Introducing Metadata

Richard P. Smiraglia

The purpose of this volume is to provide a learning resource about metadata for catalog librarians and students. Part I of the volume contains introductory, theoretical material, and original research, and Part II contains instructive material. The point of view of the volume, overall, is that in library and information science there is an ongoing convergence of cataloging and metadata, such that the community will benefit from instructional material that demonstrates this convergence. Throughout the volume, readers will find metadata discussed alongside cataloging. The rationale for this is fundamental-resource description. Metadata, in one sense, are nothing new, and yet, in another sense, metadata schemes represent a very exciting forefront in resource description.

Some basic texts on metadata are already in wide circulation, and readers might want to have them handy for reference. These are, chronologically:


The essay that follows is introductory. I have subtitled the volume “a cataloger’s primer.” A primer is a kind of textbook used (back in the day) to teach reading, arithmetic, or other basic fundamentals. This essay, therefore, contains some elements necessary for understanding metadata in the bibliographic context. A set of basic operational definitions is expressed, the evolution of metadata in the cataloging community is outlined, and basic metadata techniques are described. The concluding section of this essay is a direct introduction to the content of this volume. Let us begin by defining metadata.

**THREE DEFINITIONS**

Part I of this volume is devoted to foundational issues in metadata, and the primary issue is the search for a definition. The authors whose papers follow have devoted careful thought to the issue, and very specific definitions emerge from their labor. But we must have a beginning point, so for now, let us define metadata as structured descriptors of information resources, designed to promote information retrieval. That is, at the most basic level, metadata describe information resources—such as websites, electronic texts, digital artifacts, etc.—through the compilation of descriptors that are structured according to a specific framework (called a scheme), and are placed together to serve an indexing function in information retrieval.

The Association for Library Collections & Technical Services (ALCTS, a division of the American Library Association (ALA)) has a committee known as the Committee on Cataloging: Description and Access (or CC:DA). CC:DA governs ALA’s role in all standards for resource description, including the *Anglo-American Cataloguing Rules*. A 1999 report set out the parameters of library involvement in metadata (CCDA 1999, tf-meta3.html). The task force sought to understand and
explicate the role of metadata in the realm of library cataloging by first and foremost casting the role of metadata under the aegis of resource description:

Our RESOURCE DESCRIPTION NEEDS are grounded in the needs of our users to find, identify, select, and obtain some information thing (book, article, map, score, data set, etc.). We judge our tools—catalogs, indexes, search engines, etc.—primarily by how well they do these tasks. And not only must they perform the user tasks well, but they must make the management task or housekeeping as simple, easy, flexible, and cheap as possible.

So we see here the primary purpose of metadata described in terms of the traditionally dual roles of retrieval and inventory. Retrieval must be user-centered, inventory is institution-centered, and both require a level of efficiency and standardization. The task force understood that a new age had arrived in the catalog department—one in which traditional cataloging had given way to a multiplicity of tools and techniques for providing access both in-house and via external resources, primarily those located on the World Wide Web. A central tenet of the new age is the search of interoperability—a way to move seamlessly from one tool to another. Interoperability must also demonstrate user-centeredness in retrieval, and institution-centeredness in inventory:

Our catalogs have become one tool among many, but those many are not separate or isolated from one another. The catalog is one tool in a network of tools. THE BASIC OR NECESSARY PRINCIPLE OF TOOLS IN A NETWORK OF TOOLS IS INTEROPERABILITY . . . Or if interoperability is too high a demand, each tool in a network of tools must be compatible with each other from the user’s perspective. For example, a user employing an institution’s catalog to find data sets relating to census and voting in Hartford, Connecticut may well need to analyze the data once [they have] been located, and then format the analysis into a presentation document dominated by images not numbers. While the catalog per se is only one tool in this scenario, it should be compatible with a wide range of other tools that may be used as functional extensions of the discovery and retrieval process. What is desirable is a network of tools that are portable, flexible, agile, mappable, extensible, adaptable, a coherent network of tools. The library catalog can be part of such a coherent environment, but
only if it is designed, maintained, and used as one tool in a network of tools.

The task force submitted three formal working definitions:

- **METADATA** are structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities.

- **INTEROPERABILITY** is the ability of two or more systems or components to exchange information and use the exchanged information without special effort on either system.

- **A METADATA SCHEME** provides a formal structure designed to identify the knowledge structure of a given discipline and to link that structure to the information of the discipline through the creation of an information system that will assist the identification, discovery, and use of information within that discipline.

### **METADATA CONCEPTS**

The relationship between metadata and the knowledge artifacts these data represent (or, for which metadata serve as surrogates) is direct. In most cases, metadata are transcribed inherent data; that is, the data are taken directly from the resource and then reassembled according to the schema in such a way as to create a representation of the resource. Caplan (2003, 3) says metadata are “structured information about an information resource of any media type or format.” Key terms here are “structured” and “information resource.” There are many kinds of metadata at use in traditional library cataloging. In each instance, a structured description represents a resource—a journal article and its citation, a book and its catalog record, an electronic resource and its metadata. At the most elemental level, metadata provide structured terms that enable access to resources through information retrieval systems ranging from indexes, to catalogs, to search engines. For this purpose the metadata are intended to be user-centered, to provide the best possible access to specific resources. On another level, metadata vary by institutional purpose, and these go beyond the merely descriptive. Some are administrative, used by repositories to manage their resources, while others are structural, used for inventory control or to manipulate electronic information.
CONTENT DESIGNATION AND MARKUP LANGUAGES

Content designation is a term coined to describe the insertion of structural elements into plain text to facilitate computer manipulation. In library cataloging, content designation is used primarily to segment the components of a catalog record. In publishing, markup languages provide a form a content designation to drive the printing process. And, of course, in the World Wide Web markup languages are used to compel browsers to display content in a specific way. Markup languages provide vocabulary and syntax, which, when entered into a document, provide cues for computer manipulation of the text. It is markup language that turns normal text into a website.

International Standards for Bibliographic Description: Punctuation as Markup

The cataloging world met together in Paris in 1961 to create an international framework for the construction of bibliographic data. Even in those early days for library automation, there was great awareness that the ability to share data across linguistic and cultural boundaries would be vital for the control and dissemination of recorded knowledge. As a result of the Paris Conference, principles (known as the “Paris Principles”) were formulated and agreed to. It is these principles that underlie the structure of AACR2, and most other catalog codes today (see ICCP 1971). One of the first implementations was a new, internationally agreed-upon framework for the descriptive portion of a bibliographic record (the title transcription, through the series transcription and annotations). Disseminated in 1974 in the first generic ISBD (International Standard Bibliographic Description), these conventions quickly became the norm worldwide.

A major aspect of ISBD description was the inclusion of “prescribed-punctuation.” The purpose of prescribed-punctuation was to provide cues about the content of a bibliographic record, regardless of the user’s ability to comprehend the language. Prescribed-punctuation, then, was an early form of markup, intended to cue users (and eventually, it was thought at the time, computers) about the contents of a record. For example, look at the following bibliographic record, which is in a made-up language:

The punctuation, which always precedes an element, delineates the parts of this record. The title is followed by a statement of responsibility, which must be preceded by a space-slash-space, thus the title must be:

Rhksow fjksw bf kskk jsiusol

because the statement of responsibility is:

w Hfuyse can Lqzx.

The conventions of ISBD punctuation can be found in AACR2. A summary:

. -- (full-stop, space, dash, space) precedes a new area of description

/ (space, slash, space) precedes a statement of responsibility

: (space, colon, space) precedes the second element of an area (the publisher in area 4, the illustrations in area 5)

; (space, semi-colon, space) precedes the third element of an area (a second author in area 1, a second city or publisher in area 4, the dimensions in area 5)

**Machine-Readable Cataloging (MARC) as Markup**

No discussion of “markup” would be complete without a nod to the MARC coding language, which has fueled the great international effort to make catalogs electronic and to share catalog data worldwide via computer transmission. Essentially, catalog data are compiled according to standards (mostly AACR2) then marked up with MARC. The MARC tags, which one can view on OCLC or in “full” displays in online catalogs, but which are not visible to the searching public, designate for the computer the contents of fields and sub-fields. Their function is similar to that of the ISBD punctuation, but the language of MARC is much more complex. Here is a MARC markup of the bibliographic record from the preceding example:
MARKUP LANGUAGES IN PUBLISHING

In the early automation of publishing, markup was used to set cues within an author’s text, which would tell a typesetting program how to set the type when it printed out the book (article, etc.). A simple version might look like this:

```html
<b><t>Introduction to Markup Languages</t></b><a>by Jon Smith</a><pl>Chicago</pl><pu>Silly Press</pu><b><d>2001</d></b>
```

This markup (which I also invented) might turn that text into a title page something like this:

```
Introduction to Markup Languages
by Jon Smith

Chicago
Silly Press
2001
```

Note that each element is marked on both ends; that is, text is enclosed between a start tag “<a>” and an end tag “</a>.”

STANDARD GENERALIZED MARKUP LANGUAGE (SGML)

SGML was the first “meta” markup language. Developed to serve as a standard platform for the development of other languages, SGML provides conventions for naming the logical elements of documents, and syntax for expressing the logical relations among document components. SGML was intended to be used by specific communities to de-
velop specific markup languages, known as Document Type Definitions or DTDs. Many of the metadata schemes described in this volume, are in fact, SGML-derived DTDs.

**HYPERTEXT MARKUP LANGUAGE (HTML)**

HTML is an SGML DTD that underlies the World Wide Web. HTML is the source code that resides behind the displayed website, telling browsers how to display the text to the viewer, and serving as source data for search engines. According to Graham (1995), HTML requires a document to be constructed with sections of text marked as logical units, such as titles, paragraphs, or lists, and leaves the interpretation of these marked elements up to the browser displaying the document.

An HTML document is composed of elements, which are marked by tags. Some elements do not affect a block of text (such as a paragraph command); these are called empty elements, and do not require end tags. Element names and attributes (which instruct the browser but do not display) are case-insensitive. But the attribute value (the text that will display) is case-sensitive.

An HTML document has two main elements: HEAD and BODY. Each main element has sub-elements. The TITLE sub-element is the only required element of HEAD. The BODY has many sub-elements, such as:

- **Headings**, which come in six levels:
  - `<H1> ... words ... </H1>`
  - `<H2> ... words ... </H2>`
  - `<H3> ... words ... </H3>`
  - `<H4> ... words ... </H4>`
  - `<H5> ... words ... </H5>`
  - `<H6> ... words ... </H6>`

  These tags cause headings to display in different sizes of type, from large, bold-face (h1) to small type (h6).

- **Highlighting**, which gives special emphasis:
  - `<EM></EM>` will render the phrase in *italics*
  - `<STRONG></STRONG>` will render the phrase in bold.

- **Paragraphs**, an empty element, causes the text to break into paragraphs `<P>`.
Break is similar <BR>.
Lists cause a list to appear indented and bulleted. Lists may be unordered (ul) or ordered (ol):

<UL>
List items, each tagged with <LI>
</UL>

Horizontal Rule draws a horizontal line across the page <HR>.

Hypertext Links can be used to move between documents:

<A HREF="http://smiraglia.org">Click here for my Vita</A>

Images can be embedded in a webpage. For instance, a still image in the form of a graphical interface file (gif) can appear to be embedded in the website by using a hyperlink:

<IMG SRC="portrait.gif">

Tables format text into tabular form. The following code creates a table with three columns and two rows:

<TABLE>
<TR><TD>first data</TD><TD>second data</TD><TD>third data</TD></TR>
<TR><TD>fourth data</TD><TD>fifth data</TD><TD>sixth data</TD></TR>
</TABLE>

Here is a little exercise. Open Notepad and type the following:

<html>
<head>
<title>A Sample HTML Document</title>
</head>
<body>
<h1>A Sample HTML Document</h1>
</body>
</html>

Save the file as “all files” and give it the file extension “.html” then open the file using Netscape or Internet Explorer. You should see the line of text in large bold characters. Notice that your “title” appears in the blue bar at the top of your screen. Using the “view” menu click on “Page source” and you will see your original html file. Now go back to Notepad and add another line of text, thus:
METADATA FOR RESOURCE DESCRIPTION

Metadata such as catalog records and index citations have been used now for thousands of years (literally since antiquity). Always there has been a yearning among knowledge organization professionals to find more efficient and accurate means for providing resource description. Yet, even now, metadata mostly are compiled by individuals working with loosely-defined standards. Here we introduce several metadata projects and schemes in simple terms and in chronological order, to set the stage for the remainder of the volume.

Cataloging in Publication

In the early twentieth century (1901 in fact) the Library of Congress began to make copies of its catalog cards available for purchase by librarians. This was the real beginning of cooperative cataloging. For any book for which the Library of Congress had prepared cataloging, you (the local librarian) were freed from that effort. All you had to do was buy the cards, type added entries on top of them and call numbers in the upper left corner, and then file the cards. Savings were dramatic. As a result, standardization of cataloging spread across the United States, then North America, then throughout the English-speaking world, as cooperation grew among the Library of Congress, the British Library (then the library of the British Museum), and the National Library of Canada.
In the 1950s, there were many projects undertaken to provide copies of proof sheets for LC cards in the books libraries were buying as new acquisitions. This meant that, if your jobber participated in the program, the mere act of buying the book also brought with it the professional and standardized cataloging. This was pretty close to in-source metadata for the time. Beginning in 1961, publishers and librarians in the U.S. (and later worldwide) began to cooperate on a larger scale, implementing a project known as Cataloging in Publication, or CIP. CIP is a form of metadata literally in the resource, printed always on the verso of a book’s title page. If we extend this notion to the arena of markup languages, we arrive at the promise of self-describing resources.

Text Encoding Initiative (TEI)

The Text Encoding Initiative (TEI) is an international project to develop guidelines for the preparation and interchange of electronic texts for scholarly research, and to satisfy a broad range of uses by the language industries more generally. The TEI is sponsored by the Association for Computers and the Humanities (ACH), the Association for Computational Linguistics (ACL), and the Association for Literary and Linguistic Computing (ALLC). Major support for the project has come from the U.S. National Endowment for the Humanities (NEH), Directorate XIII of the Commission of the European Communities (CEC/DG-XIII), the Andrew W. Mellon Foundation, and the Social Science and Humanities Research Council of Canada. There is a pressing need for a common text encoding scheme researchers can use in creating electronic texts. Three organizations sponsor TEI: the Association for Computers and the Humanities (ACH), the Association for Computational Linguistics (ACL), and the Association for Literary and Linguistic Computing (ALLC). (See the TEI Homepage http://www.tei-c.org/, and Text Encoding in Libraries http://www.indiana.edu/~letrs/tei/.)

The Dublin Core Metadata Set

One prominent metadata scheme is known as the Dublin Core. It was developed using large government-funded grants that brought together teams of librarians and software and Internet experts to create a “core” set of metadata elements and coding conventions. The intent of this program was (a) to share the expertise of both communities; and (b) to prevent either community from creating a standard that would be unusable by the other.
Using Dublin Core is a usage guide by Diane Hillmann, which can be found at http://dublincore.org/documents/2001/04/12/usageguide/.

Dublin Core metadata sets reside in the header of web-resources, fueling search engines. With a metadata set a web-resource becomes, essentially, self-describing. Core elements are described at http://purl.org/metadata/dublin_core_elements and http://dublincore.org/documents/dces/. Other instructional material can be found at http://dublincore.org. The Dublin Core set is very much a work in progress, as you can see at the homepage. That is, it is a dynamic standard, being updated and revised more or less constantly. This means that keeping an eye on changes will be critical for staying up to date with the DCMI. Syntax is described in the usage guide. A basic Dublin Core set looks like this:

```
<dc:creator>Rose Bush</dc:creator>
<dc:description>Describes process for planting and nurturing different kinds of rose bushes.</dc:description>
<dc:date>2001-01-20</dc:date>
```

Encoded Archival Description (EAD)

The growth of the Web created demand for publishing of archival repositories’ finding aids on websites. To accommodate this activity, EAD was created. The EAD Document Type Definition (DTD) is a standard for encoding archival finding aids using the Standard Generalized Markup Language (SGML). The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress (LC) in partnership with the Society of American Archivists. Development of the EAD DTD began with a project initiated by the University of California, Berkeley, Library in 1993. The goal of the Berkeley project was to investigate the desirability and feasibility of developing a nonproprietary encoding standard for machine-readable finding aids such as inventories, registers, indexes, and other documents created by archives, libraries, museums, and manuscript repositories to support the use of their holdings. The project directors recognized the growing role of networks in accessing information about holdings, and they were keen to include information beyond that which was provided by traditional machine-readable cataloging (MARC) records. (See the official EAD homepage at http://www.loc.gov/ead/ and an overview at http://sunsite.berkeley.edu/FindingAids/EAD/eadmodel.html.)
Descriptive Cataloging—AACR2 and MARC 21 as Metadata

The Anglo-American Cataloguing Rules, now in advanced permutations of the second edition, provide the primary means for resource description in libraries. AACR2 is an internationally managed standard, based on the provisions of the International Standard Bibliographic Description (ISBD), and carefully extended to provide description of and access to resources in all media collected by libraries. As such, bibliographic descriptions constructed using AACR2 constitute descriptive metadata. We already looked briefly at MARC and its role as a kind of markup, and Jul (2003) describes MARC 21 alongside other markup languages. In fact, most AACR2 bibliographic records are constructed within the framework of the MARC 21 format, which facilitates their exchange between and among online systems. Several authors have described AACR2 and MARC 21, therefore, as a metadata-suite, though it is a stretch to consider them as a metadata scheme. In particular, Tillett (2003) describes ways in which AACR2 has evolved to provide metadata-like descriptions of electronic resources.

Essentially, AACR2 describes the content of a resource and MARC 21 provides the markup to facilitate record manipulation. Both AACR2 and MARC 21 make use of syntax, and the interplay of the two can be somewhat confusing. I will not describe AACR2 or MARC 21 in detail in this volume: instead readers are referred to general cataloging texts (see Wynar 1992, for example), and those interested in cataloging for the Web should refer to the recent volume by Jones, Ahronheim, and Crawford (2002). However, it is useful to understand the framework that undergirds bibliographic description in the context of the creation of metadata. Basic elements are transcription from prescribed sources. That is, catalogers do not generate free-text descriptions for the most part. Rather, the content of a catalog record is intended to provide a photo-like image of the actual item by literally copying (transcribing) inherent elements such as titles, names, etc., from specific parts of the resource. In bibliographic terms we say that cataloger transcribes descriptive detail from sources such as title pages and their versos, and then makes reference to other parts of the resource as well. For instance, the table of contents will be consulted in the analysis of subject content, and it might be partially transcribed to enhance retrieval. These data, then, are arrayed according to ISBD-proscribed syntax within eight areas of description:
Once the description is complete, access points are selected and formulated. Subject headings and classification are provided from additional tools (usually Library of Congress Subject Headings and Library of Congress Classification or Dewey Decimal Classification). These, together with the bibliographic description, constitute the descriptive metadata. Then the entire package is encoded using MARC 21 field tags and subfield indicators, and the complete record is integrated into the catalog. Unlike most of the metadata schemes, bibliographic metadata constructed in this manner reside in the catalog as parallel surrogates for the resources, rather than residing within the resource (although, as I noted above, CIP serves that role to some extent).

**PAPERS IN THIS VOLUME**

Part I includes papers that are intended to introduce readers to concepts of metadata in the bibliographic sphere. The volume opens with two papers that present essential metadata concepts and trace their history within the context of the bibliographic resource description community. Jane Greenberg opens the volume with “Understanding Metadata and Metadata Schemes.” She provides a basic framework for the study of metadata schemes, while thoroughly analyzing the definitions and attributes of metadata of all types. Next, Lynne Howarth provides a historical evaluation of the evolution of metadata and bibliographic control. Then follow four papers that present original research, to demonstrate ways in which we have only begun to attempt to exploit the potential of metadata for resource description. D. Grant Campbell uses literary concepts of metaphor and metonymy to demonstrate how metadata function as a sort of language. Leatrice Ferraioli explores the creation and use of personal metadata in a business environment. Jennifer Cwiok surveys various metadata approaches to the identifica-
tion of creators of resources. And Richard Smiraglia describes the evolution and use of content metadata for museum artifacts.

Part II of the volume is intentionally instructive, and comprises papers that introduce metadata schemes and demonstrate their use. Anita S. Coleman describes the Dublin Core, Alexander Thurman demonstrates the use of Encoded Archival Description (EAD), and Patrick Yott explains Extensible Markup Language (XML). Linda Cantara describes the Metadata Encoding and Transmission Standard (METS) designed for exchange and preservation of information packages in the digital library environment. And Michael Chopey contributes a discussion of how to plan and implement a metadata-driven digital repository. In the end the contents go beyond the definition of primer as “introductory textbook.” But the authors have collectively compiled a thought-provoking volume about the uses of metadata.

WORKS CITED


