GENERAL INFORMATION ON METADATA

What is metadata?

Before looking at the major metadata standards and their applications, you must understand what metadata is. Metadata is data about data—or better: structured data about data. Here are some more concrete approaches to the word *metadata:*

- A succinct and systematic set of information that references what can be used to efficiently and accurately retrieve a larger set of information (Morville, 1999)
- Metadata is "the sum total of what one can say about any information object at any level of aggregation" (Gilleland-Swetland, 2000)
- Data associated with objects that relieve their potential users of the need for full advance knowledge of their existence and characteristics (DESIRE project, 1997)
- Any formal scheme of resource description, applying to any type of object, digital or nondigital (Hodge, 2001)
- Structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities (ALCTS CC:DA/MARBI task force on metadata, 2000)
- Any data that aid in the identification, description, and location of networked electronic resources (Hudgins, 1999)

One misunderstanding must be corrected; it appears in the last bulleted item above. Metadata is not for electronic resources only; it is for resource description of any kind. In fact, libraries, archives, and other information organizations have been producing metadata for resources of all types for hundreds of years through cataloging.

Metadata *is treated as a singular noun in this report, contrary to the way the word data is used.* Data *is most often treated as a plural noun.*

Although cataloging doesn't have the overwhelming popularity that follows the buzzword *metadata,* the two terms can be thought of interchangeably.

Since 1971, librarians have been producing metadata using computer technology and standards enabled by the Online Computer Library Center (OCLC) located in Dublin, Ohio. This shared, online cataloging system is used by more than 41,000 libraries in more than 82 countries and territories around the world. OCLC maintains the world's largest bibliographic database; WorldCat holds more than 47 million catalog (*read* metadata) records, with a new record added every 15 seconds.

Why is metadata important?

Creating metadata is important because metadata facilitates the discovery of relevant information and resources, digital and nondigital. Metadata helps identify resources, distinguish among dissimilar resources, bring similar resources together, allow resources to be found by relevant criteria, and give location information.

Metadata promotes interoperability if accompanied by careful mapping of data elements and crosswalking of standards. Interoperability allows multiple systems to exchange data with minimal loss of content and functionality, regardless of different hardware and software platforms, data structures, and interfaces. The use of metadata allows resources to be searched seamlessly across networks through crosswalks and shared transfer protocols. Chapter 4 contains more information on the status and importance of interoperability.

Metadata ensures resources will be accessible into the future. Certain metadata efforts are focused on the long-term preservation of digital information. Special information is needed to track the lineage of a digital object (where the object came from and its current manifestation), to document its behavior, to detail its physical characteristics, and to record preservation decisions such as migration and emulation information. A section on preservation metadata efforts is included in Chapter 3.

Metadata can provide persistent and unique digital identification that assists in differentiating one object from another. These standard numbers or elements validate information and assist in locating and uniquely identifying the work or object to which the metadata refers. Chapter 3 contains a section on object identifiers.

Metadata also documents and tracks the layers of rights and reproduction information that exist for digital objects and their multiple versions, and the authenticity of version and provenance. Donor requirements, privacy concerns, and proprietary rights also can be captured through metadata.

Metadata organizes information. In the print and digital environments, metadata provides the efficient, accurate retrieval and precision of information to the user. Especially in the digital environment, where Web-based resources are growing exponentially, metadata provides the key to value-based searching already found in information centers such as libraries and archives.

Problems with polysemy (words with multiple meanings), ambiguity of meaning, and synonymy can all be alleviated by the proper application of metadata, either manually or through selected harvesting (metadata automatically generated or retrieved from a remote location by computer-generated programs). Well-formed metadata is the most efficient and effective tool for managing and finding objects in the complex information spaces that users now encounter on the Internet.

Why was metadata developed?

Traditionally, cataloging or metadata originates from three major information organizations: libraries, archives, and museums. In the past, each of these organizations developed its own standards. Those standards were rarely used by the other organizations. Each organization decided its needs and users were unique, so each rarely shared or attempted to cooperate. This attitude not only prevented joint endeavors early on to create metadata standards for the digital environment, but it also has discouraged outsiders with metadata needs to approach and work on standards with these communities.

Efforts to develop standards, or at least try to synthesize differences among the various standards, have occurred, but these efforts have been the exceptions rather than the norm.

Libraries were the first to develop consistent standards for the exchange of catalog records between computer systems. Led by the development of standards for information interchange in the early 1970s, such as ISO 2709 and the MARC formats, as well as agreement on rules regarding content and format such as *Anglo-American Cataloguing Rules* (*AACR*) and the International Standard Bibliographic Description (ISBD) series, libraries quickly became the de facto source of cataloging and metadata for much of the world's information.

Archives and museums were slow to develop metadata standards before the popularity of the Internet but have since taken the lead in encouraging the cooperation and interoperability initiatives that were disregarded in the past. The probable cause of this development is that information born digital and digital content creation is taking place on a much broader and more intense level among these information organizations than in libraries. Also, the issues of archiving and digital preservation of information are closer to the missions of these organizations than libraries, where other concerns usually take priority.

Finally, outside these information organizations, great concern arose among professional scholarly and scientific associations about the problems of addressing unique research and data. These other groups began to develop metadata standards outside, without reference to, and in replacement of the metadata practices of traditional information organizations.

Some of the current chaos stems from the attitude of institutions that were uninterested in adapting their standards or assisting others in standard development. The major reason for the existence of many standards, however, was that many of these associations felt their unique information wouldn't fit into the existing standards—to satisfy their objectives and information concerns, these associations felt they needed to develop proprietary standards.

When the Internet emerged, many metadata standards were already developed or in process—most existing in isolation from one another. Nobody felt the need to cooperate or talk about interoperability of standards, since their purposes were seemingly unrelated.

Other associations working with specific formats, such as audio-visual, sound, and film, as well as specific information communities such as government, geospatial, rights management, and education, worked quickly to develop their own metadata standards, too.

In addition, the failure of Internet search engines to adequately and sufficiently find reliable information through the harvesting of HTML by spiders and bots underlined the need for mechanisms to search metadata and cataloging data for digital information.

The metadata landscape is changing dramatically. People have begun to realize that avoiding duplication of effort is important. Cooperation among various information communities has become increasingly apparent, as the focus has shifted to combining and mapping metadata standards.

In addition, interoperability has become the key shared focus within the metadata community if multiple metadata standards are to survive. These

Mapping metadata standards involves comparison and matching of similar or related tags, and then construction of computer programs that will translate that information, so that one or more standards can interoperate, or speak, with one another. An example is "creator" in Dublin Core being mapped to "author" in MARC.

standards must be able to share, communicate, and transfer information among themselves.

A dramatic shift has occurred from standards and schemas design to more fundamental modeling approaches. This shift looks at the entire concept of information organization and description and explores new ways of presenting and finding that information.

Functional Requirements for Bibliographic Records (FRBR), www.ifla.org/VII/s13/frbr/frbr.htm

ABC Ontology model, www.openarchives.org/documents/oai.pdf

The two models under serious consideration are the International Federation of Library Associations (IFLA) Functional Requirements for Bibliographic Records (FRBR) document and the ABC Ontology model. FRBR is seriously being discussed within the professional cataloging community as a new bibliographic model for information organization and description for the future, and its integration with *AACR* is a distinct possibility. The Open Archives Initiative (OAI), which builds on the ABC Ontology model, is discussed in Chapter 3.

Different types and attributes of metadata

Four main types of metadata must be included in a metadata standard. These types have specific functions, addressing what the standard must describe:

- **Descriptive metadata** refers to the attributes of the object being described. Elements such as title, author, abstract, and keywords are examples. This type of metadata is used to describe or identify a resource for purposes such as discovery and identification. *Cataloging* is descriptive metadata.
- **Structural metadata** refers to the structure and relationships of a set of digital objects. This type of metadata is important because the structure of an information object, whether digital or not, is an important indicator of that object's meaning. How pages are ordered into a chapter or how a journal article is related to a journal, for example, are key indicators that need to be recorded by structural metadata.
- **Administrative metadata** is used to manage and administer information resources. File type, rights management, technical information, and how the resource was created are typical elements included within administrative metadata. Other elements include scanner parameters, initial capture settings, compression information, date of capture, and authentication. Minor metadata types included under administrative metadata include technical metadata and use metadata. Good administrative metadata is vital to the provision of services and systems to library users, and the long-term preservation of digital objects.
- **Preservation metadata** refers to elements related to the preservation management of an information resource. Documentation on the physical condition of the original resource, as well as information regarding efforts taken to preserve digital and physical versions of a resource, is recorded here. Some new metadata standards are focused entirely on preservation and related issues, and encompass all the above types of metadata within their schema. Chapter 3 addresses some of these initiatives.

Technical metadata is related to how a system functions or behaves; it has to do with hardware, software, and so on.

Use metadata is related to the use of information resources, such as user statistics.

Metadata creation and management have become complex mixtures of automatic and manual layers and processes that involve different people and functions at various points in the life cycle of an information object. Knowledgeable planning and organization regarding metadata types and attributes at the beginning of the process are vital to the efficient maintenance and long-term preservation of both digital and nondigital objects.

Metadata principles and practicalities

Refer to two major documents when addressing metadata principles and practicalities. One is an article by Erik Duval et al. in the April 2002 volume of *D-Lib Magazine* titled "Metadata Principles and Practicalities." The other is titled "A Framework of Guidance for Building Good Digital Collections," found at the Institute of Museum and Library Services (IMLS).

"Metadata Principles and Practicalities,"
www.dlib.org/dlib/april02/weibel/04weibel.html

"A Framework of Guidance for Building Good Digital Collections,"
www.imls.gov/pubs/forumframework.htm

The Duval article provides some excellent principles and practicalities related to metadata. In the context of metadata, principles are those concepts judged to be common to all domains of metadata and that might inform the design of any metadata schema or application. Practicalities are the rules of thumb, constraints, and infrastructure tenets that emerge from creating useful and sustainable systems.

Under principles, the authors list four truths that provide a guiding framework for the interoperability of metadata standards. These are:

- Modularity—elements from different standards that can be applied interoperably, the way LEGO™ blocks fit together
- Extensibility—an architecture that allows for a base schema that can be expanded to fit the needs of both designers and users
- Refinement—designers can choose a level of detail appropriate to a given application (the addition of qualifiers is an example)
- Multilingualism—respects cultural and linguistic diversity in construction and application of the standard

Under practicalities, the authors list nine minimum aspects related to metadata creation and management on the Internet:

- Application profiles—an assemblage of metadata elements from one or more metadata standards that are combined into a compound standard for local usage
- Syntax and semantics—an understanding about the meaning and form of metadata standards in order to share information
- Association models—these models include embedded metadata that resides in the markup of the resource, associated metadata that is maintained in separate yet available files, and third-party metadata that is maintained separately from the resource by a third party
- Identifying and naming metadata elements—coming to agreement on how to make metadata more understandable to everyone
- Metadata registries—actively maintained electronic dictionaries and databases that contain current metadata standards information, application profiles, and important projects
- Completeness of description—whether every element in a metadata standard is necessary; the debate of detailed description versus simple description of resources

- Mandatory versus optional elements—to be flexible, metadata standards must allow designers and users to further specify standards of practice in the application of metadata.
- Subjective and objective metadata—metadata can be both objective (date of creation, version, and so on) and subjective (assignment of keywords, reviews of the resource, and so on). Being able to separate these types of metadata to make the context as explicit as possible is critical.
- Automated generation of metadata—metadata created through computer-generated programs or on-the-fly through application-supported metadata, inference-based metadata, or creator-supplied metadata to reduce the cost and increase the quality of metadata descriptions

This document assists metadata standards organizations and designers in moving toward interoperability and internationalization. It also helps users locally apply and understand metadata schemas. As a conclusion, the authors state that metadata is "a key part of the information infrastructure necessary to help create order in the chaos of the Web, infusing description, classification, and organization to help create more useful stores of information."

The IMLS report was prepared by members of the Digital Library Forum to discuss issues related to the implementation and management of networked digital libraries. The group came up with four entities as indicators of good digital collections. They are:

- Collections
- Objects
- Metadata
- Projects

The IMLS authors also elaborate six principles related to good metadata:

- Good metadata should be appropriate to the materials in the collection, users of the collection, and intended, current, and likely use of the digital object (a listing of major metadata schemes follows this principle).
- Good metadata supports interoperability.
- Good metadata uses standard controlled vocabularies to reflect the what, where, when, and who of the content.
- Good metadata includes a clear statement on the conditions and terms of use for the digital object.
- Good metadata records are objects themselves and therefore should have the qualities of good objects, including archivability, persistence, unique identification, and so on.
- Good metadata should be authoritative and verifiable.
- Good metadata supports the long-term management of objects in collections.

This IMLS report includes examples of best practices in each of the principles listed above. Use the report when choosing and implementing metadata standards related to digitization and digital projects. The report's framework is geared for two audiences: for funding organizations and agencies wanting to encourage the creation of good digital collections, and for people who are working on projects and want to develop good digital collections.

Conclusion

Valid, reliable, harvest-read, and authoritative metadata is needed to address the rapidly expanding digital environment. In the mid-1990s, Internet search engines incorporated keyword searching as the primary means of searching the Web. This approach quickly proved to be flawed, indicating that both computer science and Web technology needed to incorporate the principles and practices of information science to provide precision and recall to search engine mechanisms.

Harvest-read metadata is metadata that is processed and extracted by a computer application.

The creation of metadata during the production phase of digital objects helps provide reliability and authentication. The internetworked environment raises other issues:

- The practicality of using spiders and bots to harvest and mine information for search engine indexes
- The validity and authentication of the metadata itself in HTML and XML structures
- The problems of harvesting and indexing information contained in certain software and HTML structures, such as portable document files (PDF) and Web pages incorporating HTML frame coding

Despite these concerns, the push to define and establish metadata standards among different information communities is moving forward quickly. Although some people still attempt to combine and narrow the number of metadata standards, the focus is now the interoperability and long-term preservation of digital information contained within metadata schemas. Metadata creators and organizers should choose early in the process of digitization which metadata standard(s) meets the needs of their particular information environment.