February 1999

Chaos-Show Me the Metadata!

Ed Summers

Old Dominion University Library, esummers@odu.edu

Sandra K. Paul

SKP Associates, skpassoc@internetmci.com

Albert Simmonds

SKP Associates

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

Summers, Ed; Paul, Sandra K.; and Simmonds, Albert (1999) "Chaos-Show Me the Metadata!," *Against the Grain*: Vol. 11: Iss. 1, Article 40.

DOI: <https://doi.org/10.7771/2380-176X.3746>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Chaos — Show Me the Metadata!

by **Ed Summers** (Electronic Resource Cataloger, Old Dominion University Library, Norfolk, VA) <esummers@odu.edu>

Column Editors: **Sandra K. Paul** (President, SKP Associates) <skpassoc@internetmci.com> and **Albert Simmonds** (SKP Associates)

Perhaps I have been slow to notice it, but over the past couple of years it seems as if the term “metadata” has been popping up with increasing regularity—both inside and outside the library community. Metadata was the subject of many a discussion and poster session at ALA’s most recent meeting in Washington D.C.; and ALA is currently in the process of forming a **Joint Task Force on Metadata**, that will have members from the **Committee on Cataloging: Description and Access (CC:DA)** and the **Committee on Machine-Readable Bibliographic Information (MARBI)**. In May of this year hundreds of librarians descended on **Georgetown University** to attend a joint **ALCTS/LITA** Institute on the entitled “Managing Metadata: Crosswalks or Chaos.” This institute explored the variety of metadata standards that are being used in the information industry—and the potential relationships that exist between them. And it’s not only libraries that are interested in metadata, but organizations such as the **World Wide Web Consortium (W3C)**, and corporations such as **Microsoft** are keenly aware of it, and are in fact generating much of the discussion.

So what is metadata anyway? Is it just another buzz-word that will go the way of other buzz-words, or is it here to stay? If these sorts of thoughts have crossed your mind before, then this article is for you. I intend to provide the metadata-curious with some definitions of the term, some ideas of how metadata is relevant to libraries, and some real-life examples of metadata in action. This article is based on a brief talk that I gave at **Old Dominion University** as part of a series of **Digital Awareness Seminars** this past summer.

What is metadata anyway?

One way to conceptualize metadata is to compare three different ways of searching for information using: a Web search-engine (such as **Altavista**), a Web index (such as **Yahoo**), and an online catalog (such as **WorldCat**). For example, imagine that you are looking for documents on a particular subject using **Altavista** and

Yahoo. **Altavista** is a machine-generated keyword index to millions of documents on the WWW, while **Yahoo** is a human-generated subject index to thousands of Web resources. A typical **Altavista** search on a particular subject can retrieve a list of thousands of citations, which are difficult to browse in their entirety. Although **Yahoo** does not index the magnitude of pages that **Altavista** does (**Yahoo** tends to index homepages rather than any page in a Website), **Yahoo** does provide useful subject indexing that allows you to browse and search within subject areas, while being cross-referenced to other areas of interest.

Now imagine that you are looking for documents that have a particular author. If you conduct a search for an author on **Yahoo** you will get a list of sites that are both about and authored by that person. However, if you are using **WorldCat**, you can generate a list of works that are authored by a specific person (even though **WorldCat** has more records in it than **Yahoo** has sites that it has indexed). **Altavista**, **Yahoo** and **WorldCat** each implement varying types and levels of metadata. **Altavista** uses very little metadata (keywords, and some indexing of HTML tags), **Yahoo** a bit more (with its human-generated subject indexing), and **WorldCat** quite a bit (using the **MARC** format). Actually **Machine Readable Cataloging (MARC)** is an excellent example of a well-developed metadata standard—which is part of the reason why metadata is such a hot topic in the library community. But this still begs the question, “What is metadata?”

The most succinct definition of metadata is that metadata is data about data. Some detractors of this rather brief definition might exclaim, “But wait, you just said that **MARC** is a type of metadata, and we all know that **MARC** isn’t data about data, it’s data about books!” To avoid some of the confusion that the great print/electronic divide can introduce, a more precise definition is that metadata is data about an information object (an information object could be a book, CD-

ROM, microfiche, videocassette, LP, etc.). Considered in this light one can see why **MARC** is considered a type of metadata. For example, the 245 field in **MARC** lets the computer know that the contents of the field are the title and statement of responsibility for a particular information object.

As you might expect, there are more lengthy definitions of metadata, some of which can bring to light the role that metadata plays in the information and library world. A recent draft of the A guide for simple Dublin Core (more about the Dublin Core later) gives the following definition of metadata:

“Metadata describes an information resource. The term ‘meta’ comes from a Greek word that denotes something of a higher or more fundamental nature. Metadata, then, is data about data. It is the Internet-age term for information that librarians traditionally have put into catalogs, and it most commonly refers to descriptive information about Web resources.” (A guide for simple Dublin Core, Section 1.1)

Terrence Smith, Director of the Alexandria Digital Library Project at the University of California at Santa Barbara, provides us with a broader and more holistic definition:

“[Metadata] is the set of all information services accessible to users of the library, together with all available means for coordinating the use of these services, that enable users to access, evaluate, and use any information that may be extracted from the total information resources of the library.” (Smith.)

So, even though the term metadata has come into popular usage during the relatively short lifetime of the World Wide Web, the use of metadata predates the Web by quite a bit. In fact, we can trace the use of

metadata back through the mists of time to 1800 BC when Akkadian scribes created word lists and catalogs of tablet titles for organizing and locating “information objects” in the libraries of Babylon (Logan, p. 68.). Perhaps I digress here, but the point

continued on page 80

“Metadata is the Internet-age term for information that librarians traditionally have put into catalogs, and it most commonly refers to descriptive information about Web resources.”

Chaos
from page 79

I am trying to make is that metadata is hardly a new phenomenon, and that libraries have been deeply involved with metadata for a very long time. However, the “data” in metadata certainly conjures up images of computer technology; and, as the Dublin Core definition indicates, metadata is commonly used in terms of description of information resources on the World Wide Web.

In fact, the Web and its phenomenal growth has inspired an interest in metadata that extends far outside of the bounds of the library community. Furthermore, the Web is now bringing non-library metadata standards into traditional library settings.

Daniel Greenstein of the UK’s Arts and Humanities Data Service sees three major arenas of activity that are generating an increased interest and awareness of metadata: 1) Curatorial traditions: such as libraries, museums and archives, which use their various flavors of metadata to describe, control, preserve and document the resources they manage. 2) Web-based information management: which includes bodies like the World Wide Web Consortium, where corporations such as Sun, Microsoft and Netscape hammer out new standards. 3) Network Resource Discovery: which includes business interests such as Yahoo, Altavista, NorthernLight, OCLC, etc. (Greenstein)

These definitions should provide you with a rough idea of what metadata is, at least in an abstract sense. But apart from MARC, I haven’t really given you any concrete examples of metadata standards, and how they are being used. When I was preparing my presentation for ODU I bounced some ideas off of our systems librarian who summarily said, “That’s fine, but SHOW ME THE METADATA!” — which is what I plan to do with the rest of this space. I will cover several standards that are relevant to libraries and list some URLs which you can use to check out some of the online resources that are using them. MARC is an example of a metadata standard that grew out of the library community in the late 1960s. Similarly, other professional and research communities have developed their own metadata standards. In recent years the Web has broken down many of the traditional barriers between these metadata communities by making large amounts of data available to the Internet audience—and to the patrons in your libraries. So what are these other metadata standards?

The Text Encoding Initiative (TEI)

The TEI began in 1987 as a joint venture of the **Association for Computers and**

the Humanities, the **Association for Computational Linguistics**, and the **Association for Literacy and Linguistic Computing**. The TEI set itself the task of developing guidelines for the preparation and interchange of electronic texts for scholarly research. In 1994 the **Guidelines for Electronic Text Encoding and Interchange** were released, which essentially describe how to “mark-up” electronic texts in the **Standard Generalized Markup Language (SGML)** so that they can be used for scholarly research. A key part of a TEI

“XML will probably be bringing metadata to your desktop and into your libraries even more in the near future.”

encoded document is the TEI header, which bears some resemblance to the MARC record since it includes a great deal of metadata about a text (author, title, publisher, etc.) After the TEI header comes the electronic text itself, which contains specialized markup for genres such as prose, verse and drama, as well as markup to in-

dicating textual elements such as paragraphs, quotations, changes in typeface, dates, etc. Creating electronic texts using the TEI guidelines allows the resulting text to be searched, analyzed and preserved in ways that would be impossible if it were simply a flat ASCII file.

Since 1994, more than 50 different electronic text projects that use the TEI guidelines have sprung up around the world. Notable among these are the **Oxford Text Archive** (<http://ota.ox.ac.uk/>), the **Electronic Text Center at the University of Virginia** (<http://etext.lib.virginia.edu/>), the **Humanities Text Initiative** at the **University of Michigan** (<http://www.hti.umich.edu/>), **Documenting the American South** at the **University of North Carolina-Chapel Hill** (**HYPERLINK** <http://sunsite.unc.edu/docsouth/>, <http://sunsite.unc.edu/docsouth/>), and the **Victorian Women Writers Project** at **Indiana University** (<http://www.indiana.edu/~lettrs/vwwp/>)—to name only a few. Many of these projects allow you to access their electronic texts over the Internet using the freely available **Panorama Viewer** (which allows you to view

continued on page 81

Innovations
from page 78

nies to distribute their music to a broad audience. Upcoming musicians, numbering in the thousands, are using the Internet to their advantage to create awareness in a cost-effective manner, which is clearly a threat to the major record labels’ current distribution model.”

The **RIAA** succeeded in legislating copy protection mechanisms on all recording devices with the passage of the **Digital Millennium Copyright Act**. If it is successful in its current lawsuit, the **RIAA** will tighten even further the restraints on distribution of original music. It seems that the **RIAA** wants to control the distribution not only of its own properties but also those of non-members. This is tantamount to making it illegal to disseminate music not published through a record label. Extending this analogy to print publications would mean that only publications produced and disseminated by a print publisher are worthy of distribution and that anything not so published is not worthy of consideration. Clearly, there is much important material that never gets published for a wide variety of reasons on both the author’s and the publisher’s part.

Publishers often pass on material that another publisher may accept. Poems, letters, and essays often get collected and published many years after their writing. This does not mean that they were unworthy of attention prior to publication. Consider the case of

Doris Lessing who submitted her manuscript for the *Diary of a Good Neighbor* to several different publishers under the pseudonym **Jane Somers** in the early 1980s. The publishers rejected it until she revealed her real identity whereupon it met with immediate acceptance. The success of the **RIAA** lawsuit could give legitimacy to conclusions that only published materials or those published by large companies like **Thomson**, **Bertelsmann**, **Gulf-Western**, **Elsevier**, **Pergamon**, **McGraw-Hill**, etc. are worthy of distribution.

Instead of accepting the inevitability of digital distribution and developing ways to implement it, the **RIAA** staunchly opposes the introduction of new media and new formats. The **RIAA** could allow audiophiles to create their own CDs by purchasing music on a tune-by-tune basis, just as print publishers charge by the search or by the article viewed or produced. Some of the MP3 Web sites have the capability of encoding selections to verify that nominal fees are paid upon access or downloading and for tracking pirated copies. There’s also an unremovable watermarking technology, such as that offered by **Aris Technologies Inc.** and used by companies such as **Nordic Entertainment**. This software allows Internet-scouting search engines to track down an embedded artist code to trace a file back to the original purchaser. However, instead of advocating change, the **RIAA** seems to be saying the status quo is the best way to achieve change. 🌱

[<http://www.against-the-grain.com>](http://www.against-the-grain.com)

SGML documents and can be downloaded at <http://www.softquad.com>, or by translating SGML to HTML "on the fly" so that you can view the text in your favorite Web browser. The real utility of electronic texts that are encoded, according to the TEI guidelines, is in the metadata that they provide computer systems and humans to perform fielded searching, instead of brute force keyword indexing. Bear in mind that the TEI Guidelines and SGML are fairly complex (much like MARC)—but the TEI anticipated this, and very considerably released TEI-lite (HYPERLINK <http://www-tei.uic.edu/orgs/tei/intros/teiu5.html>, <http://www-tei.uic.edu/orgs/tei/intros/teiu5.html>) that allows new users to quickly pick up the key elements of the standard, and to start using them.

Encoded Archival Description (EAD)

Like the TEI, the EAD is a standard that has its origins in the academic community. However, the EAD has grown out of an area that is particularly close to the library's heart—the archive. Work on the EAD began at UC-Berkeley in 1993 as an investigation into the feasibility of developing a non-proprietary encoding standard for machine readable finding aids. Finding aids are used in the archival world to provide access to their collections. They usually take the form of indexes, registries, inventories or guides which allow researchers to locate primary resource materials in archives, manuscript repositories, museums and libraries. Basically, the EAD effort is analogous to the development of the MARC standard to allow libraries to convert their card catalogs into machine-readable form. The **Berkeley Finding Aids** project received funding from the **U.S. Department of Education** to examine the best way to create machine readable finding aids. The BFAP participants were looking for a solution that would: a) allow for the description of archival collections; b) preserve the hierarchical relationships within and between collections; c) allow for easy navigation of collections; and d) enable search and retrieval within and across collections. They considered several technologies as possible solutions, including: GOPHER, HTML, MARC and SGML. The project concluded that the best solution to all of their requirements was to use SGML (following the same path as the TEI).

Early in 1996 an alpha version of the SGML document type definition for archival finding aids was released which became known as the EAD. The EAD was then tested, revised as a beta version, and tested again—and just this past August EAD 1.0 was released. In its short lifetime the EAD has received a great deal of support from the library and archival community—primarily because of the need it fills, and the support it has received from the **Library of Congress** and the **Society for American Archivists**. Fifteen years ago RLG began providing online access to finding aids through its **RLIN Archives and Manuscripts Database**, which used the

"SGML is a very powerful tool that the publishing community has been using for over two decades to create electronic texts that contain metadata."

MARC AMC format. This year RLG announced that it has converted its MARC AMC finding aids into the EAD format, and has made them available on the Web (for a price of course) as part of their **Archival Resources** service (HYPERLINK <http://www.rlg.org/arrhome.html>, <http://www.rlg.org/arrhome.html>). Interestingly, Archival Resources will index your EAD finding aids that are located on your institution's Web server—and will include this information in their database (at no cost). There are many other institutions using the EAD to provide access to their archival collections—a thorough list is available at the **Library of Congress Network Development and MARC Standards Office** (which maintain the EAD standard) at HYPERLINK <http://lcweb.loc.gov/ead/>, <http://lcweb.loc.gov/ead/>.

Geospatial Metadata

Another type of metadata that is of increasing importance to libraries is

"And it's not only libraries that are interested in metadata, but organizations such as the World Wide Web Consortium (W3C), and corporations such as Microsoft are keenly aware of it..."

geospatial metadata. Geospatial metadata allows researchers to locate relevant data sets for use in **Geographic Information Systems (GIS)**, in much the same way that MARC allows researchers to locate books and journals that are relevant to their research. The **Content Standard for Digital**

Geospatial Metadata (CSDGM) is the premier geospatial metadata format, and was developed by the **Federal Geographic Data Committee (FGDC)** in 1994. In 1990 the FGDC was formed as an interagency committee of the U.S. federal government to promote the use, sharing and dissemination of geospatial data in the U.S. Since then, the FGDC has played an instrumental part in the formation of the **National Spatial Data Infrastructure (NSDI)**, which was organized by **Bill Clinton** to promote the sharing of geospatial data throughout all levels of government, the private and non-profit sectors and the academic community. In fact, Clinton's **Executive Order 12906** requires that all geospatial data produced by the federal government should be described according to the CSDGM metadata standard.

The CSDGM is a fairly complex set of 334 descriptive elements to be used in describing geospatial data sets. In June of this year, the FGDC approved the second edition of the CSDGM standard, and the FGDC is currently working closely with the **International Organization for Standardization** to develop an international metadata standard for geospatial data (which more than likely will bear a strong resemblance to the CSDGM). Currently, the **Geospatial Data Clearinghouse** (<http://130.11.52.178>) allows you to conduct distributed searches across fifty different geospatial data archives across North America. This type of search is only possible because of the agreement between the various repositories on the type of metadata they will use to describe their collections (CSDGM). In fact, if you are interested in establishing a clearinghouse node for your area, the FGDC is making grant money available for just that—and offers a wide array of metadata training to set you on your way.

The Dublin Core

In 1995 OCLC (an acronym familiar to the library world) and the **National Center for Supercomputing Applications (NCSA)** hosted a workshop in Dublin, Ohio, that focused on the description of networked resources. This was the first, in what have been five workshops, which have brought together an international group of library and computer network experts to develop a core element set for description of information resources on the WWW. This set has become known as the Dublin Core, and it is aimed at improving resource discovery on the Web. The **Dublin Core Initiative** has stressed the need for a simple yet extensible metadata standard—so that it can be implemented relatively easily,

continued on page 82

while accommodating the descriptive needs of specific communities.

The Dublin Core is currently a set of fifteen optional and repeatable elements, which fall into three basic areas: Content: Title, Subject, Description, Source, Language, Relation and Coverage. Intellectual Property: Creator, Publisher, Contributor, and Rights. Instantiation: Date, Type, Format and Identifier.

The Dublin Core has proven to be quite a popular metadata standard around the world, precisely because of its simplicity, extensibility and its international/interdisciplinary focus. For an up to date list of projects that are using the Dublin Core take a look at the Dublin Core Initiative homepage (HYPERLINK http://purl.oclc.org/metadata/dublin_core/, http://purl.oclc.org/metadata/dublin_core/). Among those listed there is the **University of Wisconsin-based Scout Report**, which is using the Dublin Core to organize and provide access to its growing collection of Internet resources that are of interest to the academic community. The Scout Report has incorporated **Library of Congress Subject Headings** and **LC Classification** into the Dublin Core metadata scheme. These records are stored in a **FileMaker Pro** database, and are searchable and browsable over the Web through their **SignPost** service at (HYPERLINK <http://www.signpost.org/>, <http://www.signpost.org/>). This site is well worth a visit to see how Dublin Core can be used effectively to provide access to electronic resources. Since it is a "core set," the Dublin Core is also becoming an important standard for mapping between metadata standards. For example, if you wanted to be able to search across collections of MARC, TEI, EAD and CSDGM metadata records, the Dublin Core could be used as a common denominator between these very different metadata records. Work in this area is just beginning, and may result in centralized metadata registries which will maintain these useful mappings between metadata sets. There are already mappings between DC and MARC, DC and GILS (Government Information Locator Service), DC and CSDGM, and DC and Z39.50. A useful listing of these mappings is available at: HYPERLINK <http://vads.ahds.ac.uk/metadataaf5.html>, <http://vads.ahds.ac.uk/metadataaf5.html>.

As you can see from this brief survey, metadata standards have a tendency to grow out of specific communities: MARC from the libraries, TEI from the humanities scholars, EAD from archives, and CSDGM from the GIS community. The Dublin Core is slightly different from the others since from the beginning it has stressed the need

for an interdisciplinary metadata standard for resource discovery. However, there are many, many more metadata standards being used, such as: the Mathematical Markup Language, the Chemical Markup Language, the Platform for Internet Content Selection (which drives many Internet filtering packages), and the Synchronized Multimedia Integration Language (SMIL). An excellent resource for exploring the many different metadata standards that are available, as well as the ones already introduced here is **Metadata Clearinghouse** from IFLA at HYPERLINK <http://www.ifla.org/II/metadata.htm>, <http://www.ifla.org/II/metadata.htm>.

Afterword

You may have noticed that the computer language **SGML** has been mentioned several times during this discussion of metadata. The **Standard Generalized Markup Language** is a meta-language—yes, a language for describing languages! A whole article could be written on SGML alone; in fact many lengthy books have done just that. Suffice it to say that SGML is a very powerful tool that the publishing community has been using for over two decades to create electronic texts that contain metadata. SGML was used to define the EAD and TEI metadata standards. In the past, **Netscape** and **Microsoft** have not supported SGML document viewing in their browsers. Furthermore, the SGML learning curve is quite steep, and has typically required a great deal of organizational commitment to implement it effectively. These two factors have somewhat diminished SGML's user base—especially on the Web. However, early this year the **World Wide Web Consortium** released **The Extensible Markup Language (XML)** as a recommendation. XML is a subset of SGML, and was designed to be much easier to use than SGML, while allowing for metadata and other types of tagging which HTML doesn't support. **Netscape** and **Microsoft** were heavily involved in the creation of XML, and have committed to including support for XML in their products. The long and short of this is that XML will probably be bringing metadata to your desktop and into your libraries even more in the near future.

Useful Links

Digital Libraries: Metadata Resources. IFLA. HYPERLINK <http://www.ifla.org/II/metadata.htm>, <http://www.ifla.org/II/metadata.htm>

The Text Encoding Initiative. HYPERLINK <http://www.uic.edu/orgs/tei/>, <http://www.uic.edu/orgs/tei/>

TEI Application Page: Project Descriptions. HYPERLINK <http://www.tei.uic.edu/orgs/tei/app/index.html>, <http://www.tei.uic.edu/orgs/tei/app/index.html>

The Encoded Archival Description Official Web Site. HYPERLINK <http://lcweb.loc.gov/ead/>, <http://lcweb.loc.gov/ead/>

Encoded Archival Description Sites on the World Wide Web. HYPERLINK <http://www.loc.gov/ead/eadsites.html>, <http://www.loc.gov/ead/eadsites.html>

Archival Resources. Research Libraries Group. HYPERLINK <http://www.rlg.org/arrhome.html>, <http://www.rlg.org/arrhome.html>

The Federal Geographic Data Committee. HYPERLINK <http://www.fgdc.gov/>, <http://www.fgdc.gov/>

Geospatial Data Clearinghouse. HYPERLINK <http://130.11.52.178/gateways.html>, <http://130.11.52.178/gateways.html>

The Dublin Core Initiative. HYPERLINK http://purl.oclc.org/metadata/dublin_core/, http://purl.oclc.org/metadata/dublin_core/

Projects Using Dublin Core. HYPERLINK http://purl.oclc.org/metadata/dublin_core/projects.html, http://purl.oclc.org/metadata/dublin_core/projects.html

Extensible Markup Language (XML). World Wide Web Consortium. HYPERLINK <http://www.w3.org/XML/>, <http://www.w3.org/XML/>

The SGML/XML Web Page. HYPERLINK <http://www.oasis-open.org/cover/>, <http://www.oasis-open.org/cover/>

References

A user guide for simple Dublin Core. Draft version 5.1. Revised 31 July 1998. Online. Available: HYPERLINK <http://128.253.70.110/DC5/UserGuide5.html>, <http://128.253.70.110/DC5/UserGuide5.html>. 15 September 1998.

Greenstein, Daniel. "Metadata and middleware: a system architecture for cross-domain discovery." in *Discovering online resources across the humanities: a practical implementation of the Core*. Edited by Paul Miller and Daniel Greenstein. Bath, UK: UK Office for Library and Information Networking, University of Bath (UKOLN), 1997.

Guidelines for electronic text encoding and interchange. Edited by C.M. Sperberg-McQueen and Lou Burnard. Chicago, Oxford: Text Encoding Initiative, 1997. Online: HYPERLINK <http://etext.lib.virginia.edu/TEI.html>, <http://etext.lib.virginia.edu/TEI.html>. 15 September, 1998.

Logan, Robert K. *The alphabet effect*. New York: St. Martin's Press, 1986.

Smith, Terrence R. "The meta-information environment of digital libraries." *D-Lib Magazine*, July/August 1996. Online. Available: HYPERLINK <http://www.dlib.org/dlib/july96/new/07smith.html>, <http://www.dlib.org/dlib/july96/new/07smith.html>. 15 September 1998.