*DRTC Workshop on*
*Semantic Web*
*8th – 10th December, 2003*
*DRTC, Bangalore*

**Paper: C**

# Interoperability for Digital Libraries

**Michael Shepherd**
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada B3H 1W5
*shepherd@cs.dal.ca*

## Abstract

*Paper discusses Metadata, Syntactic and Syntax Interoperability of Digital libraries in Semantics Web environment.*

*"The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation." [1]*

## 1. Introduction

Realizing a "Semantic Web" will not be easy. The key phrase, "… *information is given well-defined meaning* …" has been a goal of indexing and classification since knowledge was first classified. Attempting to give well-defined meaning to information on the World Wide Web and in digital libraries is far more difficult to do automatically than to do manually and, assuming that it is given meaning, there still remains the problem of sharing this representation to better enable computers and people to work in cooperation.

The sharing of representations for cooperation requires "interoperability", i.e., the ability of one system to communicate and interact with another system. There are three levels of interoperability:

- Basic – common tools and interfaces that provide uniformity for navigation and access
- Middle – syntactic interoperability that allows the interchange of metadata
- Highest – deep semantic interoperability that allows users to access similar classes of objects and services across multiple sites

The Web is virtually uncontrolled – anyone with access to an ISP can put up a Web page – making it difficult attain a consistent level of interoperability. Digital libraries, on the other hand, tend to have more control over the content and a higher degree of adherence to standards. The problem of interoperability among digital libraries becomes one of sharing not within the libraries themselves, but among the many different repositories of digital content. While there are many definitions of a digital library, the call for papers of the 2002 Joint Conference on Digital Libraries [2] included (but was not limited to) the following notions:

- operational information systems with all manner of digital content;
- new means of selecting, collecting, organizing, and distributing digital content;

- distinguished from information retrieval systems because they include more types of media, provide additional functionality and services, and include other stages of the information life cycle, from creation through use.

The notion that others have included, but was not included in this list, is that digital libraries tend to serve a particular community or communities. This concept of serving a particular community is particularly important when addressing semantic interoperability (the highest level of interoperability) among digital libraries and will be discussed below.

This paper examines syntactic and semantic interoperability among digital libraries, looking at what has been accomplished and where there is work to do.

## 2. Interoperability Based on Metadata

In order for there to be interoperability among systems, there thee must be an agreed upon set of standards; both for describing the content of a digital object and for sharing the representation of the object. The content of a digital object is described by metadata, i.e., information about information. While there are many different metadata standards in use, the Dublin Core [3] is often used because of its simplicity compared to other standards. The Dublin Core has only fifteen different elements, each of which has various attributes, and may be further qualified to provide a narrower semantic meaning.

The platform for sharing metadata is provided by such Web standards as XML and RDF through the work of the W3C. [4]

## 3. Approaches to Interoperability

Within the digital library community, there are three main approaches to interoperability. In increasing order of complexity to implement, these are:

- Federated - a group of organizations decide that their services will be built according to a number of agreed upon specifications
- Harvesting – a looser grouping of digital libraries where participants make some small effort to enable basic shared services
- Gathering - no cooperation of any kind, some interoperability possible by gathering openly accessible information such as is done by Web search engines

Of these three approaches, this paper examines only the Harvesting approach as it represents the middle level of interoperability, or syntactic interoperability.

## 4. Syntactic Interoperability

Syntactic interoperability allows the exchange of metadata. The Open Archives Initiative (OAI) [5] provides standards and protocols for the exchange of metadata through the "harvesting" of metadata from repositories of digital objects, such as digital libraries. The OAI-PMH (Protocol for Metadata Harvesting) is meant to be a low-cost barrier to harvesting, i.e., it should be simple to implement and not use excessive resources.

Most digital libraries have their own internal standards and formats and the objects and metadata are often easily exchanged. The OAI-PMH addresses this issue by having each digital library "expose" the metadata of its resources in a metadata repository. This repository responds to the requests of metadata "harvesters" that gather such metadata into larger repositories where value-added services such as summarization and cross-library searching can be provided to users. The metadata must be returned to the harvester in an XML format as Dublin Core metadata, although other metadata standards are also possible.

While a powerful but light-weight protocol, there are certain issues that must be addressed:

- Granularity
- Repository synchronization
- Deep Web

Metadata is extracted or generated normally at the document level. While this is appropriate for short documents or for documents that are tightly written with respect to the subject content, it is not appropriate for documents that are quite long, not about a single topic or have scholarly arguments that depend on citing specific passages, lines, or words in a source. This granularity issue is addressed in the Perseus Digital Library project [6] by the extraction of metadata for subdocuments, such as chapters or books or sections of larger articles.

Synchronization is an issue of "freshness" of the metadata after it has been harvested. This issue is addressed by search engines (gatherers) by sending out crawlers to revisit

individual pages to see if the page has been updated since the metadata was gathered. In the OAI-PMH model the harvester only needs to visit the metadata repository for an entire digital library and request the identifiers of all resources that have been updated after a specific date.

The deep Web (sometimes called the "hidden Web" or "invisible Web") has been estimated to be 500 times the size of the visible or surface Web [7]. It includes all those Web pages that are not accessible to search engine Web crawlers and thus includes most digital libraries. The DP9 Gateway [8] addresses this issue by putting wrapper around each Web crawler request so that it looks like an OAI-PHM request, and then translates the returned XML-formatted metadata into HTML for the search engine to index.

The WEB-DL project [9] is the inverse of this, it extracts Dublin Core metadata from Web pages to be included in a digital library. There are, however, problems with this as metadata values for most Dublin Core fields cannot be found in Web pages.

## 5. Semantic Interoperability

The Grand Challenge in the 1990s was posed as "semantic interoperability across digital collections". [10] While progress has been made on the sharing of metadata (syntactic interoperability), the problem of semantic interoperability is much greater. The issues are language issues, i.e., the meaning of terms and how they are used.

The meaning of any set of terms, and the significance and utility of any taxonomy, according to Wenger [11], can be evaluated only in the context of a community whose members are involved in similar activities and share similar values. In short, semantic interoperability is tied directly to communities of practice, and to the negotiation of meaning that occurs within them. [12]

If we assume that digital libraries and built to serve a particular community or communities, then, "… the degree of interoperability between information systems may be dependent on the distance between communities whose information systems attempt to interact." [12] Thus, according to Moen [13] and Friesen [12], focused communities of practice. When such communities are relatively "close", they may form an extended community, and when they are more distant, then they become extra communities

relative to each other (Figure 1). The question for semantic interoperability is how to communicate across these communities of practice.

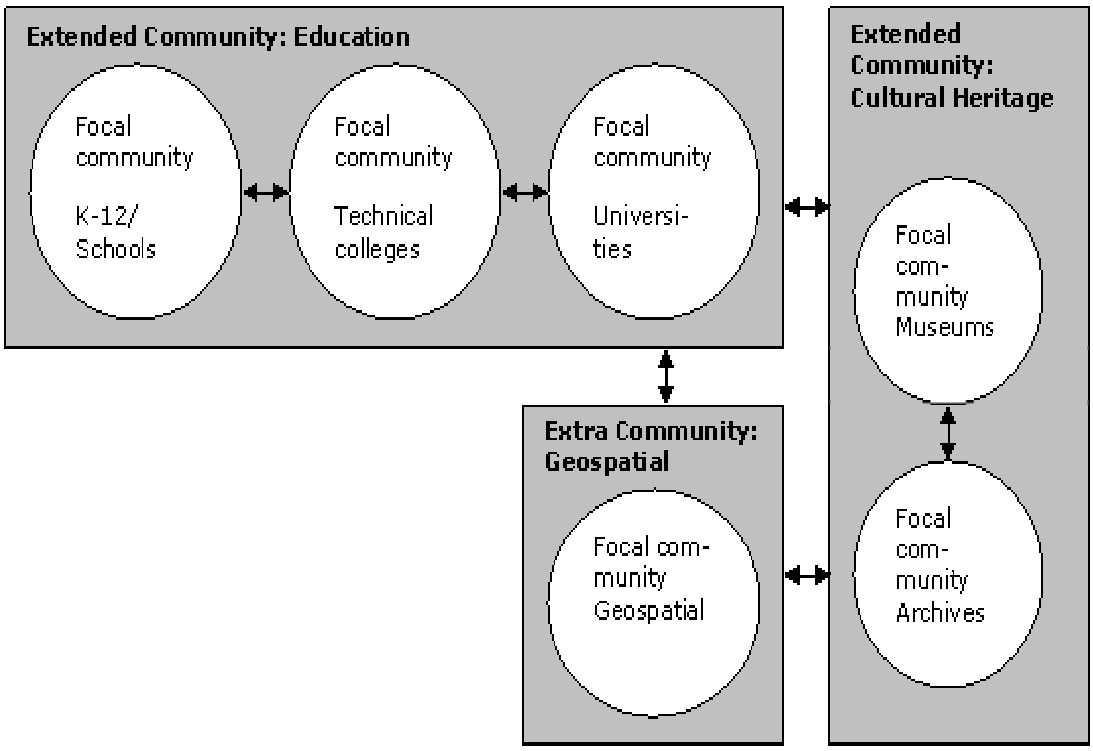


**Fig. 1:** Extended and Extra Communities of Practice

The Z39.50 Profiles standard is an attempt to address this question. Profiles indicate the behaviour of the Z39.50 client and server and the functionality needed in the underlying Information Retrieval system. They can be used to prescribe how Z39.50 should be used in a particular application environment or solve interoperability problems within a community or across two or more communities.

The Bath Profile [13] is an example of such a Z39.50 Profile. The metadata of the target systems are mapped to Dublin Core records using XML and the Dublin Core Elements are used as attributes in the Z39.50 query. Figure 2 shows such a model accessing various digital libraries.

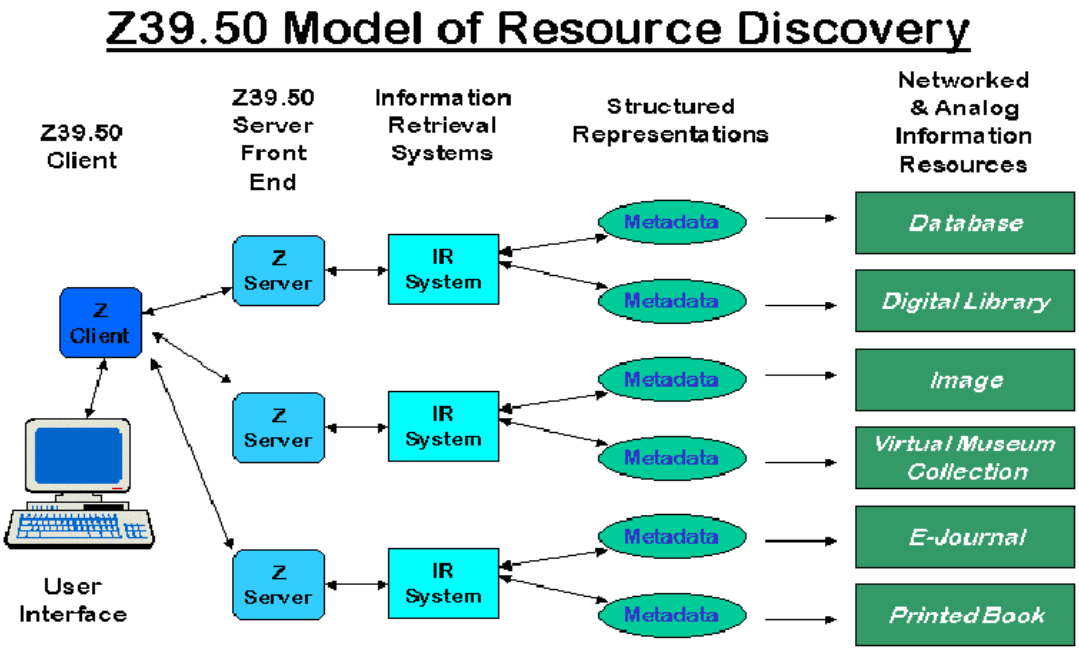


**Fig. 2:** Model of Resource Discovery

## 6. Summary

According to Schatz, while the Grand Challenge of the 1990s was semantic interoperability across digital collections, the Grand Challenge of the 2000s will be "conceptual navigation across community repositories". [14] The question will be how to navigate effectively across different collections of different objects represented by different communities at different levels.

In order to respond to this challenge, we must be able to identify boundary objects. "Boundary object" is a concept to refer to objects that serve an interface between different communities of practice [15]. Boundary objects are entities shared by several different communities but viewed or used differently by each of them. They necessarily contain sufficient detail to be understandable by both parties, however, neither party is required to understand the full context of use by the other party.

An example of a boundary object is an electronic health record. It is used by doctors, nurses, hospital administrators, insurers, government, etc. Each community of users will

have a slightly different view of and understanding of the health record, but the health record serves an interface among these communities.

As shown in Figure 3, research and development on semantic interoperability is proceeding on two fronts; the descriptive or representational front (above the arrow) and the technical front (below the arrow).

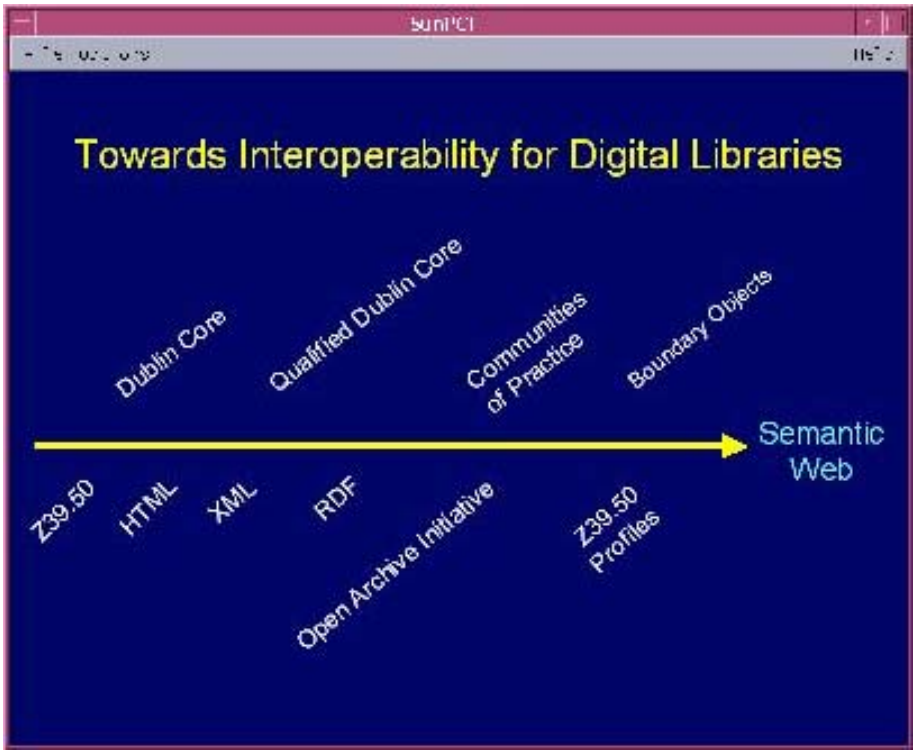


**Fig. 3:** Towards the semantic web.

In summary, we have made large advances, with respect to creating a Semantic Web, on the syntactic interoperability front, but still have some distance to go on the semantic interoperability front. In particular, more work is needed on boundary objects – how to identify them and how to use them.

## 7. References


1. Berners-Lee, T., Hendler, J. and Lassila, O.. "The Semantic Web", *Scientific American*. May 17, 2001. Available April 18, 2003 *http://www.scientificamerican.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21*

2. Joint Conference on Digital Libraries 2002. Portland, Oregon, USA July 14-18, 2002, *http://www.ohsu.edu/jcdl/*
3. Dublin Core Metadata Initiative. *http://dublincore.org/*
4. World Wide Web Consortium. *http://www.w3.org/*
5. Open Archives Initiative. *http://www.openarchives.org/*
6. Smith, David A.; Mahoney, Anne and Crane, Gregory. Integrating Harvesting into Digital Library Content. http://www.perseus.tufts.edu/Articles/oaishort.pdf
7. M.K. Bergman. The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, 7(1), 2001.
8. Xiaoming Liu, et al., DP9: An OAI Gateway Service for Web Crawlers. *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, 14-18 July 2002, Portland, Oregon, U.S.A., pp. 283-284.
9. P. Calado, et al., The Web-DL Environment for Building Digital Libraries from the Web. Proceeding of the 2003 Joint Conference on Digital Libraries. 27-31 May 2003, Houston, Texas. pp. 346-360.
10. Clifford Lynch & Hector Garcia-Molina. Interoperability, Scaling, and the Digital Libraries Research Agenda: A Report on the May 18-19, 1995, IITA Digital Libraries Workshop, August 22, 1995
11. Etienne Wenger. Communities of Practice: Learning, Meaning and Identity. Cambridge University Press, 1999.
12. Norm Friesen. Semantic Interoperability and Communities of Practice, February 5, 2002, http://www.cancore.ca/documents/semantic.html
13. Moen, W.E. (2001). Mapping the interoperability landscape for networked information retrieval. In *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, VA, June 24-28, 2001*. pp. 50-52. [Web Page]. URL http://www.unt.edu/wmoen/publications/MapInteropJCDLFinal.pdf
14. Bruce R. Schatz. Navigating the Distributed World of Community Knowledge**.** http://www.sis.pitt.edu/~dlwkshop/paper_schatz.pdf
15. Denham Grey. http://denham.typepad.com/km/2003/10/boundary_object.html