

## METADATA

Francis Miksa © 2000

### **Definition.**

Metadata as a term has come to be widely used only during the past half dozen years. As already noted, metadata means data about data. But, there is just enough vagueness in that phrase that it will be useful to point out how its meaning has developed.

The term metadata had its origin during the 1970s in database construction where it came to be used as a way to differentiate between two kinds of data found in the database—data that referred directly to the objects that the database was about, and data that represented the category or field (or subfield) names into which the first kind of data was partitioned. . For example, in a database of information about students in a university, one would find the actual names of students divided, perhaps, into last name, first name, and middle initial(s). But, then one would also find field names for those same categories, often in shortened form for programming purposes—for example, Lastn, Firstn, and Midin. (An alternative would be simply to number the fields (e.g., F02,F03, and F04) and then keep a control list or ‘dictionary’ of their meanings—i.e., F02 = Last name; F03 = First name; F04 = Middle initial(s).

Within the foregoing context, the latter three shortened terms serve as data that is about the data recorded as the actual last names, first names, and middle initials of the students whose names are entered in the fields. In order to differentiate the category field names from the actual data being recorded about students, they were as a group named metadata, and were defined as data (in the form of category field names) about data (in the form of attributes of the objects that the database was about).

This division of data into two kinds or layers, data and metadata, both of which occurred within the database, worked well when the objects that the database included information about were not primarily information entities—in short, when they consisted of objects such as persons, products, processes, and so on. However, when the objects being listed in a database became information entities such as books, periodicals, sound recordings, etc. (a computerized catalog or index is, in fact, a “bibliographic database” or an “information entity database”) then a new situation arose, because information entities are themselves comprised chiefly of data. Technically, this produced no less than three layers of data—the data that exists in the information entities themselves, their texts, graphics, etc., the data that exists within the computer database and that consists of category names/fields, and the actual data within the computer database that refers to the information entities. The solution among information entity organizers as to how to distinguish different kinds of data has been to refer to all of the data within the database as metadata. Something of this complexity can be seen in the following table:

### Object Database

(For example, a database of information about people  
where the objects are not information entities)

<u>[Object]</u>	<u>Computerized Database</u>	
People _____	<u>Metadata</u>	<u>Data</u>
_____	(Field or subfield <u>category names/#s</u> )	(Actual data <u>values</u> )
Person #1	Lastn	Smith
	Firstn	David
	Midin	L.

### Information Entity Database

(For example, a computerized catalog or index)

<u>[Info. Entity]</u>	<u>Computerized Database</u>	
<u>Data</u>	<u>Metadata</u>	<u>Data</u>
(consisting of text, <u>graphic, etc. data</u> )	(Fields or subfield <u>category names/#s</u> )	(Actual data <u>values</u> )
Info. Entity #1	Author	Puzo, Mario
Info. Entity #1	Title	Omerta

In the first database above where the objects are people, both the data and the metadata that are about the data are within the database itself. In the second database, where the objects are information entities, data is outside the database in the form of the information entities themselves, and all data within the database are considered metadata.

In past decades, the same information now put into a computerized database and called metadata had its own names, names that were dependent upon the tradition of practice using it. For example, in library cataloging, what is now called metadata was then simply called cataloging information; in indexing, either indexing data or descriptors; in archival work, collection information, finding aid information, calendar information, etc.

### **The Language of Metadata.**

Metadata for information entities comes from the information entity itself, namely, from the attributes or characteristics of the information entities. But, as Svenonius points out, some of these attributes are found in or on the information entities, whereas others are devised and simply assigned to the entity as an attribute. For example, for many information entities what is considered a “title”—that is, a name of the entity—and what is considered the name of an “author”—that is the name of the creator of the intellectual or artistic content of the entity—will be found directly in or on the entity itself. In contrast, many others attributes are not recorded in or on the entity itself, or are recorded on some information entities but not on others. The latter include such things as the size of the information entity, or the fact that it has illustrations, or the fact that it is in a certain medium, or a full and sufficient characterization of its subject, or its subject converted to the form of a call number, etc. If these kinds of attributes are important, they must be devised and then assigned to the entity.

Normally, when an attribute is used in exactly the way it appears in the information entity itself, it is called *natural language* metadata. [This sense of natural language should be distinguished from the normal and natural language that people speak. The latter is also natural language, but it is spoken natural language, not the language specifically used in information entities to refer to themselves.] However, if the language of an attribute is either taken from the information entity but then manipulated into a normalized format or created de novo to represent the attribute, it is called *controlled vocabulary* metadata. In order to gain some symmetry in speaking about these two kinds of metadata and in line with her conclusion that all metadata is an expression of special languages, Svenonius calls natural language metadata by the name “uncontrolled vocabulary,” and she calls metadata that is manipulated and normalized by the name, “controlled vocabulary.”

Normalizing the vocabulary of attributes is typically achieved in the following ways:

1. Choosing to use one form of a term among two or more variant forms that are available, and referring from the non-used form to the used form.

Examples: Choosing to use Francis L. Miksa,  
and referring from Fran Miksa, F. Miksa, Francisco Miksa, and F. L. Miksa to  
the form Francis L. Miksa.

Choosing to use KILLER CELLS,  
and referring from K CELLS to KILLER CELLS.

Choosing to use POLYBROMINATED BIPHENYLS,  
and referring from POLYBROMOBIPHENALS to POLYBROMINATED BIPHENYLS

[1a. Some changes are relatively minor. For example, in the Anglo-American cataloging tradition of practice, title words are ordinarily converted to lower case orthography except for the first word and for proper nouns, as in,

Example: The closing of the American mind, *not* The Closing of the American Mind

And, under certain circumstances, an initial article will be omitted, as in,

Example: Seat of the soul, *not* The seat of the soul]

2. Choosing to use one form of a term among two or more different forms that are available and referring from the non-used form(s).

Examples: Choosing to use Mohammed Ali,  
and referring from Cassius Clay to Mohammed Ali.

Choosing to use TADARIDA AFRICANA,  
and referring from GIANT AFRICAN FREE-TAILED BAT to .TADARIDA AFRICANA.

[2a. As an alternative, using all variant forms of a term and referring among them.

Example: Plaidy, Jean.

For works of this author entered under other names, search also under Carr, Philippa; Ford, Elbur; Holt, Victoria; Kellow, Kathleen; Tate, Ellalice.]

3. Manipulating the syntax of a term to achieve a normalized order.

Examples: Using Miksa, Francis L., *not* Francis L. Miksa

Using Rad, Gerhard von, *not* von Rad, Gerhard

Using La Fontaine, Henri de, *not* De la Fontaine, Henri

Using MARKETING, VERTICAL, *not* VERTICAL MARKETING

4. Adding data to a term that clarify it or that distinguish it from another term written the same way (i.e., a homograph.) [Additions are shown here in parentheses]

Examples: Smith, David L. (1901-1975)

Smith, David L. (1938- )

Grace (Theology)

Grace (Aesthetics)

Nine Inch Nails (Musical group)

Apollo 12 (Spacecraft)

Apollo Quartet (Vocal group)

Apollo Smile (Musical group)

Washington (D.C.)

Washington (State)

Washington (Ind.)

Catalogs, Classified (Universal decimal)

Catalogs, Classified (Dewey decimal)

5. Creating complex structured titles or name and title combinations that collocate versions of works by one or another special attributes:

Examples: Actual title: L'Apocalypse de Saint John.

Specially formulated title: Bible. N. T. Revelation. French

Actual title: Trio, op. 66, Piano, Violin & Violoncello.

Specially formulated title: Trios, piano, strings, no. 2, op. 66, C minor

Actual series title:	Publications du Centre National D'histoire des Sciences
Specially formulated title:	Publications (Centre national d'histoire des sciences (Belgium))

The importance of distinguishing between natural language vocabulary metadata and controlled vocabulary metadata and understanding their relative merits cannot be overemphasized. Natural language metadata is easier to determine because it involves nothing more than copying it from information entities. In fact, when an information entity exists entirely in electronic form, programming can be created that identifies such data almost entirely in an automatic manner. Its weakness is that it contains all of the vagueness that virtually thousands of creators of information entities assign to it, and without at least some controls it will easily lead to loss of access or even erroneous access. For example, the natural language title of a work written by Barbara Tuchman is *A Distant Mirror*. Were one to accept the title words as being indicative of the subject of her book and use those words for its subject attribute, then, one supposes the entity would be grouped with other books on optics or mirrors. In reality, the book is about the Europe in the fourteenth century. In a similar manner, her book entitled *The Proud Tower* is not about towers, but rather about the beginnings of World War I.

In contrast, controlled vocabulary is much more difficult and expensive to create. It is labor intensive because all forms of a term must be checked out and, where necessary, linked in the system. And, in addition, one must keep a record of terms being officially used and terms that are only referred from. The latter is called "Authority work" at least in library cataloging, and a record of such decisions is called an authority file. At the same time, controlled vocabulary is a powerful tool because it has the capacity to collocate. In short, by normalizing vocabulary, a system will bring together the various information entities (or their surrogates) to which such normalized vocabulary has been assigned.

### **The Functions of Metadata.**

Metadata has three basic functions. The first, already obvious, is that it is used to represent information entities in information entity access systems. The second is to aid in the structuring of the system. And the third is to provide a basis for the visual display of the metadata.

**Metadata as a Means of Representation:** Given the function of metadata as a means of representation, metadata must be chosen and worked with carefully in order that the system not lose control of it and end up not remembering in a systematic way how its metadata has been employed. It is this realm that Svenonius most directly addresses in her two chapters, 4 and 5 on bibliographic languages and principles of description where she speaks of standards, rules, principles, and other matters directly related to the use of metadata. In short, metadata chosen to represent information entities must have been chosen according to established and rationalized purposes, standards and principles, with detailed guidelines at all stages of its creation.

**Metadata as an Aid to Structuring a System:** The second function of metadata is to aid in structuring the information entity access system. In this respect, the very kinds of metadata

chosen will often serve as a framework for how to organize a system. For example, large information entity databases such as those found in the Online Computer Library Center (OCLC) or in the Research Libraries Information Network (RLIN) are generally structured in the form of an inverted file system, where one file exists (a ‘master’ file) that includes the complete record of data of each information entity—i.e., all of its discursive and controlled vocabulary data—and other files, built by copying information out of that first or ‘master’ file, serve to “index” the first file. The latter are generally updated at night when the system is not in great demand. When one looks for information entities in the online catalog, one actually searches only the index files. If a search yields more than one hit, intermediate displays will usually provide an abbreviated list of the hits. When, finally, a particular record of a particular information entity is chosen, then the entire record from the first or master list is displayed.

Alphabetical card catalogs have a traditionally different approach to data structure. They often function as “term-on-item” systems in which every card in the system contains the complete record of a particular information entity, the only differences being that different copies of the same record are filed under different access points. This kind of file is usually kept in a single alphabetical order and does not have anything but alphabetically derived entries in it. In short, it does not usually include any long sequences of purely numeric kinds of access points or descriptors, such as ISBNs or Music numbers, etc.

A classified card catalog is also ordinarily a term-on-item system; but in contrast to an alphabetical catalogs, it is arranged numerically (rather than alphabetically) by the classification notations of the classification system used that have been assigned to individual information entities. A classified catalog also normally has assigned multiple classification numbers to an individual information entity, just as in an alphabetical catalog an individual information entity has been assigned different verbal subject descriptors, so that the entry or description will appear in different subject sections of the system. A classified card catalog, by being organized by classification notations, is in effect a subject catalog system. One cannot look directly in it for names or titles. Because of this, a classified catalog needs a second and separate file which is alphabetical, so as to list the same entries alphabetically under names and titles. And it will also require a third separate file—one that allows the user to search for subject terms alphabetically so as to find out what their notational equivalents are if that equivalent relationship is not known. Once the notational equivalent of a verbal term is found, one can then go to the main section of the classified catalog and conduct a search by classification number.

**Metadata as a Basis of Visual Display of Information:** The final function of the metadata is to provide a basis for the display of information about information entities for the user of the system to read. This should be so obvious as not to need mentioning. However, the reverse of it is worth note. If in representing information entities, one or another kind of attribute does not have metadata representations in the system, or if in collecting the metadata some sub-aspect of the metadata is not differentiated formally and systematically, then systematically displaying the data will be impossible (if the metadata has not been collected) or will be very difficult (if the metadata has not been differentiated). The latter case is particularly important. If, in collecting metadata for information entities, one does not differentiate, for example, subtitles from main titles, or main class number from book number, by any consistent punctuation or subfield indicator; then it will be nearly impossible to either label or display those sub-elements in some

special way. In short, if they are not specially marked, they will not be able to be manipulated for the purposes of display.

### **Categories of Metadata.**

Categories of metadata will generally be found to be tailored to practices in the various traditions of information entity access control. In short, each such tradition has its own master list of such categories, these being in turn based on how the tradition of practice has evolved and according to what kinds of information entity attributes have come to be viewed as valuable.

Library cataloging has generally used the following categories, although as systems have become computerized, some of the distinctions here have become broken down.

#### Content metadata:

*Names* associated with creation of the content:

*Persons* names

*Corporate body* names (including the names of Conferences, Meetings, etc.)

*Titles* of works, regardless if only one work is in the entity or if there is more than one work in the information entity.

*Subjects* of information entities or of specific works (including names and titles also as subjects, and extending to both verbal subject terms and classification numbers)

*Forms* of information entities and of works (incl. genre, medium, etc.)

*Audience* of information entities or of specific works (i.e., for what class of persons has an information entity content been created?)

*Relationships* to other works or information entities (including links, their titles, etc.)

For example, *Monographic series titles*, *Titles of original versions*, *Collective titles* of information entities that contain a work along with other works, etc.

*Alphanumeric and numeric identifiers* of information entities and of works

#### Container metadata:

*Collective titles*, (the name of an information entity, but not also the name of a work)

Publication data, including, variously, place published, manufactured, or acquired from, name of publisher, manufacturer, or generator, and date of publication, manufacture, or generation.

*Extent*, including number of physical pieces, length, running time, etc.

*Dimensions*

*Other physical details* according to the medium used or the special type of material being represented (i.e., for example, maps, printed music, etc.), sometimes written in controlled vocabulary and at other times written as discursive notes, including statements about bibliographical apparatus, history of the item, special features of the item, etc.

*Alphanumeric and Numeric identifiers* of containers

All of the foregoing may well be represented both in specially devised codes as well as in the form of verbal, alphanumeric, or numeric terms.

Archival and Museum practice generally also add metadata for the following attributes:

Provenance (i.e., source, or information on the origination of the materials)  
 Special conditions of use and care  
 Donors

The foregoing is incomplete by any measure and serves mainly to get one thinking about information entity attributes and their metadata equivalents. A more systematic approach to the matter would be to examine metadata formats to see what has developed by the way of standards in the various information entity access control traditions of practice. A useful discussion of metadata formats of all kinds will be found in Taylor, chapter 4, and a discussion of metadata formats especially relevant to library cataloging will be found in Taylor, chapter 5.

### **Metadata Formats.**

When representing information entities in a given information entity access control system, the metadata collected for each information entity will be compiled into what is called an information entity “record.” (In former days, such records were simply called entries. The term “record” itself comes from having moved to computerized databases for such information.) A record can be defined in turn as the complete set of metadata pertaining to a single information entity (no matter how the latter is defined) within a given information entity access control system.

The purpose of a metadata format is to provide an established order for the categories of metadata that accumulate for any one information entity. An established order is needed especially in a computerized system because the computer, being, so to speak, as blind as a bat and as dumb as a post, cannot “think”: about the metadata that is in a record nor can it identify this or that category of metadata by thinking about it. A computer does not think. It simply processes instructions given to it. Thus, in order to perform operations on any collection of such records, one must organize the metadata in a normative order with appropriate signals (codes of one sort or another) that indicate the beginnings of kinds of metadata, the ends of kinds of metadata, the extent of kinds of metadata, relationships between kinds of metadata, sources of metadata, characterizations of metadata, special instructions for manipulating metadata, etc.

In addition to the foregoing purposes, metadata formats also have particular functions. First, they provide a basis for communicating the metadata from one electronic system to another without human intervention except to write the programming instructions to accept or send such communications. Second, metadata formats, along with the metadata categories themselves, also provide a basis for visually displaying the metadata for users to read. Third, metadata formats, along with the metadata categories themselves, provide aid in structuring the information entity access system itself.