DLs for DLs:
Description Logics for Digital Libraries

Christopher A. Welty
Vassar College
Computer Science Dept.
Poughkeepsie, NY, U.S.A.
weltyc@cs.vassar.edu

1 Introduction

We are working on several aspects of using description logics for digital libraries. Our main goal is to enable robust retrieval of information stored in a digital library. We have found that the use of description logics has also served to assist in ensuring the integrity of the data. In addtion, the recent release of XML has made us aware of an enormous opportunity to leverage description logic technology, *in reality* against the widely recognized problems with web searching.

The three main thrusts of the project are:

1. Developing an ontology for *card catalog* data, traditionally known as meta-data, including a browsable and queryable hierarchy of subject classifications, similar to, yet substatially deeper and more complex than, the taxonomies which exist today. [1]
2. Developing techniques for representing mark up elements (or tags) specified in an SGML or XML DTD, and instantiating those representations on fully marked-up texts.[2]
3. Developing standardized extensions to XML that can be used by a description logic based web search engine. We can not over-emphasize the potential magnitude of the possibilities this opens up for DL technologies.

These efforts are in conjunction with a large digital libraries group with diverse expertise in areas such as library science, text encoding, linguistic analysis, database query optimization, literature, history, etc. Members of the group represent Vassar, Brown University, CNRS (France), The University of Illinois, AT&T, INSO Corp., and others. The group's central objective is to demonstrate the value of *smart texts*, that is, texts exibiting highly detailed levels of markup. Some members of the group have large amounts of exhaustively marked-up texts according to standards laid down ten years ago by the Text Encoding Initiative (TEI)[6].

The description logic aspects of this work are currently being done in Classic.

2 Card Catalog Ontology

The work on card-catalog based retrieval has been going on for some time. See [5, 4] and [1]. One of the major challenges of the ontological development has been to achieve the goal of enabling *taxonomic* hierarchies of subjects for narrowing searches and for browsing, while accounting for the ambiguities and in general the overloading of the notion of a subject.

For a description logic system to utilize the taxonomy to narrow searches, the subject hierarchy must be terminological. Traditionally, however, a subject can be something conrete, i.e. *A book about Ernest Hemingway.* In this particular case, Hemingway is also an author and thus there are compelling reasons to represent him as an individual.

Subjects can also, traditionally, be purely categorical, i.e. *A book about Artificial Intelligence.* In this case, it makes perfect sense to represent the subject as a concept.

These two cases are at odds with each other, because the subject of the first book is an individual, and the subject of the second is a concept, and this leads to an ambiguous usage of subjects in the ontology.

In addition, current keyword search techniques imply to some that subject is a role. That is, keywords are typically represented as the equivalent of role fillers for the role `subject`. Even if this approach is not used, user interfaces will need to give the illusion that this is true to support users accustomed to the current method.

3 Representing DTDs

The incentive for moving this work into the broader context of the large-scale digital library project is recent collboration with members of the text encoding community. This collaboration is described at a high level in [3], and a more detailed description will appear this year [2].

The encoding community, who are responsible for SGML [7], the TEI [6] and more recently XML [8], has been focusing almost exclusively on syntax-based tools for specifying and extracting markup. A DTD, or Document Type Definition, is in fact little more than a language for specifying context free grammars. Any semantics have, up to now, been represented as documentation, e.g. "The `<PERSNAME>` tag will be used to mark up a person's name," or as code, e.g. code that displays text marked up up with the `<I>` tag in italics.

There has never been any effort to formalize (in a KR sense) the semantics of these tags, and as various groups with large commitments to markup and large amounts of marked up data look to actually put their data online and make it accessible, the deficiencies of a syntax-centered view are becoming apparent. This is not to say that a semantic-centered approach alone would have served better, only that the union of the two can provide for the full functionality that is desired.

This functionality is normally described as the ability to pose advanced queries and to manufacture virtual documents. These facilities are described in detail in [2].

4 The DL Tagset

With the recent release of XML [8], we have developed an interest in exploring the creation of a tagset for specifying information usable by a description-logic based web search engine.

The goal of XML is to provide all the capabilities of HTML and augment this with some of the capabilities of SGML. SGML provides the capability to specify content models (that is, context-free grammars specifying what tags can appear within others and how). In other words, XML will allow documents to provide their own tags. Browser providers will make promises to render certain standard extended tagsets, and search engine providers will make promises to recognize certain extended tagsets. In the latter case, it is expected that most of the tags that specialized search engines will recognize will represent *meta-data*, that will not appear in a browser and thus will not need to be rendered.

Such a vision of the future of the web opens up a huge opportunity for the description logic community to develop a standard set of extended tags for XML that can be utilized by a class of DL-based search engines.

5 The Enrico Question

Finally, the question, "Why are you using a description logic for this," comes up frequently from within and without the project. We attempt here to provide some of our justifications.

Most importantly, description logics are among a class of systems that support well though-out ontologies, and the digitial library community, in particular the text encoding community, have been so long without any rigorous or methodical formal approach to representing their semantics (outside of rudimentary subject classification systems like dewey decimal), that an approach with a strong theoretical and philosophical basis is quite welcome. In fact, librarians are quite familiar with the notion of ontology, in a less formal but still quite practical sense, and are quite excited about the possibilities of being able to formally specify their ideas.

The notion of a taxonomy is an essential feature that is lacking in current text encoding tools, both for data entry and retrieval. Again, description logics are merely one among a class of systems that provide for the specification of taxonomies. In addition, we have found numerous examples where specifying not just the taxonomic links, but their justifications, have been quite effective in ensuring some level of quality control during data entry. For example, one can quite easily define the concept *autobiography* as follows:

$\mathsf{autobiography} \doteq \mathsf{biography} \sqcap \forall \mathsf{subject{:}author}$

During the entry of actual book data, an autobiography which is underspecified as simply a biography can be automatically re-classified if the subject data is accurate. We have found roughly 20 such examples thus far, without actually having done the data entry (into Classic) yet.

A large part of the actual data we will be using are manuscripts, letters, and other documents that are older (in some cases significantly) than 100 years in age. One of the obvious advantages of encoding these documents is that access to them can be provided without endangering the physical integrity of the documents themselves. It is common, however, in such old documents, for information common to a particular type of document to be missing. A letter, for example, in which the recipient's name is not identifiable, but the address is, might be considered, *A letter to someone in Boston*. In a DL we could define the letter as

$\mathsf{letter01} \in \mathsf{letter} \sqcap \forall \mathsf{recipient}{:}(\mathsf{person} \sqcap \mathsf{address} : \mathsf{Boston})$

In a representation system without a terminological component, it is not possible to represent the notion of *someone in Boston* without creating an individual. Creating an individual, of course, in some ways presumes that the recipient of the letter is *not* any of the existing people represented in the system that live in Boston. While this may seem a minor philisophical point to some (hopefully not in the DL community), the librarians in our group are quite motivated by the idea of being this precise, and again we have identified roughly ten similar examples that arise from the need to represent data that is sometimes incomplete.

Acknowledgements The idea of *a letter to someone in Boston* was originally postulated by Alex Borgida. Nancy Ide and Tim McGraw of Vassar College have been part of the work related to the ontology for DTDs.

References

[1] Welty, C. The Ontological Nature of Subject Taxonomies. To appear, *Proceedings of the 1998 International Conference on Formal Ontology in Information Systems (FOIS'98)*. IOS Press. June, 1998. Preliminary version available at http://www.cs.vassar.edu/faculty/welty/papers.

[2] C. Welty and N. Ide. Using the Right Tools: Representing Semantics of Marked-Up Documents. To appear, *Journal of Computers in the Humanities* Special Issue on the Tenth TEI Workshop. July, 1998.

[3] N. Ide and T. McGraw and C. Welty. Representing TEI Documents in the CLASSICKnowledge Representation System. *Proceedings of the Tenth workshop of the Text-Encoding Initiative.* November, 1997. Available at http://www.cs.vassar.edu/faculty/welty/papers.

[4] C. Welty. Intelligent Assistance for Navigating the Web. *Proceedings of The 1996 Florida AI Research Symposium.* May, 1996. Available at http://www.cs.vassar.edu/faculty/welty/papers.

[5] C. Welty. A Knowledge-Based Email Distribution System. *Proceedings of the 1994 Florida AI Research Symposium.* May, 1994.

[6] W. Plotkin and C.M. Sperberg-McQueen. *The Text Encoding Initiative Home Page.* Available at http://www.uic.edu/orgs/tei.

[7] R. Cover. *The SGML Web Page.* Available at http://www.sil.org/sgml/sgml.html.

[8] R. Cover. *The XML Web Page.* Available at http://www.il.org/sgml/xml.html.