Introduction to Metadata *Pathways to Digital Information*

http://www.getty.edu/research/institute/standards/intrometadata/

Setting the Stage

Anne J. Gilliland-Swetland

Defining Metadata

Metadata, literally "data about data," is an increasingly ubiquitous term that is understood in different ways by the diverse professional communities that design, create, describe, preserve, and use information systems and resources. As these communities — and also repositories and information and communication technologies — come together to make the information age a reality, it is essential that we understand the critical roles that different types of metadata can play in the development of effective, authoritative, interoperable, scaleable, and preservable cultural heritage information and recordkeeping systems. Until the mid-1990s, "*metadata*" was a term most prevalently used by communities involved with the management and interoperability of geospatial data, and with data management and systems design and maintenance in general. For these communities, "metadata" referred to a suite of industry or disciplinary standards as well as additional internal and external documentation and other data necessary for the identification, representation, interoperability, technical management, performance, and use of data contained in an information system.

Perhaps a more useful "big picture" way of thinking about metadata is as "the sum total of what one can say about any information object at any level of aggregation." In this context, an *information object*[1] is anything that can be addressed and manipulated by a human or a system as a discrete entity. The object may be comprised of a single item, or it may be an aggregate of many items. In general all information objects, regardless of the physical or intellectual form they take, have three features — content, context, and structure — all of which can be reflected through metadata.

- *Content* relates to what the object contains or is about, and is *intrinsic* to an information object.
- *Context* indicates the who, what, why, where, how aspects associated with the object's creation and is *extrinsic* to an information object.
- *Structure* relates to the formal set of associations within or among individual information objects and can be *intrinsic* or *extrinsic*.

Cultural heritage and information professionals such as museum registrars, library catalogers, and archival processors are increasingly applying the term *metadata* to the value-added information that they create to arrange, describe, track and otherwise enhance access to information objects. Library metadata development has been first and foremost about providing intellectual and physical access to content. *Library metadata* includes indexes, abstracts, and catalog records created according to cataloging rules and structural and content standards such as MARC (MAchine-Readable Cataloging format), as well as authority forms such as LCSH (Library of Congress Subject Headings) or the AAT (Art & Architecture Thesaurus). Such bibliographic metadata has been cooperatively created since the 1960s and made available to repositories and users through automated systems such as bibliographic utilities, online public access catalogs (OPACs), and commercial online databases.

A large component of archival and museum activities has traditionally been focussed on context. Elucidating and preserving context is what assists with identifying and preserving the evidential value of records and artifacts in and over time; it is what facilitates the authentication of those objects, and it is what assists researchers with their analysis and interpretation. *Archival and manuscript metadata* includes accession records, finding aids, and catalog records. Archival descriptive standards that have been developed in the past two decades include the MARC Archival and Manuscript Control (AMC) format published by the Library of Congress in 1984 (now integrated into the MARC format for bibliographic description); the General International Standard Archival Description (ISAD (G)) published by the International Council on Archives in 1994; and Encoded Archival Description EAD), adopted as a standard by the Society of American Archivists (SAA) in 1999. While archival metadata has primarily existed in print form until recently, it is increasingly distributed on line through resources such as the RLIN AMC (the Research Libraries Information Network Archival and Mixed Collections file,[2] Archives USA,[3] and EAD-based archival information systems such as the Online Archive of California.[4] Consensus and collaboration have been slower to build in the museum community, where the benefits of standardization of description such as shared cataloging and exchange of descriptive data are less readily apparent.

An emphasis on the structure of information objects in metadata development by these communities has perhaps been less overt. However, structure has always been important in information organization and representation, even prior to computerization. Documentary and publication forms have evolved into industry standards and societal norms and have become an almost transparent information management tool — when users access a birth certificate, they can predict what the content is likely to be. When they use a scholarly monograph, they intuitively understand that it will be organized with a table of contents, chapter headings, and an index. Archivists use the physical structure of their finding aids to provide visual cues to researchers about the structural relationships between different parts of a record series or manuscript collection. Archival description also very much exploits the hierarchical arrangement of records according to the bureaucratic hierarchies and business practices of the creators of those records.

The role of structure has been growing as computer processing capabilities become increasingly powerful and sophisticated. Information communities are aware that the more highly structured an information object is, the more that structure can be exploited for searching, manipulation, and interrelating with other information objects. Capturing, documenting, and enforcing that structure, however, require specific types of metadata.

In short, in an environment where a user can gain unmediated access to information objects over a network, metadata:

- certifies the authenticity and degree of completeness of the content;
- establishes and documents the context of the content;
- identifies and exploits the structural relationships that exist between and within information objects;
- provides a range of intellectual access points for an increasingly diverse range of users; and
- provides some of the information an information professional might have provided in a physical reference or research setting.

There is more to metadata than description, however. A more inclusive conceptualization of metadata is needed as information professionals consider the range of their activities that may end up being incorporated into digital information systems. Repositories also create metadata relating to the administration, accessioning, preservation, and use of collections. Acquisition records, exhibition catalogs, and use data are all examples of these, even though they are still largely created in paper form. Today, integrated information systems such as virtual museums, digital libraries, and archival information systems include digital versions of actual collection content as well as descriptions of that content. Incorporating other types of metadata into such systems reaffirms their importance in administering collections and maintaining their intellectual integrity both in and over time. Paul Conway alludes to this metadata capability when he discusses the impact of digitization on preservation:

> *The digital world transforms traditional preservation concepts from protecting the physical integrity of the object to specifying the creation and maintenance of the object whose intellectual integrity is its primary characteristic.*[5]

When applied outside the repository, the term *metadata* acquires an even broader scope. An Internet resource provider might use *metadata* to refer to information being encoded into HTML metatags for the purposes of making a Web site easier to find. Individuals digitizing images might think of metadata as the information they enter into the header field for the digital file to record information about the image, the imaging process, and image rights. A social science data archivist might use the term to refer to the systems and research documentation necessary to run and interpret a magnetic tape containing raw research data. An electronic records archivist might use the term to refer to all the contextual, processing, and use information needed to identify and document the scope, authenticity, and integrity of an active or archival record in an electronic recordkeeping system. Metadata is critical in personal

information management and for ensuring effective information retrieval and accountability in recordkeeping — something that is becoming increasingly important with the rise of electronic commerce and digital government. In all of these diverse interpretations, metadata not only identifies and describes an information object; it also documents how that object behaves, its function and use, its relationship to other information objects, and how it should be managed.

As this discussion implies, theory and practices vary considerably owing to the differing professional and cultural missions of museums, archives, libraries, and other information and recordkeeping communities.[6] Many additional highly detailed metadata standards are now emerging (such as EAD and the Australian Recordkeeping Metadata Schema (RKMS)) that attempt to articulate these communities' mission-specific differences as well as to facilitate mapping between common data elements. By contrast, the Dublin Core Metadata Element Set (DC) identifies a small, simple set of metadata elements that can be used by any community to describe and search across a wide variety of information resources on the World Wide Web. Such metadata standards are necessary in order to ensure that different kinds of descriptive metadata are able to interoperate with each other and with metadata from non-bibliographic systems of the kind that the data management communities and information creators are generating. Indeed, not since the latter part of the nineteenth century has there been such an exciting — but also potentially bewildering — array of organizational and descriptive schemas from which information professionals can choose.

Categorizing Metadata

All of these perspectives on metadata become important in the development of networked digital information systems, but they lead to a very broad conception of metadata. To understand this conception better, it is helpful to break it down into distinct categories — administrative, descriptive, preservation, use, and technical metadata — that reflect key aspects of metadata functionality. Table 1 defines each of these metadata categories and gives examples of common functions that each might perform in a digital information system.

Table 1. Different Types of Metadata and Their Functions

Type	Definition	Examples
Administrative [7]	Metadata used in managing and administering information resources	- Acquisition information - Rights and reproduction tracking - Documentation of legal access requirements - Location information - Selection criteria for digitization - Version control and differentiation between similar information objects - Audit trails created by recordkeeping systems
Descriptive	Metadata used to describe or identify information resources	- Cataloging records - Finding aids - Specialized indexes - Hyperlinked relationships between resources - Annotations by users - Metadata for recordkeeping systems generated by records creators
Preservation	Metadata related to the preservation management of information resources	- Documentation of physical condition of resources - Documentation of actions taken to preserve physical and digital versions of resources, e.g., data refreshing and migration
Technical	Metadata related to how a system functions or metadata behave	- Hardware and software documentation - Digitization information, e.g., formats, compression ratios, scaling routines - Tracking of system response times - Authentication and security data, e.g., encryption keys, passwords
Use	Metadata related to the level and type of use of information resources	- Exhibit records - Use and user tracking - Content re-use and multi-versioning information

In addition to there being different types of metadata and metadata functions, metadata also exhibits many different characteristics. Table 2 indicates some of the key attributes of metadata, with examples.

Table 2. Attributes and Characteristics of Metadata

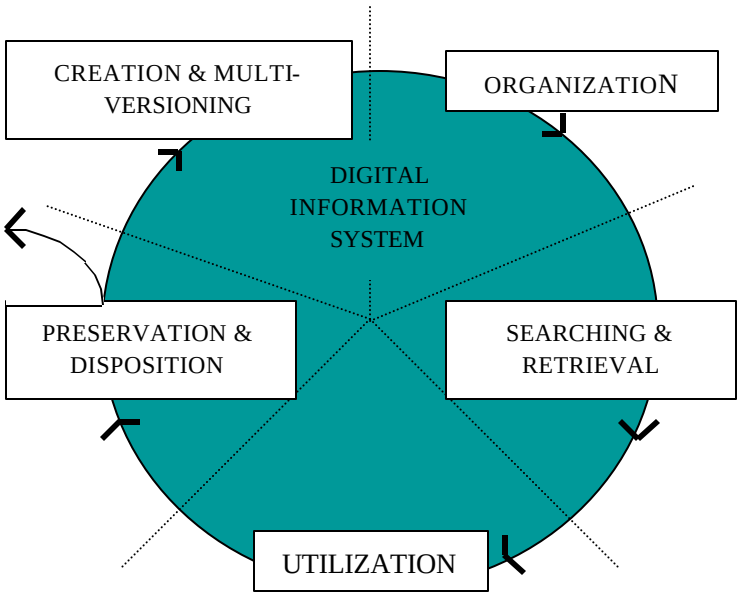
Attribute	Characteristics	Examples
Source of metadata	Internal metadata generated by the creating agent for an information object at the time when it is first created or digitized	- File names and header information - Directory structures - File format and compression scheme
	External metadata relating to an information object that is created later, often by someone other than the original creator	- Registrarial and cataloging records - Rights and other legal information
Method of metadata creation	Automatic metadata generated by a computer	- Keyword indexes - User transaction logs
	Manual metadata created by humans	- Descriptive surrogates such as catalog records and Dublin Core metadata
Nature of metadata	Lay metadata created by persons who are neither subject nor information specialists, often the original creator of the information object	- Metatags created for a personal Web page - Personal filing systems
	Expert metadata created by either subject or information specialists, often not the original creator of the information object	- Specialized subject headings - MARC records - Archival finding aids
Status	Static metadata that never change once they have been created	- Title, provenance, and date of creation of an information resource
	Dynamic metadata that may change with use or manipulation of an information object	- Directory structure - User transaction logs - Image resolution
	Long-term metadata necessary to ensure that the information object continues to be accessible and usable	- Technical format and processing information - Rights information - Preservation management documentation
	Short-term metadata, mainly of a transactional nature	
Structure	Structured metadata that conform to a predictable standardized or unstandardized structure	- MARC - TEI and EAD - local database formats
	Unstructured metadata that do not conform to a predictable structure	- Unstructured note fields and annotations

Semantics	Controlled metadata that conform to a standardized vocabulary or authority form	- AAT - ULAN - AACR2
	Uncontrolled metadata that do not conform to any standardized vocabulary or authority form	- Free-text notes - HTML metatags
Level	Collection metadata relating to collections of information objects	- Collection-level record, e.g., MARC record or finding aid - Specialized index
	Item metadata relating to individual information objects, often contained within collections	- Transcribed image captions and dates - Format information

Metadata creation and management have become a very complex mix of manual and automatic processes and layers created by many different functions and individuals at different points in the life of an information object. Figure 1 illustrates the different phases through which information objects typically move during their life in a digital environment.[8] As they move through each phase, the objects acquire layers of metadata that can be associated with the objects in several ways. This metadata can be contained within the same envelope as the information object — for example, in the form of header information for an image file, e.g. Dublin Core, or through some form of bundling, for example, with the Universal Preservation Format (UPF). Metadata can also be attached to the information object through bi-directional pointers or hyperlinks, while the relationships between metadata and information objects, and between different aspects of metadata, can be documented by registering them with a metadata registry.[9] However, in any instance where it is critical that metadata and content coexist, then it is recommended that the metadata become an integral part of the information object and not be stored elsewhere.

As systems designers increasingly respond to the need to incorporate and manage metadata in information systems and to address how to move them forward through time, many additional mechanisms for associating metadata with information objects are likely to become available. Metadata registries and schema recordkeeping systems are also more likely to develop as it becomes increasingly necessary to document schema evolution and to alert implementors to version changes.

Figure 1. The Life Cycle of Objects Contained in a Digital Information System



Creation and multi-versioning: Objects enter a digital information system by being created digitally or by being converted into digital format. Multiple versions of the same object may be created for preservation, research, dissemination, or even product development purposes. Some administrative and descriptive metadata may be included by the creator.

Organization: Objects are automatically or manually organized into the structure of the digital information system and additional metadata for those objects may be created through registration, cataloging, and indexing processes.

Searching and retrieval: Stored and distributed objects are subject to search and retrieval by users. The computer system creates metadata that track retrieval algorithms, user transactions, and system effectiveness in storage and retrieval.

Utilization: Retrieved objects are utilized, reproduced, and modified. Metadata related to user annotations, rights tracking, and version control may be created.

Preservation and disposition: Information objects undergo processes such as refreshing, migration, and integrity checking to ensure their continued availability. Information objects that are inactive or no longer necessary may be discarded. Metadata may document both preservation and disposition activities.

Other Little-Known Facts about Metadata

1. *Metadata does not have to be digital.* Cultural heritage and information professionals have been creating metadata for as long as they have been managing collections. Increasingly, such metadata are being incorporated into digital information systems.
2. *Metadata relates to more than the description of an object.* While museum, archives, and library professionals may be most familiar with the term in association with description or cataloging, metadata can also indicate the context, management, processing, preservation and use of the resources being described.
3. *Metadata can come from a variety of sources*. It can be supplied by a human (a creator, information professional, or user), created automatically by a computer, or inferred through a relationship to another resource such as a hyperlink.
4. *Metadata continuse to accrue during the life of an information object or system.* Metadata is created, modified, and sometimes even disposed of at many points during the life of a resource.
5. *One information object's metadata can simultaneously be another information object's data*.

Why is metadata important?

As illustrated by the preceding discussion, metadata consists of complex constructs that can be expensive to create and maintain. How then can one justify the costs and efforts involved? The development of the World Wide Web and other networked digital information systems has provided information professionals with many opportunities, while at the same time requiring them to confront issues that they have not had occasion to explore previously. Judiciously crafted metadata element sets, wherever possible conforming to national and international standards, have become the tools that information professionals are using to exploit some of these opportunities, as well as to address some of the new issues:

Increased accessibility: Effectiveness of searching can be significantly enhanced through the existence of rich, consistent metadata. Metadata can also make it possible to search across multiple collections or to create virtual collections from materials that are distributed across several repositories, but only if the descriptive metadata are the same or can be mapped across each site. Digital information systems and emerging metadata standards developed by different professional communities but incorporating some common data elements, such as Encoded Archival Description (EAD), the Text Encoding Initiative (TEI), and the Dublin Core are making it easier for users to negotiate between descriptive surrogates of information objects and digital versions of the objects themselves, and to search at both the item and collection level within and across information systems.[10]

Retention of context: Museum, archival, and library repositories do not simply hold objects. They maintain collections of objects that have complex interrelationships among each other and associations with people, places, movements, and events. In the digital world it is not difficult

for a single object from a collection to be digitized and then to become separated from both its own cataloging information and its relationship to the other objects in the same collection. Metadata plays a critical role in documenting and maintaining those relationships, as well as in indicating the authenticity, structural and procedural integrity, and degree of completeness of information objects. In an archive, for example, by documenting the content, context, and structure of an archival record, metadata in the form of an archival finding aid is what helps to distinguish that record from decontextualized information.

Expanding use: Digital information systems for museum and archival collections make it easier to disseminate digital versions of unique objects to users around the globe who, for reasons of geography, economics, or other barriers, might otherwise never have had an opportunity to view them. With new communities of users, however, come new challenges concerning how to make the materials most intellectually accessible to them. These new communities of users may have significantly different needs to those of the traditional users for whom many existing information services have been designed. For example, teachers and school children may want to search for and use information objects in quite different ways than scholarly researchers do. Metadata can document changing uses of systems and content, and that information can in turn feed back into systems development decisions. Well-structured metadata can also facilitate an almost infinite number of ways to search for information, present results, and even manipulate information objects without compromising the integrity of those information objects.

Multi-versioning: The existence of information and cultural objects in digital form has heightened interest in the ability to create multiple and variant versions of those objects. This process may be as simple as creating both a high-resolution copy for preservation or scholarly research purposes and a low-resolution thumbnail image that can be rapidly transferred over a network for quick reference purposes. Or it may involve creating variant or derivative forms to be used, for example, in publications, exhibitions, or schoolrooms. In either case, there must be metadata to link the multiple versions and capture what is the same and what is different about each version. The metadata must also be able to distinguish what is qualitatively different between variant digitized versions and the hard copy original or parent object.

Legal issues: Metadata allows repositories to track the many layers of rights and reproduction information that exist for information objects and their multiple versions. Metadata also documents other legal or donor requirements that have been imposed on objects — for example, privacy concerns or proprietary interests.

Preservation: If digital information objects that are currently being created are to have a chance of surviving migrations through successive generations of computer hardware and software, or removal to entirely new delivery systems, they will need to have metadata that enables them to exist independently of the system that is currently being used to store and retrieve them. Technical, descriptive, and preservation metadata that documents how a digital information object was created and maintained, how it behaves, and how it relates to other information objects will all be essential. It should be noted that for the information objects to

remain accessible and intelligible over time, it will also be essential to preserve and migrate this metadata.

System improvement and economics: Benchmark technical data, much of which can be collected automatically by a computer, is necessary to evaluate and refine systems in order to make them more effective and efficient from a technical and economic standpoint. The data can also be used in planning for new systems.

Conclusion and Outstanding Questions

Metadata is like interest — it accrues over time. To stretch the metaphor further, wise investments generate the best return on intellectual capital. Carefully designed metadata results in the best information management in the short and long-term. If thorough, consistent metadata has been created, it is possible to conceive of it being used in an almost infinite number of new ways to meet the needs of non-traditional users, for multi-versioning, and for data mining. But the resources and intellectual and technical design issues involved in metadata development and management are far from trivial. For example, some key questions that must be resolved by information professionals as they develop digital information systems and objects include:

- identifying which metadata schema or schemas should be applied in order to best meet the needs of the information creator, repository and users;
- deciding which aspects of metadata are essential for what they wish to achieve, and how granular they need each type of metadata to be — in other words, how much is enough and how much is too much. There will likely always be important tradeoffs between the costs of developing and managing metadata to meet current needs, and creating sufficient metadata that can be capitalized upon for future, often unanticipated uses;
- ensuring that the metadata schemas being applied are the most current versions.

What we do know is that the existence of many types of metadata will prove critical to the continued physical and intellectual accessibility and utility of digital information resources and the information objects that they contain. In this sense, metadata provides us with the Rosetta Stone that will make it possible to decode information objects and their transformation into knowledge in the cultural heritage information systems of the twenty-first century.

---

[1] An information object is a digital item or group of items, regardless of type or format, that can be addressed or manipulated as a single object by a computer. This concept can be confusing in that it can be used to refer both to actual content (such as digitized images) and to content surrogates (such as catalog records or finding aids).

[2] http://lcweb.loc.gov/z3950/rlinamc.html

[3] http://archives.chadwyck.com/

[4] http://www.oac.cdlib.org/

[5] Conway, Paul. *Preservation in the Digital World*. Washington, DC: Commission on Preservation and Access, 1996. http://www.clir.org/pubs/reports/conway2/index.html

[6] For a more detailed discussion, see Anne J. Gilliland-Swetland, *Enduring Paradigms, New Opportunities: The Value of the Archival Perspective in the Digital Environment* (Washington, D.C.: Council on Library and Information Resources, 2000).

[7] For an example of a detailed specification of administrative metadata, see the *Making of America II White Paper, Part III, Structural and Administrative Metadata* http://sunsite.berkeley.edu/MOA2

[8] Modified from the Information Life Cycle, *Social Aspects of Digital Libraries: A Report of the UCLA-NSF Social Aspects of Digital Libraries Workshop*, Los Angeles, CA: Graduate School of Education & Information Studies, November 1996: 7.

[9] Joint Workshop on Metadata Registries, Workshop Report, Draft 1.6, 1997. http://www.lbl.gov/~olken/EPA/Workshop/report.html.

[10] Gilliland-Swetland, Anne J. "Popularizing the Finding Aid: Exploiting EAD to Enhance Online Browsing and Retrieval in Archival Information Systems by Diverse User Groups" *Journal of Internet Cataloging* 4 nos. 1/2 (2000) (forthcoming).