Facet Analytical Theory as a Basis for a Knowledge Organization Tool in a Subject Portal

Abstract: The paper examines the way in which classification schemes can be applied to the organization of digital resources. The case is argued for the particular suitability of schemes based on faceted principles for the organization of complex digital objects. Details are given of a co-operative project between the School of Library Archive & Information Studies, University College London, and the United Kingdom Higher Education gateways Arts and Humanities Data Service and Humbul, in which a faceted knowledge structure is being developed for the indexing and display of digital materials within a new combined humanities portal.

1. Introduction and background to the project

The work described in this paper is currently being carried out at the School of Library, Archive & Information Studies at University College London (SLAIS), under a grant from the United Kingdom Arts and Humanities Research Board. Partners in the research with the School are the Arts and Humanities Data Service (www.ahds.ac.uk) and Humbul (www.humbul.ac.uk) two large government funded subject gateways for the humanities, serving the UK higher education community. The Arts and Humanities Data Service (AHDS) comprises several distinct digital collections, including archaeology and the visual and performing arts; Humbul has a more conventional humanities content with strong collections in history, philosophy, theology, literature and the classics. The two gateways are soon to merge into a single Humanities portal.

The object of the research is to develop a subject tool for the management of the new portal with two functions envisaged. Firstly the system will be used to organize the front end of the portal, and will provide a structure for the first point of entry to the site. A directory style layout will be provided with subject headings developed from the system, together with browsable indexes derived from the inversion of the subject headings. The knowledge tool has the additional function of providing a vocabulary for use in subject metadata.

Currently the two gateways create catalogue data for all of the digital objects in their collections using the Dublin Core. For the subject field (and for organization of the display of resources) Humbul uses the Dewey Decimal classification, and AHDS Library of Congress Subject Headings. Neither of these systems have proved entirely satisfactory, and for the combined portal both organizations decided that they wanted to explore the possibilities of a new and more satisfactory indexing tool, logically structured and based on modern theories of subject organization, and designed with the indexing needs of digital materials specifically in mind.
Although it might be argued that the problem of identifying the semantic content of an information carrier does not differ with the nature of that carrier, nevertheless there are some distinctive features of digital materials that do affect the indexing process. Firstly the intellectual content of the materials can be very complex, and the level of indexing is required to be high, in terms of both the specificity and exhaustivity. The fairly broad level of subject description provided by the conventional systems currently used has created difficulties, particularly for AHDS which consists of several discrete subject oriented databases. Where there are only one or two levels of hierarchy available for index description the resulting lack of specificity often means that the item will not be retrieved by a searcher starting from the context of another discipline, since the content relevant to his discipline will not have been identified and tagged. For example, a resource dealing with nineteenth century political cartoons deposited in the Visual Arts Data Service database will not be retrieved by a searcher with an interest in nineteenth century political history if the item has only been described as ‘visual arts – cartoons – nineteenth century’. This cross discipline searching is a particular feature of AHDS which will migrate to the combined portal, where it is envisaged that the problem can only become magnified.

A second indexing problem related to the digital nature of the materials is the need to identify non-semantic properties of the objects i.e. properties of electronic format, media, and so on, since these may well be sought terms.

There is also a desire to exploit the major tool for ‘cross-referencing’ in a digital environment – hypertext. Hypertext is able to facilitate retrieval in at least two ways. Firstly within the visual display where it can be used to expose successive layers of a hierarchy (see below) maintaining the detail of the developed structure without exposing too much of it to the user at once. Secondly it supports the notion of multiple points of access to the network of resources and reinforces the regularity of the structure of the system.

During discussions with staff at SLAIS it was decided that a new indexing tool, built on faceted classification principles might meet the specific indexing needs of the situation better than the conventional systems currently being employed.

2. The nature of classificatory structures

At the end of the nineteenth century the conventional bibliographic or library classification scheme was the first means of organization of, and thus access to, information on a subject basis. Since the advent of the web we have seen a number of cases where traditional bibliographic classification schemes have been employed for the organization of web resources, usually within organized gateways or digital libraries, and usually within the academic sector. The Universal Decimal Classification is probably the most widely used, closely followed by the Dewey Decimal Classification, with Library of Congress Subject Headings and various national classification systems also represented. The recent HILT (HILT, 2001) project, which investigated means of subject access to digital materials in the UK information sector, worked with forty-two stakeholders representing libraries, archives, museums and electronic information providers; of these, the majority were using locally created systems. There is thus a proliferation of systems of subject organization on the web.
Despite this, the classification scheme is often seen as ineffective as an organizer for electronic resources (Koch, 1998). Some obvious reasons for this can be found in the lack of logical structure in older systems of classification. Traditional classification schemes are built on the basis of a tree structure, with the emphasis on downward subdivision into smaller and more specific classes; often, the only relationships acknowledged are those of super- and subordination and there is no provision for syntactic relations. As a result the classification is usually relatively broad and there may be limited facilities for combining between classes, or for expressing complex semantic content. As seen in the example given above, this can cause difficulties when searching in a multi-disciplinary environment, or when dealing with objects of a complex nature.

Classifications built on the facet analytical model work on different principles, starting from the constituent concepts of the subject, rather than from a view of knowledge in its entirety. It is important to establish what precisely we mean by facet analysis, since there are a number of interpretations of this theory, some of them very different from the original work of S. R. Ranganathan. The situation is complicated by the current interest shown by computer scientists in this area, and the frequent use of the term faceted classification to mean any system of subject organization in which analytico-synthesis is used. This notion is also often employed in the world wide web literature on facet analysis (Maple, 1995), and also in the United States usage of the term when compared with the British or European tradition. A recent commentator on types of classification scheme (Koehler, 2000) states that “Universal classification schemes .... tend to classify works by a single characteristic.”, contrasting them with faceted schemes which are considered to express more than one aspect of a subject.

In this paper, and in the accompanying project, facet analysis is interpreted in a much more restricted way; it is taken to mean that rigorous process of terminological analysis whereby the vocabulary of a given subject is organized into facets and arrays, resulting in a complex knowledge structure with both semantic and syntactic relationships clearly delineated.

3. Applications of faceted classifications

Classifications of this type provide effective tools for vocabulary management, document description and retrieval. Facet analysis provides a method which is in theory appropriate for the management of terminologies and concepts in a variety of environments, although its applications have so far mainly been limited to the conventional print based document collection.

As an ordering device for initial access to resources in a managed site the faceted classification has some advantages over traditional schemes. The predictable format of the classification is usually very evident to the casual searcher, and the structure can also be used to generate subject headings and a browsable index, as alternatives to, or in support of, a directory style front page.

Although primarily used for the physical organization of print based materials the potential of facet analysis for the management of documents in an automated environment has already been recognized. More than ten years ago Godert (Godert, 1991) and Ingwersen and Wormell (Ingwersen & Wormell, 1992) testified to their performance when used in conjunction with online databases for both document description and the framing of queries. Ingwersen and Wormell...
(Ingwersen & Wonnell 1992, p.199) were able to state with confidence that “... the discussion demonstrates the suitability of the faceted categorization, not only for textual documents, but also with other forms of carriers of information. Faceted categorization may provide multi-dimensional and hence structured entry points to document contents, and thus give intellectual access to generated and stored knowledge.” More recently the specific application of facet analysis to searching the World Wide Web has been discussed by Ellis and Vasconcelos (Ellis & Vasconcelos, 2000) who conclude that “... because ... the classification is derived inductively from the concepts or terms used in the subject field, it can alleviate some problems in searching the World Wide Web by being applied to using subject directories or search engines.”

4. Facet analytical methodology

The system currently under development at UCL is built on those classificatory principles developed by the Classification Research Group in the mid twentieth century out of the original work of S. R. Ranganathan. The internal logic of the system is based on rigorous analysis of vocabulary, whereby terms are sorted into a standard set of functional categories. Within these categories a range of semantic relations are acknowledged, and problems of vocabulary control addressed. A sophisticated system syntax provides for the ordering and combination of terms both intra- and inter-category. There is consequent improved performance in the accommodation of complex subjects, the predictability of location, and in the effectiveness of retrieval. This methodology has been brought to a high level of sophistication in the second edition of the Bliss Bibliographic Classification (Mills & Broughton, 1977 - ) where it has been used to create a new general scheme of classification (although built on the infrastructure of the original scheme (Bliss, 1940-1953)).

Taking the BC2 methodology as a standard, the new system begins from the literary warrant of the objects to be classified. The digital objects within the humanities collections are used to establish the terminology to be worked on. A set of standard categories is used for analysis which follows the ‘standard citation order’. These categories are functional in nature; they comprise thing/entity – kind – part – property – material – processes – operations – patient – product – agents – space – time. Although developed initially within the areas of science and technology (so that they resemble a production process), in the process of the BC2 work these standard categories have been applied successfully to the vocabulary of subject fields in almost all areas of knowledge. As one moves across the spectrum of disciplines from science to the arts it is true that the number of these categories that are used becomes smaller; notions such as materials, products and agents are less often found in the humanities. It is also true that some domain specific categories can occur. Commonly recognised examples include those of genre and form in literature and music. We expect that new categories can be determined and that the techniques of facet analysis can be applied in new ways and to new material. It is the methodology that is here being tested, and not any specific structure that has been created in the past.

Once the terms have been sorted into categories, further analysis structures each category and clusters the terms into arrays or groups of terms which share a common property or characteristic (known as the characteristic of division).
Various principles of ordering are used to determine the sequence of arrays within the category, and the collocation of terms within the array; chronological, developmental, spatial and physical contiguity, and the commonly used classificatory principles of increasing concreteness, increasing speciality, and increasing complexity. The organization of the categories or facets into a single sequence normally follows standard citation order i.e. the order given above, but inverted; this brings the more concrete facets to the end of the sequence, and facilitates the location of compounded terms.

The classificatory structure built in this way will consist of single simple terms arranged in a systematic way according to the logic of the system – a pure faceted structure in Ranganathan's understanding. In practice we find that the terminology does not always lend itself to absolute reduction of this kind. There are constant occurrences of terms which represent compounds and combinations of all sorts, but which require to be included in the schedules (and perhaps more significantly in the index) because of their status within the literature and their existence as sought concepts. Consequently the classification consists of rather more than the ‘skeleton’ of simple concepts in mutually exclusive sets. As examples of compound concepts are used to populate the classes the structure grows in complexity. Where, as is usually the case, the compound represents the intersection of two or more categories, the system syntax (or rules of combination) control the precise location of the compound and a much expanded classification is generated by the interaction of the syntax and the conceptual structure. The basic citation order can be repeated as necessary at any point within the classification, and subjects of considerable complexity can be accommodated in a highly logical, predictable and regular fashion.

A fully developed faceted classification on the model of BC2 can by this means generate a very complex knowledge structure of n-dimensionality and great logical regularity, and with deep levels of hierarchy. The resultant structure can be utilised in a number of ways; as an ordering device, as a source of index terms and subject headings, and can also be converted to a thesaurus.

The principles of categorization implicit in facet analytical theory can be extended to organize any set of properties of objects in any domain, and of the domain itself: The theory is sufficiently well established to allow variation in the classical form, and the compiler of a faceted structure need not feel restricted to the categories and combinatorial rules of standard citation order. Viewed in this way, facet analysis should be thought of as a powerful methodological tool rather than a specific arrangement of topics in a given field. Any attributes of objects which are significant for retrieval can be incorporated into the categorical formulation; hence facet analytical theory can accommodate concepts descriptive of a digital environment and the objects in it when building a structure to function as a retrieval tool in that environment. Its capacity to cope with objects with complex combinations of attributes in a logical and predictable fashion permits the creation of structures which reflect the multidimensionality of such objects. There also appears to be potential to exploit hypertext to generate links across such structures, and to increase the points of entry to the semantic network in an innovative manner.
5. Implementation of the knowledge tool and operational testing of the system

As we have seen above, the effectiveness of browsing across a series of semantically discrete databases is important to the success of the combined portal; therefore experiments will be carried out to test its use in cross-disciplinary browsing and retrieval of digital resources within the digital collections. In a test-bed implementation for the research, AHDS and Humbul are applying the knowledge structure to the Portal’s planned metadata repository for all the digital objects in their collection. Initially a pilot project will build a structure for the discipline of history and test its effectiveness over this part of the collection. This area has been chosen because of the pervasiveness of the discipline, and its particular suitability for testing cross-disciplinary searching and retrieval. Extensible markup language (XML) will be the vehicle for implementing the knowledge structure, and the potential of this combination of the knowledge structure and the markup language seems to be very considerable.

6. Markup languages and semantic content

It is certainly true that, to date, markup languages have not been utilised to index the intellectual content of documents. Hypertext markup language is engaged only with the format and display of documents. Extensible markup language has taken things a stage further in considering document structure and provides more information to the searcher in terms of the constituent parts of the document, although from the viewpoint of traditional retrieval this leaves us only with the equivalent of the author-title catalogue organized by a descriptive cataloguing standard. The problems of intellectual content of the object have yet to be addressed.

There is currently much interest in the development of the ‘Semantic web’. In an article in *Scientific American* Tim Berners-Lee (Berners-Lee, 2001) looks at how the emphasis in web searching might be switched from a provenance based way of handling documents to one linked to the semantic content; in this new approach to document analysis the ontology has an essential role to play, and it is clear that Berners-Lee’s understanding of the ontology bears a close relationship to the kind of structure that we are developing at SLAIS. He says (Berners-Lee, 2001 p 34) “The most typical kind of ontology for the Web has a taxonomy and a set of inference rules. The taxonomy defines objects of classes and relations among them …..Classes, subclasses and relations among entities are a very powerful tool for Web use...Inference rules in ontologies supply further power.” He also describes (Berners-Lee, 2001 p 37) what is essentially the process of mapping and of the use of a classification with a notation to enable information exchange “An essential process is the joining together of subcultures when a wider common language is needed. … The relations allow communication and collaboration even where the commonality of concept has not yet led to a commonality of terms.”

The capacity of the classification scheme or controlled indexing language to perform this function of intermediary in the exchange of information across languages and cultures is well represented in the professional literature. We are also clear that the use of a faceted structure has advantages in terms of its efficiency as a retrieval tool, and that it has the potential to handle complex digital objects and facilitate cross-disciplinary searching. It remains to be seen whether the
implementation of a system with these properties using XML might provide us with a tool with far wider application.

Reference


Godert, Winfried (1991) Facet classification in online retrieval International Classification 18(2) 98-109


Ingwersen, Peter and Irene Wormell (1992) Ranganathan in the perspective of advanced information retrieval Libri 42 184-201

Koch, T. (1998) The role of classification schemes in Internet resource description and discovery; work package 3 of Telematics for Research project DESIRE (RE1004) www.ukoln.ac.uk/metadata/desire/classification (Accessed 29.08.01)


Mills, J. and Vanda Broughton Bliss Bibliographic Classification 2nd edition London; Butterworth and Bowker-Saur 1977