Mustafa Elhadi, W. paper 346-353, ref, 354
MUSIC, Metrics for Usability Standards in Computing 232
N
Nagao, M. ref, 63
National Library of Medicine (NLM) 51, 81
National Research Institute for Metals in Japan 336
Natural Language Processing Systems, to build terminological databases 346-353
NASA Thesaurus 369
Navarro see Esteban Navarro
Ndi-Kimbi, A. 369
Neeleman, A. ref, 86, et al 80
Negrini, G. 277, 282, paper 275-283, ref, 283, 292
Nelson, S.J. ref, 86
NetFirst Database of OCLC with DDC numbers 249
Record, example of, fig. 250
Networked computers impact on society 320-327
Now, G.R. paper 386-395, ref, 385, 395
Nielsen, J. 235, ref, 240, et al 235
Niki, K. ref, 345, et al 344
Nippon Cataloging Rules 48
Nishida, T. 262,
paper 261-267, 336-345
ref. 267
NISS Information Gateway
(UDC), example to access
collections in the Internet
249
Node identification and link
building, automatic 89
Nodell, M. paper 198-205
Nohr, H. ref. 222
Nonaka, I. 213, ref. 215
Norman, D.A. 235, ref. 240
Notation
for a link 70
should be expressive in
online environment 124
Noun phrases
construction of 332
play a fixed role with fixed
relation 329
Novak, J.D. ref. 94
Novikoff, A.B. 157, 162,
ref. 164
N O EM IC
based on a componential
analysis model 349-
Number building in DDC
complicated 402
Nyc, A. 303, ref. 312
O
O'Hara Conway, M. ref. 128
Object areas main classes of
ICC 157
Obsolescence of knowledge
388
Ockerse, T. ref. 181, et al 176
OCLC Online Computer
Library Center 171, 248
OCLC Online Union Catalog
(OLUC)
includes LCSH mapped to
Dewey numbers 383
OECD Macrothesaurus as
elementary example 142
Ogonowski, A. ref. 354,
et al 349
Olshain, E. 59, ref. 64
Olson, H.A. paper 302-312
Olson, N.B. 95, ref. 105
Online access to save time for
the classifier 27
Online classification
potential uses for 26-
implications for classifying
and retrieval 249-253
Ontological categories
distinct from logical
categories 316
Ontologies,
formal, vs. systems for
knowledge organization
18
weakly structured, example,
fig. 269
Ontology
for Knowledge Organization
313-319
four general theses on 313-
314
relationship between AI and
313
Ontology-based information
capturing
from the Internet 261-
267
Open Text Search, search
engine in the WWW 242
OPAC user studies
basis for present system
design 95
Ordering of terminology
like ordering of knowledge
355
Osborne, J.E. 299, ref. 300
Otal, J.L. ref. 94, et al 89
Otlel and LaFontaine
used Divide-Like Rule of
Dewey 123
Otel, P. 407, 410
P
Pacey, Ph. 303, ref. 312
Packer, H. ref. 240
Palmer, C. 95-, ref. 105
Palmer, J. ref. 240
Panosky, E. 189, ref. 197
Pappano, L. ref. 173
Parallel subdivision
causing a lot of trouble today
123
Parker, E.B. 192, 195,
ref. 197
Parser, bottom-up
used for Prometheus 329-
Paton, N.W. 239, ref. 240
Pattern adjustment
LCC Class K, Law 40
Patterson, J. ref. 180
PATR notation 330
Peijtersen, A.M. 179, 196,
ref. 181, 197
Penel, D. 353, ref. 354
Peritz, B.C. 46, ref. 50
Pesce, M. 259, ref. 260
Peticot, J. 314, ref. 319
Phenomenology as bases for
change oriented KO 21
Piatetsky-Shapiro, G. 19,
ref. 24
Picht, H. 356, 359, ref. 362
Pictograph
visual image of an object
referring to the object 175
Pierce, S.J. 295, ref. 300
Pint 
Pinakes of Callimachus,
Alexandria library 43
Pinkerton, B. 261, ref. 267
Pinin romanization system 56
Paternick, A.B. 47, ref. 50
PLCPC Subject Guide 108
Poli, R. paper 313-319,
ref. 319 et al 313
Poliziano 156
Pollard, A.E.C. 410
Pollard, R. ref. 94
Pollitt, A.S. 231, ref. 240,
paper 231-240
Polson, P.G. ref. 241
Popper, K. 213, 296,
ref. 215, 300
Porter, M.F. 75, ref. 79
Post-modern library,
DDC and 411
Poststructuralism, def. 304
POPSI
Postulate-based Permutated
Subject Indexing 330
Prabha, Ch. 107, ref. 117
Prasad, A.R.D. paper 329-335,
ref. 334
Pratt, G.F. 294, ref. 300
Predictability, needed for a
workable order 125
Predictable arrangement for
compound concepts 125
Preferred terms, choice of, fig.
373
Prebish, C.I. ref. 49
Principles for HIV/AIDS
vocabulary 296-
-Probyn, E. 307, ref 312
Procedures terms in MeSH 82
Prometheus
an automatic indexing
system 329-335
Properties, figurial of an
effective icon 178
Prytheach, R. 147, ref. 150
PRECIS 88, 122
Putnam, Herbert 35, 41
Q
Qiu, X. ref. 64
Quorum searching 231

R
Ragucci, M. paper 284-292
Ranganathan school of thought 330
Ranganathan's five laws of library science to be extended 411
Ranganathan, S.R. 123, 410, ref. 164, 334
mapping of basic classif. and depth schedules CC 126
Rao, K.S. ref. 63
RABBIT, an intelligent database assistant 91
RCC, Ranganathan's Colon Classification 155-163
Reality, levels of 315-316
Recategorization 387
ruled out at LC 38
Reed, E.S. ref, 223
Reimet, U. ref, 94, et al 90
Relational inventory organizational structure within 76-78
Relational structures inventory of basic 75-
Relational thesaurus development of 72-79
Relationships see also Links binary, of prepositions and conjunctions 74
degree of, def. 73
in thesaurus, desiderata 72
internal, of conceptual entities 73
non-hierarchical associative in MeSH 80-86
procedures 83
standardization of 65-71
Relative index, of DDC numerous entries for English outside USA 381-2
Relativity in terminology, hypothesis of 359
Renger, R. ref. 240
Repertorium bibliographicum (Hain) 321
Restrictors speed up searches 398
special indexes, allow a search limited to a specific main class 398
Retrievability enhanced by a multilingual thesaurus 149
Retrieval, seen as a multi-step process 243
Retrieval increase by use of classification numbers, Fig. 114
Retrievals, too many, too few 107-118
Reverse engineering of knowledge resources 208
Revision, needed in parts of all traditional schemes 123
Rhyno, A.W. 259
Rice, R.E. ref. 260
Richardson, E.C. 410
Richmond, Ph. 95, ref. 106
Riegler, M. 44-, 48, ref. 50
Rieman, J. ref. 241
Riesthuis, G. 149, 371, ref. 150, 377
Rieu, A.-M. 21, ref. 25
Riley, C. 59, ref. 64
Rivlin, E. ref. 93
Rodriguez, D.R. ref. 106
Roget's thesaurus verb groupings, used for relation research 75
Roget, P. ref. 79
Romanization systems variants of existing 56
Ronner, D.M. ref. 188
Rooks, D. 259-, ref. 260
Rose, G. ref. 311
Rosenberg, S.A. ref. 300
Rosenberg, V. 99, ref. 106
Rossney, R. 325, ref. 328
Rosigard, F. 45
Roulin, C. ref. 58
Rowley, J.E. 174, 59, ref. 64, 181
ROOTThesaurus (BSI) 370
Rueter, H. ref. 181
Rui de Mondoza, F.J. ref. 94
Russian National Library for Sci.& Technol. made translation of DDC21 384
S
Sager, J. ref. 223, 354, 369, et al 364
Sailer, A. ref. 94, et al 90
Salton, G. et al 263, ref. 267
Sando, M.A. ref. 300
Sapir-Whorf hypothesis 359
Saracevic, T. 224, ref. 230
Sayers, W.C.Berwick 305, ref. 312
SAM ScriptApplier Mechanism 344
SANe. The Skill Acquisition Network, a computational cognitive architecture 233
Schank, R. 330, ref. 334
Schatz, B. ref. 106, et al 99
Schedule of DDC, displayed through Dewey for Windows 401
Schipper, J.D. 10, ref. 25
Schmitz-Esser, W. ref. 94
Schneider-Hufschmidt, M. ref. 23
Schreinemakers, J. paper 213-215
Science Library Catalog Project 259
Scientific research as autonomous field, structur of 276-281
Scott, J.W. 308, ref. 312
Screen display of ICC 160Swit
SCI-MAP Software System 23
SCORPIO a browse call command 104
Search terms selection of 224-230
sources of 225
Search window of DDC 396
See also reference, no rules 67
Selection of search terms a meeting place of different discourses, fig. 229
Self, P.C. ref. 300, et al 294, 297
Self-organizing knowledge resources, formal meta-principles for 212
Self-organizing teams for new organization, concepts of companies 213-4
SemWeb development of 171
System for Integrated Access to Knowledge 165-173
Template, frame slots on concepts and terms 166-168
Senez, D. ref. 173
Sengupta, I.N. ref. 301, et al 294, 297
Seymour, M. 407
SEEK 349-
SGML Standard Generalized Markup Language 20
Sharoff, S. 21, ref. 25
Shatford Layne, S. ref. 174, 181, 182, 188-9, 197
Shaw, D. paper 174-180
Shelflist online 28
Shneiderman, B. 232, ref. 93, 241
Shoham, S. paper 43-50
Shreve, G. paper 216-222, ref. 223
Shute, St.J. ref. 106
Signs constitute a one-to-one mapping of s.th., 174 175
Simandl, D. ref. 163
Simons, P. ref. 319
Simplicity trap of in enterprises of ontological nature 316
Siochi, A.C. ref. 240
SI Art Image Browser 182-188 evaluation of 186
Small, J.P. 196, ref. 197
SmartMarks 16
Smith, B. ref. 319
Smith, M.P. paper 231-240, ref. 240 et al 104
Smith, PhJ. ref. 106
Smith, R.A. 326, ref. 328
Soergel, D. 67, 80, paper 165-173, ref. 71, 86
Sonnenwald, D. 229, ref. 230
Sowa, J. 313, ref. 319
Soyibo, A. ref. 301, et al 296
Spink, A. ref. 230
Spivack, N. ref. 25
SR Explain service (SR: Search and Retrieval) 140
Stability of a classification system 119
Standard Citation Order needs to be used universally 123
Standardization, needed for indexing purposes 40
Standardized definition creation of, fig. 368 structure of, fig. 365
Stein, Gertrude 19
Steinberg, F. 303, ref. 312
Steinberg, S.H. 44-, ref. 50
Steyer, H.C. paper 198-205
Stotts, P.D. ref. 94, et al 91
Strehlow, R.A. 365, ref. 173, 369
Strout, R.F. 43, ref. 50
Structural relationship needed to construct a concept system 285
Structured notation to permit searching on one symbol for one concept 125
Structured vocabulary Browse (SVB) 139-
Structured Vocabulary Scan for future subject access in the Internet 155-164 mechanism establishment, organizational problems 161-
process implementation with ICC 161
System, principles of a broad - 156-
Symbolic languages an integrated set of graphic images 175 ability to represent hierarchical structure 179 as classificatory structures 178
Symbolic of graphic images to represent classific structure 179 of a visual languages, refers to the level of representativeness 175 Symbols, are ideograms with abstract referents 175 Synonyms, in descriptions of UDC notations 148 Syntactic categories identification of, in Prometheus 331 Syntax of Prometheus 330 Synthesized numbers classifier to trace formation of 27 in the index for compound concepts or terms 123-4 Synthetic devices to be built into classification schemes 125 Systematifier application in several areas 275-283 applied in cereals science 284-292 nine facets of 158 Systems comparison diagram 351 T Table entries display in Dewey for Windows 401 Takawashi, T. ref. 50, et al 48 paper 261-267 Takeda, H. ref. 2678 Talmudical books titles as surrogates for authors names 48 Talmy, L. 314, ref. 319 Tanaka, K. ref. 345 Taxonomy of visual representations 178 TAIGA 353 Term Banks 347 construction, automatic 350 structure, problems of traditional related 67 Terminological change, levels of with reasons for 356-
control for better hypertext systems 89
databases, built by natural language processing systems 346-353
definitions, domain dependent 363
definitions for indexing and retrieval thesauri 363-369
thesauri 363
Terminology
of anatomy in MeSH 80-86
standardization and updating of, LCC 40
building tools, state of the art 347
dynamics, types of 358
management, def. 355
Organization and change 355-361
problems of caption headings for end users 252
source for UDC thesaurus, medicine 372
Terms
clarification of, fig. 373
sources of search -, fig. 226-228
Text categorization using an ontology, Fig. 271
TEAM 347
Thesauri
definitions in 363 and hypertext have precedents in encyclopedias and dictionaries 89
Thesaurification, multilingual 148-
Thesaurofacet 370
Thesaurus
added facility for UDC 127
construction, a practical manual 371
derivation from a restructured UDC 370-377
iconographique 191- in information retrieval 138
multilingual, based on UDC 144-154
of Educational Descriptors, used to verify completeness for HIV/AIDS vocabulary 298
of Graphic Materials, LC 191-
relationships are n-ary 73
source of 374-
Thesaurus-aided searching 137-143
Thomas, P. ref. 354
Tillett, B. ref. 253
access control model 252
Title informativity increase in 20th century 46
Title page, established by 1500 45
Titles
context, cultural and technological changes 43-50
from the Bible 48
of medical information articles, English-Chinese, Chinese-English translation 59
use in information retrieval, history of 43-50
Tocatlian, J.J. 47, ref. 50
Tocqueville, A.de 308, ref. 312
Todd, P. ref. 181
Tokunaga, T. ref. 79
Tolle, J.E. 107, ref. 117
Tou, F.N. ref. 94, et al 91
Traut, M. ref. 181
Travis, I.L. paper 198-205
Trefler, Florian 45
Treglown, M. 236, paper 231-240, ref. 241
Trotter, R 396, ref. 404
Truncation and mask symbols to be used via the Search window of DDC 396
Tsuiji, J.-K. ref. 63
Tsutomu Sh. ref. 50
Turner, B.S. ref. 230
Turner, J.M. 189, ref. 197
Tyckoson, D.A. 255, ref. 260

U
Udell, J. 273, ref. 274
UDC 155-
as an online browsing tool for subject access 97
based multilingual thesaurus 144-154
Class 61, Medical sciences 370-restructured 371
Class 8, experiments with 149
Class 9, revision 124
Consortium of publishers 127
Descriptor group display, example 153
French Medium Edition 127
maintenance 122-136
Master Reference File 127
notations, source for terms and scope notes 148
restructured, thesaurus derivation from 370-377
revisions, fully faceted 123
revision work, a compromise at best 376
structure, to be mapped on the Dewey base 126
Thesaurus, alphabetical example 151
used for access in the Internet 249
Ullman, J.D. ref. 353
UMLS
Unified Medical Language System 51-, 81, ref.173
Knowledge Sources
Documentation ref. 86
system for the names of herbs needed 55
to build knowledge base, for data about herbs 58
Unified
Agricultural Thesaurus UAT 291
Medical Language System see UMLS
Universal Decimal Classification, see UDC
language to overcome individual diversity 304
Subdivisions, used for standard subdivisions 297
University
of Bristol, social science network, use of UDC 123
of Leeds, Computer Science Dept., use of UDC online 123
of Michigan LCC catalog, used for searches 108-109
of Surrey, team to build term banks 350
UNESCO ref. 164
UNISIST
Guidelines for the Establishment and Development of Multilingual Thesauri 148-
Report 156
Urr, C. 89, ref. 94
Usability problems regarding HIBROWSE 236
Use of titles, history of 43-50
User and public knowledge organization, interaction with user profiles, views, and plans 216-222
education plan (UEP), an educational package 220
profile, components of, fig. 218
view, a projection of user profile, fig. 219
User-thesaurus interaction 137
Users participation in all stages of knowledge work 20
USMARC Classification Format, structure familiar to other USMARC formats 29
for Authority Data ref. 173
for Classification 40
for Classification Data, 18, 26, ref. 173
adaption of DDC to 383

V
Van Dijk, H. ref. 181
Van Gelder, T. ref. 25
Van Tassel, D.D. ref. 41
Variation, accommodated by DDC 307
Veltman, R. ref. 63
Verbs, relational thesaurus elements, structure of 73-
Veronica 273
Vickery, A. ref. 335
Vickery, B.C. et al 330
ref. 128, 335
on notation 124
View-based searching, language of 237
utilises thesaurus displays 231-2

Viewpoints
Dewey’s suggestions on 308
Virtual hypertext systems in the Internet 90
Vizine-Goetz, D.
paper 249-253, ref. 253, 385, 405, et al 396
Vocabulary controlled, for HIV/AIDS 293-301

displays, need for better 97

and hypertext browsers, need for better interaction 98
for HIV/AIDS, principles articulated 296-
selection mechanisms, movement between retrieval and 104
Volberding, P.A. ref. 300
VUBIS System 146

W
Wade, St.J. paper 231-240
Wade-Giles Romanized system 56
Wagman, F.H. 37
Wahl, W.S. paper 198-205
Wajenberg, A. 95, ref. 106
Walker, N. ref. 181
Walker, St. 107, ref. 117,
et al 108, 114
Walker, Th.D. paper 320-327
Wall, R.A. 98, ref. 106
Walker, V.A. ref. 259
Wang, G. ref. 64
Warren, M. 390
Watters, C. ref. 94
WALS 273
Web dialectics and perceptions of knowledge 322-
WebCrawler
search engine in WWW 242
tool to search for info on the Internet 261
Weibel, St. ref. 188, et al 187
Wells, H.G. 325, ref. 328
Welsh, W.J. ref. 42
Westland, J.C. 89, ref. 94
Weyuker, E. ref. 362
Wharton, C. ref. 241,
et al 235-6
Whitaker, R. 20, ref. 25
Whitten, D. ref. 25
Whole, the Whole as such 317-318
Whorf, B.L. 359
Wiberley, St.E. Jr. 107,
et al 118
Wiesacker, F. 37, ref. 41
Wille, R. ref. 94
graphical representation of concept maps 92
Williamson, N. 127, 370
paper 370-377
study of LCC 30
ref. 32
Witten, W. ref. 180
Wittgenstein, L. 214, ref. 215
Wolf, St. 303, ref. 312
Wolf, Christian 37

Word identifiers fig. 170
WordNet ref. 173
World Bank
Information Management Architecture 198-205
Information Engineering Unit ref. 205
World Encyclopaedia 326-
World Wide Web
see also WWW
address of database of an image system 183
brilliant example for nets of hypertext documents 90
search engines 242
uses icons 180
search tools, study on by E. Yoder 206
Wright, S.E. ref. 173
WWW 171
as a world brain 325
information gathering, example, fig. 270
uses of DDC 382
Virtual Library, access in a -
Worm, tool to search for info on the Internet 261
Wyllie, J. ref. 94
Wynne, B. ref. 41

X
Xerox Corporation
attaches business attributes to information 212
Xiao, P.G. 54, ref. 58
Xiao, Y. ref. 223

Y
Yahoo 324
popular Internet search service 249
search engine in WWW 242
Yin, B. ref. 64, et al 59
Yitzhaki, M. 46, paper 43-50,
et al 50
Yoder, E. 206, ref. 212

Z
Zeng, M. ref. 64,
paper 51-58
Zensei, O. ref. 50
Zhang, M. ref. 64, et al 59
Zhao, T. ref. 64
Zhou, M. ref. 64
Zimmermann, H. ref. 173
Recent publications from INDEKS Verlag

ADVANCES IN KNOWLEDGE ORGANIZATION
ISSN 0938-5495, an irregular series, published by the
International Society for Knowledge Organization (ISKO)

60 papers on topics such as online retrieval and retrieval technologies, knowledge organization by universal and special systems, thesaurus issues, algorithmic text analysis, indexing, hypermedia, etc.

20 papers on construction and programming of documentary languages, classification systems and their use in databases, and natural languages and automatic indexing.

53 papers arranged according to the conference program in 12 sections covering: Quality in Knowledge Organization (KO), Theory of Knowledge, Linguistics in KO, Concept representation in systems design, KO in specific domains, Communication and KO, OPACs, Knowledge-based systems, Tools and techniques for KO, Thesauri facing new technologies, Restructuring classification schemes and thesauri, Future prospects. - The vol.includes an extensive name and subject index and a list of the authors.

KNOWLEDGE ORGANIZATION IN SUBJECT AREAS (KOSA)
ISSN 0946-9389, irregular, published by ISKO

Contains 34 papers and 7 abstracts as well as the Recommendations as the result from the papers of the conference and the discussions during the final session. There is also a list of participants as well as a name and subject index in each volume.

***

TEXTBOOKS FOR KNOWLEDGE ORGANIZATION
ISSN 0944-8152, irregular series

Synthesis of lectures held at the University of Indiana, Bloomington, School of Library and Information Science in the summers of 1992 and 1993. In 6 chapters the following topics are treated: Information and information systems, Information system survival power, Theoretical considerations on information storage and retrieval, Indexing, Technology of information supply. Glossary of terms used and a systematic and a "basic index" to the book.

The book is about knowledge structures as found in information handling tools, classification and indexing systems, as represented in the human mind and in Artificial Intelligence systems. It looks at concepts and relationships in these structures.

INDEKS VERLAG, Woogstr. 36a, D-60431 Frankfurt
Tel.: 069-52 36 90, Fax: 069-52 05 66
Payments by VISA, DINERS Club, MASTERCARD or EUROCARD are accepted
Advances in Knowledge Organization (AKO)