PLA Tech Note: Metadata

Metadata: Always More Than You Think

Metadata is often defined offhandedly as "data about data" or "information about information" which doesn't actually tell you anything. The prefix "meta" comes from the Greek and can indicate change, as in *metamorphosis*; or it can mean beyond or after, as in *metaphysics*. In information technology usage, the word metadata has come to be used as a definition or description of data: a small indicator that encompasses and points to a larger piece of information. The library card catalog is the standard metaphor for metadata: each card represented and led the user to a much larger body of information, the book or other item cataloged.

So, "metadata is a succinct and systematic set of information that references, and can be used to efficiently and accurately retrieve, a larger set of information" (Robert DeCandido, in his chapter on metadata, *Internet Searcher's Handbook, 2nd Ed*). Metadata includes indexing and cataloging but goes far beyond those as it is applied to electronic resources: metadata can be used to describe geospatial information, music, art images. There is as yet no single standard for metadata, although various groups, both library and non-library related, are working furiously to develop one.

What does metadata do?

Metadata enables searchers to find information in cyberspace and to see how it relates to other information. The traditional methods of cataloging do not work for the Web: the proliferation of data makes that impossible. So the use of metadata, what Milstead and Feldman call "cataloging by any other name," provides a way to organize Web information in a way that makes it retrievable.

Librarians have a professional understanding of the need for standardization which may not be shared by others working on metadata. Milstead and Feldman write: "Information professionals ... are concerned both with how to write down the descriptive information and what to write down. In contrast, most current non-library approaches to this problem address the structure of data rather than its contents, and this represents a severe shortcoming. In other words, they are concerned with what fields are established, but not particularly worried about which terms to put in them."

What does metadata look like?

The metadata set we are going to look at comes from the Dublin Core, which was created for text-based Web documents by an OCLC sponsored group. The Dublin Core offers papers, workshops, discussions and working groups "to facilitate the discovery of electronic resources." The fifteen element Dublin Core set is intended to be usable across a broad spectrum, with no more complexity

than that of a library catalog card. The Dublin Core Elements are Title, Creator, Subject, Descriptions, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights, as described by Sutart Weibel in a report in April 1999.

The Dublin Core Metadata Initiative regularly produces a status report, and its latest to date is April 2001.

Some of these elements are familiar to anyone who has ever seen a bibliographic record: title, author, subject. Some of these elements are technical in nature, covering information crucial for Web-based documents, like file size. Still others are indicative of the brave new world of cyberspace, like what entity holds the rights to the material.

The developers of the Dublin Core and the World Wide Web Consortium, another group working on standardization, are closely involved with each other, but anyone can invent a metadata scheme, and groups whose focus is on art images or visual modeling, for example, are working on completely different element sets. Metadata is being developed for different kinds of material, but metadata can also be mapped to other metadata schemes. What that means is, for example, that the Dublin Core <title> tag can be related to the MARC record title fields. This kind of mapping holds out some hope for the many metadata schemes being developed for different needs in differing disciplines. A related OCLC project is CORC, the Cooperative Online Resource Catalog, a project exploring the creation and sharing of metadata by libraries.

On the Web, if you View Source in Netscape, metadata elements may look like this (below are some of the Dublin Core metadata elements for this Tech Note, <http://www.pla.org/technotes/metadata.html>)

<META NAME="DC.Title" CONTENT="PLA Tech Note: Metadata">
<META NAME="DC.Subject" CONTENT="Metadata, Public Library Association, SGML, XML">
<META NAME="DC.Description" CONTENT="A basic description, with links, of what metadata is and does, for the Public Library Association">
<META NAME="DC.Publisher" CONTENT="Public Library Association, a division of the American Library Association">
<META NAME="DC.Creator" CONTENT="GraceAnne A. DeCandido">

These tags provide some basic metadata - some basic cataloging information - about what you are reading. You will recognize elements that look like a catalog record: title, author, publisher (PLA).
Metadata can be created when the object it describes is created or it can be added later, as in traditional cataloging. The profusion and proliferation of Web pages seem to almost force the former rather than the latter.

Search engines and metadata

So far, none of the major search engines reads Dublin Core metadata. Some search engines that do (with a little tweaking) include Ultraseek, Swish-E, Microsoft's Index Server, Autonomy Knowledge Server, Blue Angel Technologies MetaStar, and Verity Search 97 Information Server. Search engines work by matching query terms to the words in a document often through a complex algorithm, as Milstead and Feldman point out. If the terms do not match, the engine doesn't find the document. The use of metadata elements can resolve that.

The use of metadata elements, by providing consistent tagging with their own controlled vocabulary, also resolve three specific language problems, according to Milstead and Feldman:
Polysemy: words that have multiple meanings like "spring" or "pitcher." Metadata will enable searchers who want information about water sources, those who want information about seasonal planting, and those looking for a new mattress to find what they are looking for without a lot of extraneous hits. The same is true for those looking for baseball stats or vessels for orange juice.
Synonymy: words that represent the same concept but with different shades of meaning, like "plump" or "fat" or "obese."
Ambiguity: think pitcher, or spring again. Controlled vocabularies or standardized terms used in metadata schemes can retrieve documents precisely even if the actual term is never used in its text.

Garden-variety, basic HTML does offer a meta-tag. Meta-tagging allows the creator to insert words that search engines may pick up, to increase number of hits and to provide a wider net of words for searchers to find, but this is an uncontrolled vocabulary indeed. Meta-tags with irrelevant words, often scurrilous or sexual in nature, can be added to a Web document in a meta-tag in order to attract more users to the site. Because of this kind of faux-metadata, called spamdexing, some engines ignore meta-tags completely.

Metadata elements like the Dublin Core offer a possible method of organizing and accessing the information on the Web. The challenge is not only to create a set of elements that can be universally agreed upon, but to foster widespread usage of those elements.

Large blocks of text, images, correspondence, and other materials that libraries and researchers wish to place on the Web for access and use by all are being made so by SGML, a metalanguage that describes a collection of data in retrievable ways.

SGML, XML
In general discussions of metadata, SGML and lately, XML always come up. SGML and XML are metalanguages (A Gentle Introduction to SGML) and may include metadata elements.
SGML stands for Standard Generalized Markup Language. It employs ASCII code to describe text; virtually all computers understand this code. HTML is a

very simple form of SGML. Using SGML, names, titles, subjects, links between texts and images, can all be identified and searched. "There is almost no limit to the amount of information and intelligence that can be added to text [with SGML] giving it the structural advantages of a database without losing the discursive advantages of text." (Robert DeCandido).

A type of SGML for encoding archival and library finding aids has been developed, called Encoded Archival Description, or EAD. EAD promotes consistency in finding aids in the same institution and among institutions, so that researchers can search many finding aids at once regardless of their location -- an excellent example of how metadata can be a vital tool in the quest for information.

SGML, however, can only be read by specialized browsers, so it must be translated to HTML if it is to have widespread accessibility. SGML's limitations as a tool for the Web may be addressed by XML (Extensible Markup Language). XML is SGML streamlined, relying on tags that almost always come in pairs, and on a new standard called Unicode that supports text in all major languages (see also the PLA Tech Note on Unicode). XML tagged documents can be poured, via stylesheets, into audio format, print, or Web pages.

Talking about metadata sometimes feels like talking about faerie. Information about information about information moves into a deeply abstract realm, more related, perhaps, to epistemology or physics than real words on the page. Chris Armstrong wrote puckishly, in the January 1999 *Information World Review*, that writing about metadata could be described as "metametadata". But metadata may hold the key to finding the precise piece of information on the Web for which we were searching.

Bibliography

Milstead, Jessica and Susan Feldman. "Metadata: Cataloging by any other name ..." in *Online*, Jan/Feb 1999, p24-31.
A very lucid description of what metadata is and does, what the term means, and what the challenges are in implementing metadata schemes.
"Metadata: Projects and standards" same authors, same issue.

DeCandido, Robert, "Metadata: What's it to you?" in *The Internet Searcher's Handbook, 2nd Edition,* Neal Schuman, 1999.

Bosak, Jon and Tim Bray, "XML and the Second-Generation Web," *Scientific American*, May 1999, p89-93.
What XML is and does, by two people who worked on its development.

The indefatigable Robin Cover maintains the most complete coverage of SGML/XML issues.

Chepesiuk, Ron. "Organizing the Internet: The 'Core' of the Challenge," *American Libraries*, January 1999, p60-63.

About the Dublin Core and CORC, the Cooperative Online Resources Catalog.

Dorman, David. "Metadata Musings" *American Libraries*, January 1999, p102.
One straightforward page, clearly presented.

Weibel, Stuart. The State of the Dublin Core Metadata Initiative April 1999.
Bulletin of the American Society for Information Science, June/July 1999.
p18-22.
A shorter version of the complete text in D-Lib Magazine.

Christensen, Deborah. "Golden Retrievers" *School Library Journal*, November 1999. Nice introduction to metadata concepts.

David Dorman, "The Season of Metadata at the Annual Dublin Core Workshop in Ottawa" in *Computers in Libraries*, January 2001. p26-29.

Prepared by GraceAnne A. DeCandido for the Public Library Association, June 13, 1999; reviewed April 2000; links updated May 2001. ladyhawk@well.com

The Public Library Association's Tech Notes project grew out of the desire to continue the work of *Wired for the Future: Developing Your Library Technology Plan* by Diane Mayo and Sandra Nelson, published for PLA by ALA in 1999. Each of the Tech Notes is a Web-published document of 1500-2500 words, providing an introduction and overview to a specific technology topic of interest to public libraries at a particular point in time. Topics were identified by PLA's Technology for Public Libraries Committee. Each Note is marked with the date of its completion and posting, and updates are noted.

Readers' comments and suggestions are welcome and should be addressed to pla@ala.org. Please use *Tech Notes* in your subject line.

PLA home page

Tech Notes Index page

E-Books: I Sing the Book Electric

Filtering: No Easy Answers

Geographic Information Systems (GIS): Mapping the Territory

Video Teleconferencing: Here, There, and Everywhere

Metadata: Always More Than You Think