2000

Representation and Organization of Information in the Web Space: From MARC to XML

Jian Qin
Syracuse University

Representation and Organization of Information in the Web Space: From MARC to XML

Jian Qin
School of Information Studies
Syracuse University

jqin@syr.edu

Abstract

Representing and organizing information in libraries has a long tradition of using rules and standards. As the very first standard encoding format for bibliographic data in libraries, MAchine Readable Cataloging (MARC) format is being joined by a large number of new formats since the late 1980s. The new formats, mostly SGML/HTML based, are actively taking a role in representing and organizing networked information resources. This article briefly describes the historical connection between MARC and the newer formats for representing information and the current development in XML applications that will benefit information/knowledge management in the new environment.

Keywords: MARC, information representation, information organization, XML schemas

Introduction

The notion of information representation and organization traditionally means creating catalogs and indexes for publications of any kind. It includes the description of the attributes of a document and the representation of its intellectual content. Libraries in the world have a long history in recording data about documents and publications; such practice can be dated back to several thousand years ago. Indexes and library catalogs are created to help users find and locate a document conveniently. Records in the information searching tools not only serve as an inventory of human knowledge and culture but also provide orderly access to the collections. Just like every other business and industry, the representation and organization of information in the network era has gone through dramatic changes in almost every stage of this process. The changes include not only the methods and technology used to create records for publications, but also the standards that are central to the success and effectiveness of these tools in searching and retrieving information. Today the library catalog is no longer a tool for its own collection for the library visitors; it has become a network node that users can visit from anywhere in the world via a computer connected to the Internet. The concept of indexing databases is no longer just for newspapers and journal articles; it has expanded into the Web information space that is being used for e-publishing, e-businesses, and e-commerce.

The heart of such a universal information space lies in the standards that make it possible for different types of data to be communicated and understood by heterogeneous platforms and systems. We all know that TCP/IP allows different computer systems to talk to each other and to understand different dialects of networking language; in the world of organizing information content, the content is represented by terms either in natural or controlled language or both. The characteristics of its container (book, journal, film, memo, report, etc.) will be encoded in certain format for computer storage and retrieval. Libraries in the world have used MAchine Readable Cataloging (MARC) (Library of Congress, 1999) to encode information about their collections. In conjunction with cataloging rules, such MARC format standardized the record structure that describes information containers, i.e., books, manuscripts, maps, periodicals, motion pictures, music scores, audio/video recordings, 2-D and 3-D artifacts, and microforms. The Online Computer Library Center (OCLC) in Dublin, Ohio is the largest and the busiest cataloging service in the world. Almost 33,000 libraries from 67 countries now use OCLC products and services and more than 8,650 of them are OCLC members. As e-publishing thrives and Web information space grows, libraries have expanded conventional cataloging of their collections into organizing the information on the Web. In the early 1990s, OCLC started the Internet cataloging project, in which librarians from all types of libraries volunteered to contribute MARC records they created for Gopher servers, listserves, ftp and Web sites, and other net-

worked information resources (OCLC, 1996). Another major undertaking in organizing information on the Web is OCLC's Metadata Initiative (Dublin Core Metadata Initiative, 1999) inaugurated in 1995, which proposed a metadata scheme containing 15 data elements. Among them are title, creator, publisher, subject, description, format, type, source, relation, identifier, and rights. The metadata scheme was named after the city where OCLC is located: Dublin Core Metadata Element Set (Dublin Core for short). Since its debut, it has become an important part of the emerging infrastructure of the Internet. Many communities are eager to adopt a common core of semantics for resource description, and the Dublin Core has attracted broad ranging international and interdisciplinary support for this purpose.

Metadata and Metadata Creation

The term "metadata" refers to "machine-understandable information about Web objects" (Swick, 1997). It is the "documentation about documents and objects. They describe resources, indicate where the resources are located, and outline what is required in order to use them successfully" (Younger, 1997). Metadata schemes, such as Dublin Core, entail a group of codes or labels that describe the content and/or container of digital objects. When the metadata is embedded in hypertext documents, they can accommodate automatic indexing for digital objects and thus provide better aids in networked resource discovery. Several terms have been used interchangeably in describing the digital objects that a user views through various interfaces (e.g., a Web browser). They are given names such as Web document, Web object, digital object, hypertext, and hypermedia.

Post-Publishing Representation

Post-publishing representation is a method in which a special type of computer program generates metadata from digital objects already published. These programs are known as spiders, knowbots or automatic robots, Webcrawlers, wanderers, etc. Using these programs, metadata are extracted from the objects that were made available on the Internet. Many of the Web search engines, e.g., Excite, Lycos, AltaVista, employ the post-publishing representation method to collect metadata and build their metadata bases for networked information discovery purposes. This fully automated process of metadata generation is "a mixed blessing": it requires little or no human intervention, but the methods used to extract metadata are too simple and far from effective in resource discovery. Lynch indicates that automatic indexing is "less than ideal for retrieving an ever-growing body of information on the Web" for several reasons: the inability to identify characteristics of a document such as its overall theme or its genre, lack of standards, and inadequate representation for images (Lynch, 1997). However, post-publishing representation has its merits. The most appealing advantage is probably that updating a metadata base can be done automatically and as frequently as one desires. This advantage makes it possible for popular search engines such as Yahoo! AltaVista, and HotBot to create dynamic metadata in response to queries. Since they do not generally retrieve the metadata content, results are created on the fly to answer users' queries (Schwartz, 1998). Another advantage comes with this automatic indexing process: the labor costs tend to be low because little or no human intervention is involved in the metadata harvesting process.

Pre-Publishing Structuring

One way to compensate for the shortcomings in post-publishing representation is through pre-publishing structuring, i.e., attaching structured metadata to the digital objects so that automated indexing programs can collect this information in a more efficient way. Earlier efforts in pre-publishing structuring of metadata have taken place in various domains. The Text Encoding Initiative (TEI) (University of Virginia, 1994) was one of the pioneers. It is basically an encoding scheme consisting of a number of modules or Document Type Declaration (DTD) fragments, which include 3 categories of tag sets: (1) core DTD fragments; (2) base DTD fragments; and (3) additional DTD fragments. Another project, the Encoded Archival Description (EAD) (Library of Congress, 1996) is an SGML document type definition for encoding finding aids for archival collections. Other domain-specific projects include the Content Standards for Digital Geospatial Metadata (CSDGM) (Federal Geographic Data Committee, 1998) and the Government Information Locator Service (GILS) (OIW/SIG-LA, 1997). As of April 1998, there were over 40 projects in more than 10 countries that either use Dublin Core or are developing their own metadata element set that are based on Dublin Core.

The common element among these projects is that they embed the structured metadata into the Web objects prior to or after their "publication." The structured metadata consists of components that allow establishing relationships among data elements with other entities, and these components are usually categorized into several different "packages" or "layers." Newton (1996) maintains that "[meta]data elements must be described in a standard way as well as classified. Attribute standardization involves the specification of a standard set of attributes, and their allowable value ranges, independently of the application areas of data elements, tools, and implementation in a repository." Her five categories of attributes include identifying, definitional, relational, representational, and administrative, reflecting a complex structure in metadata elements. Bearman and Sochats (1996) propose a reference model for business-acceptable communication. They define clusters of data elements that would be required to fulfill a range of functions of a record. The functions of records are identified as:

- The provision of access and use rights management
- Networked information discovery and retrieval
- Registration of intellectual property
- Authenticity, including: handle, terms and conditions, structural, contextual content, and use history

Metadata and Digital Information Repositories

Among the key concepts in digital information repositories, metadata plays two important roles: as a handler (i.e., identifier) and as points of access to data/document content (Kahn & Wilensky, 1995). As a locator, metadata helps users obtain the data or document by providing the exact location. As access points, metadata supplies information about the content of resources. The demand for effective organization of information does not diminish with powerful information technology, but rather, people nowadays have higher expectations for networked resources. The success of a digital information repository in meeting such high expectations depends largely on the quality and scale of metadata, which, in turn, depends on a whole set of information processing standards and quality control management.

Metadata and XML

The dilemma of post-publishing representation and pre-publishing structuring reflects the inadequacy of describing unstructured data/documents coded with HTML. Given the shorter publishing cycle and huge volume of information, any method requiring heavy manual intervention in creating metadata records would be impractical. If data or documents can be structured with meaningful tags at the time they are created, it would greatly increase the flexibility of these data/documents to be exchanged and understood over the network systems. The structured documents can make it easier to extract information about them to build metadata repositories. This is where the eXtensible Markup Language (XML) (Cover, 2000) comes in to play.

XML describes a class of data objects called XML documents and partially describes the behavior of computer programs that process them. It is an application profile or restricted form of SGML, the Standard Generalized Markup Language. XML allows large-scale Web content providers to perform such tasks as industry-specific markup, vendor-neutral data exchange, media-independent publishing, one-on-one marketing, workflow management in collaborative authoring environments, and the processing of Web documents by intelligent clients. XML applications for creating metadata involve a wide range of activities: sitemaps, content ratings, stream channel, definitions, search engine data collection (web crawling), digital library collections, and distributed authoring. There are several parallel efforts in developing XML-based metadata applications. One of them is the Resource Description Framework (RDF) developed at W3C (Lassila & Swick, 1999). RDF "is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. RDF emphasizes facilities to enable automated processing of Web resources." RDF uses XML as syntax to express the semantics in the RDF data model. A simple example is diagramed in Figure 1 to demonstrate how RDF/XML structures data elements. This diagram represents that "the individual referred to by employee id 85740 is named Ora Lassila and has the email address lassila@w3.org. The resource http://www.w3.org/Home/Lassila was created by this individual." In RDF/XML, it will be represented as:

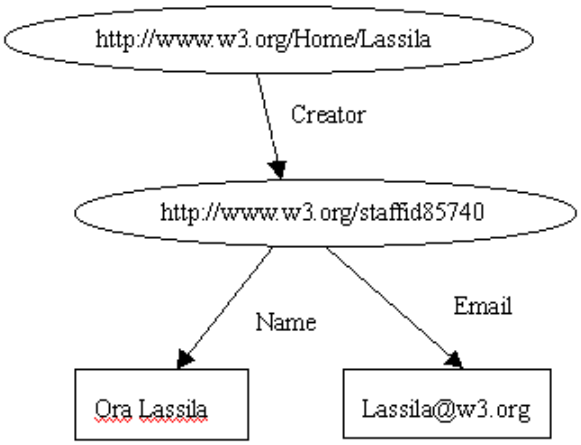

Figure 1. Structured value with identifier (Source: http://www.w3.org/TR/1999/PR-rdf-syntax-19990105/)

```
<rdf:RDF>
  <rdf:Description about= "http://www.w3.org/
        Home/Lassila">
    <s:Creator rdf:resource=http://www.w3.org/staffId/85740/>
  </rdf:Description>

  <rdf:Description about="http://www.w3.org/staffId/85740">
     <v:Name>Ora Lassila</v:Name>
     <v:Email>lassila@w3.org</v:Email>
  </rdf:Description>
</rdf:RDF>
```

This example bears four important elements: description, property element, property attribute, and data type, which comprise the RDF data model. Theoretically, if all documents and data adopt this type of structure when they are created, then it will greatly increase the quality of metadata and reduce the cost in generating metadata databases due to the exchangeability and interoperability of metadata. The potential

in XML syntax-based metadata opens up opportunities for a wide range of applications not only in e-publishing and digital libraries, but also in e-businesses and e-commerce.

XML Namespaces

One of the requirements for organizations these days is to have effective information systems that can quickly respond to information needs of ad hoc nature or for decision-making. XML can contribute to build such a system by quickly generating both data-centric and document-centric documents. The so-called "data-centric" documents are characterized by "fairly regular structure, fine-grained data (that is, the smallest independent unit of data is at the level of a PCDATA-only element or an attribute), and little or no mixed content… The document-centric documents often have irregular structure; larger grained data (that is, the smallest independent unit of data might be at the level of an element with mixed content or the entire document itself" (Bourret, 1999). It becomes a reality now that almost all the information flowing within and between organizations can be represented as one of these two kinds of documents (marked up by XML), stored in databases, and communicated through network systems.

A recent statistical survey found that up to October 1999, a total of 179 initiatives and applications emerged (Qin, 1999). Many of these applications propose specialized data elements and attributes that range from business processes to scientific disciplinary domains (Figure 2). Businesses and industry associations are the most active developers in XML initiatives and applications (Figure 3). The burgeoning of these specialized XML applications raises a critical issue: how can we be sure that data/documents marked up by these specialized tags can be understood correctly cross different systems in different applications? It is well known that different domains use their own naming conventions for data elements in their operations. For example, the same data element "Customer ID" may be named as "Client ID" or "Patron ID." Besides the same data may be named differently, the same term may also mean different things, such as "title" may be referring to a book, a journal article, or a person's job position. To further complicate the issue, future XML documents will most likely contain multiple markup vocabularies, which pose problems for recognition and collision.

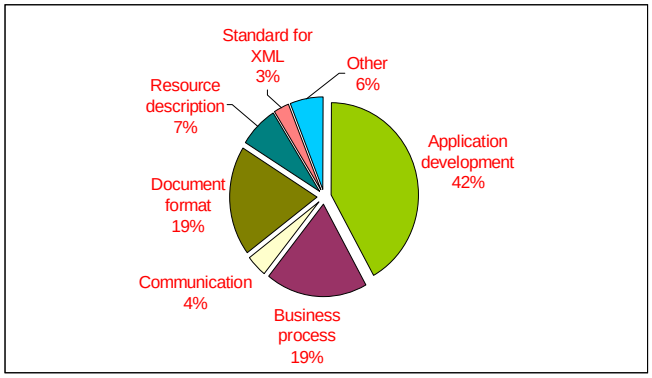

Figure 2. Areas of XML application

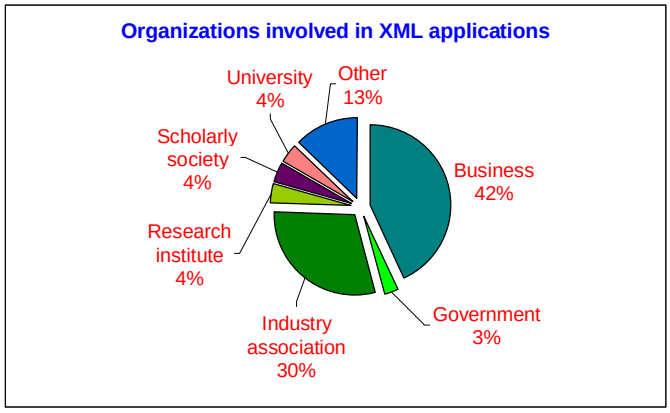

Figure 3. Organization categories involved in XML applications

Solutions to the problems related to XML namespaces lie largely in the hands of the library and information science community who, over the years of research on information/knowledge representation and organization, have developed a whole spectrum of methodologies and systems. An immediate example is that the techniques used in thesaurus construction and control can be applied to standardize the naming of data elements in various XML applications and map out semantics of data element names in namespace repositories. With more and more XML applications sprouting, the demand for namespace control and management will also increase.

Conclusion

When libraries began to use MARC format for their library catalogs back in the late 1960's, they mainly converted their printed records into electronic form for storage and retrieval. The materials represented by these records are physical and static. In the Web space, there is not much physical, nor static--the material is virtual and the information is dynamic. The library's role today has more emphasis in being as a "path-finer" than a "gatekeeper." All these grant the library and information profession a wonderful opportunity to take a significant part in this information revolution, as well as a great challenge to demonstrate the value of library and information science and its potential contribution to e-organizations and e-enterprises.

References

Bearman, D. and K. Sochats. (1996). *Metadata Requirements for Evidence*. Accessed January 30, 2000: http://www.lis.pitt.edu/~nhprc/BACartic.html.

Bourret, R. (1999). *XML and Database*. Accessed January 30, 2000: http://www.informatik.tu-darmstadt.de/DVS1/staff/bourret/xml/XMLAndDatabases.htm

Cover, R. (1999). *The SGML/XML Web Page: Extensible Markup Language (XML).* Accessed January 30, 2000: http://www.oasis-open.org/cover/xml.html.

Federal Geographic Data Committee. (1998). *Content Standard for Digiotal Geospatial Metadata (CSDGM).* Accessed January 30, 2000: http://www.fgdc.gov/metadata/contstan.html .

Kahn, R. and R. Wilensky. (1995). *A Framework for Distributed Digital Object Services*. Accessed January 30, 2000: http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html.

Lassila, O. & Swick, R. R. (1999). *Resource Description Framework (RDF) Model and Syntax Specification.* Accessed January 30, 2000: http://www.w3.org/TR/1999/PR-rdf-syntax-19990105/.

Library of Congress. (1999).[1]*MARC Standards*. Accessed January 30, 2000 http://lcweb.loc.gov/marc/

Library of Congress. Network Development and MARC Standards Office. (1996). *Encoded Archival Description (EAD) DTD*. Accessed January 30, 2000: http://lcweb.loc.gov/ead/.

Lynch, C. (1997). Searching the Internet: Organizing Material on the Internet. *Scientific American*, 276, March, 52-56.

Namespaces in XML. W3C Working Draft 16-September-1998. Accessed February 10, 2000: http://www.w3.org/TR/1998/WD-xml-names-19980916.

Newton, J. (1996). Application of Metadata Standards. In: *Proceedings of the First IEEE Metadata Conference, April 16-18, 1996, Silver Spring, Maryland*. Accessed January 30, 2000: http://www.computer.org/conferen/meta96/newton/paper.html.

OCLC. (1996). *Internet Cataloging Project Call for Participation: Building a Catalog of Internet-Accessible Materials*. Accessed January 30, 2000 http://www.oclc.org/oclc/man/catproj/catcall.htm

OCLC. (1999). *Dublin Core Metadata Initiative.* Accessed January 30, 2000 http://www.oclc.org/oclc/research/projects/core/index.htm

OIW/SIG-LA. (1997). *Application Profile for the Government Information Locator Service (GILS).* Version 2. Accessed January 30, 2000: http://www.usgs.gov/gils/prof_v2.html.

Qin, J. (1999). *Discipline- and industry-wide metadata schemas: Semantics and Namespace Control*. Paper presented at the ASIS Annual Meeting 1999.

Schwartz, C. (1998). Web Search Engines. *Journal of the American Society for Information Science* 49, 973-982.

Swick, R. (1997). *Metadata: A W3C Activity*. Accessed January 30, 2000: http://www.w3.org/Metadata/Activity.html .

University of Virginia. Electronic Text Center. (1994). *TEI Guidelines for Electronic Text Encoding and Interchange* (P3). Accessed February 10, 2000: http://etext.virginia.edu/ TEI.html.

Younger, J. A. (1997). Resources Description in the Digital Age. *Library Trends*, 45, 462-481.