

Towards an Ontology for Library Modalities

Christopher A. Welty

Vassar College Computer Science Dept.
weltyc@cs.vassar.edu

1 Introduction

The AXIOM project is a collaboration between researchers in knowledge representation (KR), markup languages such as XML, linguistic analysis, text encoding, databases, humanities computing, and other related fields, whose goal is to develop technologies that will enable a large scale digital library. The central theme of the project is the exploitation of “intelligent tools” and “intelligent texts.” The project also includes several large text encoding projects associated with existing libraries that have been diligently encoding “humanities” data over the past ten years.

2 Knowledge Representation for Card Catalog Systems

Among the many places we are applying KR techniques is in the evolution of existing card catalog systems. While improvements to the existing systems are easy to envision, library scientists are deeply motivated to proceed carefully and thoughtfully, and do not make changes lightly.

From a KR perspective, library scientists are in several ways the perfect collaborators. They are probably, as a profession, the first practicing ontologists, as evidenced by the dewey decimal system and other similar classification systems. They are accustomed to taking a principled approach to representing information, and are always concerned about having things

in the right place. They are also accustomed to thinking taxonomically, again as evidenced by the dewey decimal and subsequent systems. They do not, as a result, require any convincing that the right way to extend the card catalog system is to carefully model all the information that will be represented.

There are many KR issues involved in the evolution of card catalog systems, and principle among these is representing the variety of data a digital library may be expected to contain. The present technology centers on things that can be shelved, principally books, and these systems are being severely taxed by the addition of electronic source material. Marked up texts, which offer the potential for truly advanced and expressive queries, are not exploited any more than printed materials.

There are many KR issues involved even in this more focused view of dealing with these broadened data types, see [Welty, 1994], [Welty, 1996], [Ide, et al., 1997], and [Welty and Ide, 1998]. This paper focuses briefly on the work we have been doing in developing an ontology of modalities for libraries. This ontology is founded on a simple, yet profound, observation: many of the different entities within a library will be versions of the same thing. Some of these may simply be identical copies of the same book, journal, etc., but even in the present day the majority of these cases are different modalities of the same thing.

The prototypical example was formulated in [Welty, 1994]. A paper in the library, which has attributes such as author, abstract, date,

what it was published in, etc., may be available in several formats, such as HTML, Postscript, PDF, etc. The key here is that the postscript file, the HTML file, and the PDF file, are all the same paper. They have the same author, abstract, etc., yet there are attributes of the different formats that are distinct, and are meaningful in the role of those files as different views of the paper. Obviously each has a different location in the filesystem, in addition the HTML file may have an `html-version` attribute, the postscript file may have a “translated by” attribute, etc. Finally, the paper may well exist in paper form as well, and this is merely another view of the paper. The paper form would have attributes like location (such as its library catalog number - perhaps inherited from the book or journal the paper was published in).

This has important implications for the card catalog system. Many views of the same document can be stored in the catalog without requiring different entries. A search for the paper will have a single result, and then give the user the option of “delivery formats”. In fact, part of the AXIOM project includes research into a query language that will allow users to construct new documents from parts of existing ones (this is enabled by the deep markup of the texts).

Of course a digital library will have many different information formats, or modalities, that information may be stored in. 1984, for example, is a book that is available in printed form, on audio tape, as a movie, and in a variety of electronic formats including fully CES (Corpus Encoding Standard) compliant marked up texts in ten languages, accomplished by the MULTEXT project [Véronis, 1996].

The goal of the library is obviously not simply to represent the different modalities, but to deliver them, when possible. This requires that the interface to the library be able to exploit modality information intelligently. In some cases, delivery through the interface will not

be possible, e.g. if the user is interested in the printed book, or the VHS videotape version of the movie, and in these cases the interface should provide information about the location. This and other knowledge about the forms of interaction will need to be specified as part of the modality classes.

The modality ontology has not been the central focus of our KR efforts, and we have at this time we have only a rough ontology. Our main interest in this workshop will be to explore aspects of different interaction modalities to augment the existing ontology. An outline of the current ontology is provided below. It is important to keep in mind that these are modalities for the objects we would expect to find in a digital library, where the principal type of object is a document.

3 Ontology

There are two major (disjoint) modality types: internal-modality and external-modality. The internal and external are with respect to the digital library system: an internal modality is one for which the interface can deliver the object, such as a postscript paper or an MPEG movie. An external modality is one that the interface can not deliver, and thus must present access information to the user.

All modalities are also broken into four basic types: text, graphics, audio, and video. A modality of a particular object must be classified under one of these four types and must also be either internal or external.

Internal text modalities come in two types, formatted text (such as HTML, RTF, Postscript and PDF), and unformatted text (such as plain ASCII). In addition, all the internal text modalities fall under one of the following types: source text (such as HTML, RTF, and ASCII), and image text (such as postscript and PDF). The purpose of the latter disjoint categories is to distinguish text that can effectively be altered, perhaps by the software itself, from

that which can not (or isn't really supposed to be). The source text category is actually fairly deep, with a complete taxonomy of all the mark up languages (including HTML).

The internal graphics formats are GIF, JPEG, TIFF, etc.

The internal audio formats are WAV, real audio, etc.

The internal video formats are AVI, MPEG, etc.

The latter three categories are not well thought out at the time, since we have principally been dealing with text data.

The external text modalities are of two types: bound (such as books, proceedings, journals, etc.) and loose (such as letters, manuscripts, etc.).

The external graphics modalities are photo and microfilm.

The external audio modalities are cassette, CD, record, etc.

The external video modalities are VHS cassette, 8mm film, 16 mm film, etc.

Again, the latter three categories are not well thought out, and in general we have not spent a great deal of time yet with the external modalities because the focus of the project is electronic texts. Our ultimate goal, however, is to fully integrate a digital library with a conventional print library, and the external modalities will be important to complete.

At this point, identifying the modality types and getting a few examples is as far as we have come. While it is clear that the different modalities will have their own attributes, such as the publisher of a bound external text or the URL of an HTML file, we have not analyzed the domain for a complete set of attributes.

References

[Ide, et al., 1996] Ide, N., Priest-Dorman, G., and Véronis, J. *Corpus Encoding Standard*. CES Document CES-1. Available at <http://www.cs.vassar.edu/~ide/CES/CES1.html>

[Ide, et al., 1997] Ide, N., McGraw, T., and Welty, C. Representing TEI Documents in the CLASSIC Knowledge Representation System. *Proceedings of the Tenth Workshop of the Text-Encoding Initiative*. November, 1997.

[Véronis, 1996] Véronis, J. *Multext: Multilingual Text Tools and Corpora*. Available at <http://www.lpl.univ-aix.fr/projects/multext/>

[Welty, 1994] Welty, Chris. Knowledge Representation for Intelligent Information Retrieval. *Proceedings of the CAIA-94 Workshop on Intelligent Access to Digital Libraries*. March, 1994.

[Welty, 1996] Welty, Chris. Intelligent Assistance for Navigating the Web. *Proceedings of the 1996 Florida AI Research Symposium*. May, 1996.

[Welty and Ide, 1998] Welty, C., and Ide, N. Knowledge Representation for Text Markup. *The International Journal of Computers and the Humanities*. To appear.