

Metadata: a New Word for an Old Concept

Amin Yousefi

Department of Library and Information Science
Faculty of Psychology and Education
Allameh Tabatabaee University
Tehran, Iran

Shima Yousefi

Department of Medical Library and Information Sciences
Faculty of Management and Medical Information Science
Iran University of Medical Sciences
Tehran, Iran

Introduction

Metadata, or "data about data," is a new word based on an old concept. In libraries, cataloging is the process of creating metadata. A card-catalog containing information about a book is a simple example of metadata describing characteristics of an information resource. Regardless of old concepts, the term "metadata" is used particularly in the context of modern information systems and electronic networks.

Defining Metadata

Metadata has been defined in various ways. Tim Berners Lee defined metadata as "machine-readable information about electronic resources or other things" (1997). This definition addresses metadata applied to electronic resources and refers to "data" in a broader scope that includes not only textual, but non-textual information such as graphics, music, or anything likely to appear in an electronic format. It

is clear that metadata can be deployed for non-digital objects too. But as mentioned, it most commonly refers to digital information especially on the Web.

Another definition of metadata is that assigned by the DESIRE project: "Data associated with objects which relieves their potential users of having to have full advance knowledge of their existence and characteristics" (2000). The basic purposes of metadata are covered by this definition, including a wide range of operations such as discovery, description, management, and long-term preservation of information resources. Metadata also facilitates and improves the information retrieval process (when examined with a view towards recall and precision criteria), by identifying the major concepts of the information resource.

Main Types of Metadata

The abovementioned definitions address three main types of metadata. According to the North Carolina ECHO (Exploring Cultural Heritage Online) Guidelines for Digitization (2006), these are:

1. **Descriptive metadata** describes a resource for purposes such as indexing, discovery and identification. It can include elements such as title, abstract, author, and keywords.
2. **Structural metadata** includes information employed to display and navigate digital resources; also includes information on internal organization of the digital resource. Structural metadata might contain information such as the structural divisions of a resource that indicates how compound objects are put together—for example, how pages are ordered to form chapters, or information about sub-object relationships such as individual diary entries in a diary section.
3. **Administrative metadata** provides information to help manage a resource, such as the data and the state in which the resource was created, file type and also right management information (which deals with intellectual property rights). Administrative metadata might include technical information, such as the resolution at which the images were scanned, the hardware and software used to produce the image, compression information, pixel dimensions, etc. Administrative metadata may also assist in the long-term preservation of digital resources (which contains information needed to archive and preserve a resource). It is mentionable that sometimes Rights management and

Preservation information are listed as separate metadata types (NISO, 2004).

Other categorizations of metadata exist. One of them is as follow: Administrative, Descriptive, Preservation, Technical, and Use metadata (Gill, Gilliland, & Woodley, 2000).

The essential information that metadata gives about a resource is: how it was gathered, the purpose of its gathering, manifestation and manipulation, intellectual properties, and content descriptions such as title, subject, and abstract. This information is represented by a limited number of elements. Each element can take one or more values. These elements are originally defined by one of the metadata schemas. The elements must be embedded in an encoding structure—such as HTML or XML—in one of two ways: in the object itself or separately.

Dublin Core

There are several metadata schemes that were designed to meet the unique needs of specific users, and the number is growing rapidly, but the most popular schema, *Dublin Core*, has been accepted as a sort of standard.

In March 1995, a group of librarians, archivists, information professionals, and other parties interested in describing Internet resources, attended a workshop of the National Center for Supercomputing Applications (NCSA) at the Online Computer Library Center (OCLC) in Dublin, Ohio. Their original objective was to create a core set of elements that could be used for categorizing Web-based resources. The outcome of this workshop was 13 core elements, later increased to 15: title, subject, description, source, language, relation, coverage, creator, publisher, contributor, rights, date, type, format, and identifier.

These elements are continually extended for simplicity, and the level of details is increasing to meet the needs of specialized groups. All elements are optional and repeatable. The continuing development of the Dublin Core is managed by the Dublin Core Metadata Initiative (DCMI).

Although the Dublin Core elements are limited and simple, they can be mapped in more complex systems such as MARC. Also the elements can be added for site-specific purpose or specialized fields.

Thus the major advantages of the Dublin Core are its usability and flexibility. In addition to the 15 elements, Dublin Core also has 3 qualifiers that give additional information for interpretation of elements and enable it to function in an international context:

1. Language: specifies the language of the element value (and not the resources itself). Example: Title LANG=en.
2. Scheme: specifies a context for the explanation of a given element. This qualifier indicates the set of regulations, standards, conventions or norms from which a term in the content of the element has been taken. Typically this will be a reference to an accepted standard. For example: Subject SCHEME=LCSH. (this indicates that the Library of Congress Subject Heading is used to identify the subject keywords)
3. Sub-Element: Refines the meaning of element. It specifies a facet of a given field. For example a sub-element for "title" can be "journal.title = Library Philosophy and Practice."

With these three qualifiers, Dublin Core also meets higher level scientific and subject-specific resource discovery needs. In the last few years, there has been a motion within the Dublin Core community toward use of the Dublin Core Metadata Element Set for more complex and specialized resource description tasks, and toward developing mechanisms for incorporating such complexity within the basic element set. Made possible by using above qualifiers, this has generally been called *qualification of Dublin Core*. Dublin Core, in the hands of information professionals, is expected to provide an alternative to more developed description models such as AACR2/MARC cataloging.

Some other Metadata Element Sets

Dublin Core, though popular, is not the only metadata scheme being used. A few of the most common ones include:

1. **Global Information Locator Service (GILS)**. Formerly known as Government Information Locator Service, GILS was created by the US Federal Government to provide a means for locating information generated by government agencies. Although its original goal was to provide high-level locator records for US government resources, it has in various forms been adopted by other governments and for international projects, leading to its current designation, *Global Information Locator Service* (NISO, 2004). Part of GILS is a complex metadata scheme influenced by

- MARC and designed for Z39.50 servers and clients. GILS has a Core Element Set much larger than that of Dublin Core. It contains separate fields for details on the point of contact and the provenance of the information, administrative fields, and fields for copyright and other access constraints (Milstead and Feldman, 1999) , but generally its emphasis is on availability and distribution rather than on description.
2. **Text Encoding Initiative (TEI)**. TEI attempts to define the encoding of texts. The Text Encoding Initiative Guidelines were published in 1994, the result of a project funded jointly by the US National Endowment for the Humanities and the European Union 3rd Framework Program for Linguistic Research and Engineering. TEI is now a joint project sponsored by three professional bodies: the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing. Burnard (1994) describes the goal of the project as follows: "To define a set of generic guidelines for the representation of textual materials in electronic form, in such a way as to enable researchers in any discipline to interchange and re-use resources, independently of software, hardware, and application area." The TEI initiative aimed to reach agreement on encoding text across a range of disciplines. Giordano (1994) says, "It represents a major milestone—before the TEI it had not been possible to reach consensus among research communities about encoding conventions to support the interchange of electronic texts." The TEI Guidelines, despite their origins in the humanities and linguistics were designed to form an extensible framework that could be used to describe all kinds of texts. It is mentionable that the word "text" should not be read too literally—the TEI is equally concerned with both textual and non-textual resources in an electronic form, whether as constituents of a research database or components of non-paper publications (Burnard, 1994).
 3. **Encoded Archival Description (EAD)**. The EAD standard was developed to allow finding aids to be searched and displayed online. According to Caplan (2002): " Unlike the TEI...the EAD was designed as an electronic finding aid to resources that would not necessarily be available in electronic form. While the EAD can be used to describe web-accessible collections, its primary purpose is to improve awareness of archival holdings in all formats" (p. 3). The EAD standard is maintained jointly by the Library of Congress and the Society of American Archivists (see <http://www.loc.gov/ead/>). Hodge (2001) notes that "although it

is easier to put finding aids on the Web by simply marking them up in HTML...libraries and archives investing in EAD creation hope that using this metadata scheme will encourage consistency in encoding and give them some measure of search interoperability" (p. 7).

Conclusion

The number of metadata projects is growing rapidly. Probably the biggest obstacle in the way of development of metadata is the variety of different metadata projects. Any group may create its own metadata standards to meet its own specific needs, and creators are free to use whatever elements come to mind. Even if common metadata elements are used, the content of the elements will not be compatible. It seems essential to use a global controlled vocabulary system for all metadata element sets. But this raises another question: would the use of controlled vocabularies make searching less efficient? There is some voluntary coordination between projects at the very top level and developers of these projects have been active in developing "crosswalks" between their systems (Milstead and Feldman, 1999). It is this coordination that may be the key to ensuring future compatibility.

References

Baca, Murtha (ed.), Tony Gill, Anne J. Gilliland, and Mary S. Woodley. (2000). Introduction to Metadata: Pathways to Digital Information. Online Edition, Version 2.1. From: http://www.getty.edu/research/conducting_research/standards/intrometadata/

Berners-Lee, Tim (1997). "Metadata Architecture". From: <http://www.w3.org/DesignIssues/Metadata.html>

Burnard, Lou (1994). "The Text Encoding Initiative Guidelines". From: <http://ftp.sunet.se/pub/etext/ota/TEI/doc/teij31.sgm>

Caplan, Priscilla (2002). "International Metadata Initiatives: Lessons in Bibliographic Control". From: http://www.loc.gov/catdir/bibcontrol/caplan_paper.html

Desire project (2000). "A Review of Metadata: a Survey of Current Resource Description Formats". From: http://www.desire.org/results/discovery/cat/meta_des.htm

Giordano, Richard. (1994). The documentation of electronic texts using Text Encoding Initiative headers: an introduction. *LRTS 38(4)*

Hodge, G. (2002). Metadata made simpler. From <http://www.niso.org/standards/resources/UnderstandingMetadata.pdf>

Milsted, J. & Feldman, S. (1999). Metadata: Cataloging by Any Other Name ... *ONLINE*. (January)

NISO (National Information Standards Organization), (2004). "[Understanding Metadata](http://www.niso.org/standards/resources/UnderstandingMetadata)". From: <http://www.niso.org/standards/resources/UnderstandingMetadata>

North Carolina ECHO (2006). Guidelines for Digitization. From: <http://www.ncecho.org/guide/metadata.asp>

Acknowledgment

Very special thanks to [Dariush Alimohamadi](#) for his contributions. The authors gratefully thank him for his revisions on this paper.