Authorial Index Browsing in an XML Digital Library

Abstract: In this paper, we present the CODEX project which aims at designing a system for ETDs access based on a XML architecture. XML offers a possible delivery format for controlled vocabularies. It can also be used as an alternative to offer a single interface to the different controlled vocabularies and classification systems. We use these XML facilities to implement a system based on the authorial indexes as a browsing tool.

1. Introduction

The widespread use of the Web among the scholarship community and the increasing availability of thesis as electronic documents raise new opportunities for a wider dissemination of large scientific documents in electronic form. The true benefit of an ETD (Electronic Thesis and Dissertation) digital library, however, comes not from the replication of traditional library functions, but rather from the ability to make possible tasks that would not be possible outside the electronic environment, such as the hypertextual linking of related texts, full text searching of holdings, and the integration of filtering information tools and data visualisation.

Large corpora of ETDs on the Web also gives the researchers a new environment to consider the question of knowledge classification and organization. As a matter of fact, the question of information structuration within the ETDs and, specifically, the matching process between the document produced by the author and the informational need that may have a reader has been renewed by the use of new techniques of classification and description.

Diferent strategies are available to access the information contained in ETDs. Classical - even not easy to implement - ones are based on classifications, thesaurus or ontologies which give the user (i.e., the reader) keys to go in the informational nodules at different levels of granularity. These strategies share the fact that they all are pre-defined views of a knowledge domain. Pre-defined views are also tools which are all conceived by library and information science professionals who share some common concerns - at least practical if not cultural - about what is a representation. In this context, readers need to access the information in the documents via these pre-defined representation tools.

Another way to consider the matching process between the document and the reader is to examine what happens when using the author-made index instead of the pre-defined representation tools. The author abstract and index (or indexes) may be considered as content representation tools with specific features. Authorial indexes directly classify the ETDs according a personal view (the author one) of the content and give the reader a direct access to the document. One may consider that when indexes are directly produced by the authors, documents are faster available on the Web and at a reduced cost. Finally, and this is the central point for this article we
want to focus on, authorial indexes can be used as devices for dynamic browsing within the documents.

In this paper, we describe the CodeX project *(Consultation et Organisation de Documents Électroniques à parisX)* which aims at studying and evaluating technologies for implementing an XML-based digital library of dissertations. We first introduce and define the term “multiple views” which is the core of our approach and we then present the technical choices we made to develop this user oriented approach. For most users, there are two familiar methods to access the content of the document: the kind of index at the back of the document (often called back-of-book indexes) and the table of content (TOC). In this paper, we focus on *authorial indexes* and we present the CodeX XML-based capabilities which allow the reader to browse the document.

2. Principles

When dealing with large documents, search functions usually do not provide enough information to decide whether or not a given document is relevant. On the other hand, if the user is provided with the entire document (usually in the form of Postscript or PDF documents) he/she is confronted with problems encountered when reading in the online condition.

As it has been pointed out by several studies comparing the reading of paper versus on-line documents in the psychological and ergonomics (Dillon, 1991, 1992), (Hass, 1992) (O’Hara, 1997) (Van Dijk, 1980), documents are not easily readable on a monitor. As a matter of fact, reading in the on-line condition is slow, laborious, and detracted from real conditions of reading. For Landoni (2000) “Table of Contents and Index(es) are essential features. These cannot be simply replaced by search facilities whose complexity makes the reading process more confusing for the user”.

For most of users, the most two familiar access method are those that's been around the longest: the kind of index at the back of a book and table of content (TOC). There are two general classes of indexes: indexes that are intended for human readers to browse (often called *back-of-book* indexes) and indexes that are intended for use with information retrieval software.

They are some work that addresses the problem of automatically creating an index:

- Turner (1997) presents an empirical evaluation of four algorithms for automatically extracting keywords and keyphrases from documents. The four algorithms are compared using five different collections of documents.
- Woods also used phrase analysis in addition to a large knowledge base to organise terms into a concept hierarchy (1997).
- Evans (2001) present Intellindex whict is a software tool that automatically identifies a list of index terms.
- Muñoz (1996) uses an unsupervised learning algorithm to discover two-word keyphrases. The algorithm is based on Adaptive Resonance Theory (ART)
neural networks. Muñoz's algorithm tends to produce a large list of phrases, so it has low precision.

The main feature that distinguishes a back-of-book index from a keyphrase list is length. As Nakagawa (1997) observes, a document is typically assigned keyphrases, but a back-of-book index typically contains index terms. Also, keyphrases are usually intended to cover the whole document, but index terms are intended to cover only a small part of a document. The usability of electronic indexes has been investigated by (Milstead, 1994).

Having an author write, an index is often less expensive. Authors also have a greater understanding of the audience, the vocabulary, and the theory of the material included in the text. In Codex, the authorial index are structured with XML capabilities.

XML offers a possible delivery format for controlled vocabularies. XML can be also used as an alternative to offer a single interface to the different controlled vocabularies and classification systems (Light, 2001). Light (2001) presents some controlled vocabularies encoded in XML. VocML (Vocabulary MarkUp Language) is a DTD under development to support structured representation of a wide range of knowledge organization resources, "including authority files, hierarchical thesauri (including those with polyhierarchies), classification schemes, digital gazetteers, and subject heading lists. Arnold (2001) presents an architecture to place index reference from control vocabularies in XML documents. Luk et al. (2002) provide a survey of current indexing and searching techniques for XML Documents. The indexing techniques described are divided into three groups: flat-file, semistructured, and structured indexing paradigms.

As most electronic thesis present an index (generated automatically by software like Word), Codex create a set of links from the authorial index to the words in the ETD document. Such index is represented in XML. In this approach, the design of the automatic hypertext generation is based on the content-based analysis of the ETDs. This new kind of view does not derive from the logical structure of documents but relies on the textual content. Content browsing applications provide information seekers with access to texts. Structured lists of index terms give a preview of the content of an ETD.

Our feeling is that the benefits of using authorial indexes in a XML Digital Libraries are:

• Less time spent reading: the list of indexes produces a direct access to specific and interrelated passages of the document.
• Less time spent searching: searching with index-terms is a much easier task than doing full-text searching which requires knowledge of queries formulation.
• Better use of an often neglected source of information (user-created index).

3. Technical aspects of the CodeX prototype

During the Word to XML conversion, the position of each user-created index entry in the document logical structure is computed and recorded using XPath expressions. Based on this information, Java servlets running on a Apache server can then build index-based views. Users specify the text they wish to browse; CodeX locates the XML file and associated metadata, determines which portion of the file contains the desired text section, applies appropriate styling rules, and presents the text in HTML form.
Currently, the CodeX system consists of three modules:
- a converter: the converter is a batch tool used to transform our collection of word documents (students thesis) into well-formed XML documents;
- a metadata creation tool: the metadata creation tool is a simple electronic form which is used to generate an TEI-based XML header associated with the documents;
- a visualization module: it uses the concept of "multiple views" to display documents generated by the two other modules in a wide variety of visual forms according to the informational needs of the users. Among the views defined in the first phase of the project (Chaudiron, 2000), there are the following:
  - visualisation of the content table;
  - visualisation of the introduction, conclusion and content table;
  - visualisation of the abstract made by the author;
  - visualisation of the abstract made by the author and the bibliography.

The search and retrieve procedure has the following steps:
- searching metadata database using the web form (see figure 1)
- selecting a particular document
- browsing and visualising the content of selected document (see figure 2)

4. Conclusion and Future Work

The Codex project demonstrates that the XML format can successfully be used to provide electronic access to theses and dissertations. XML provides a standard way for information providers to add custom markup to information-rich documents, in order to publish complex documents in a dynamic way. Codex tools were developed to extract structural and descriptive metadata from these documents and deliver document fragments on demand.

This paper presents also a methodology for indexing collections of ETDs to help users find and browse information more efficiently. Our hypothesis is that browsing ETDs with authorial indexes improves user satisfaction and reduces time of reading. In future, we will compare the three methods for providing index terms (indexers, users, automatic).
Figure 1: Codex search interface

Figure 2: The user browses the structured document entitled "Besoins et usages de résumés" (upper frame) by requesting all occurrences of the user-created index entry "analyse de surface" to be displayed (lower frame).
Notes
http://xml.coverpages.org/vocML.html

References
Arnold, Stephen & Spenla, Brad (2001). Indexing into Controlled Vocabularies with XML,
http://www.infoloom.com/gcaconf/SWEB/TOC/philadelphia99toc.htm".
(available from http://flood.nhm.ac.uk/mcg/talks/rl/).