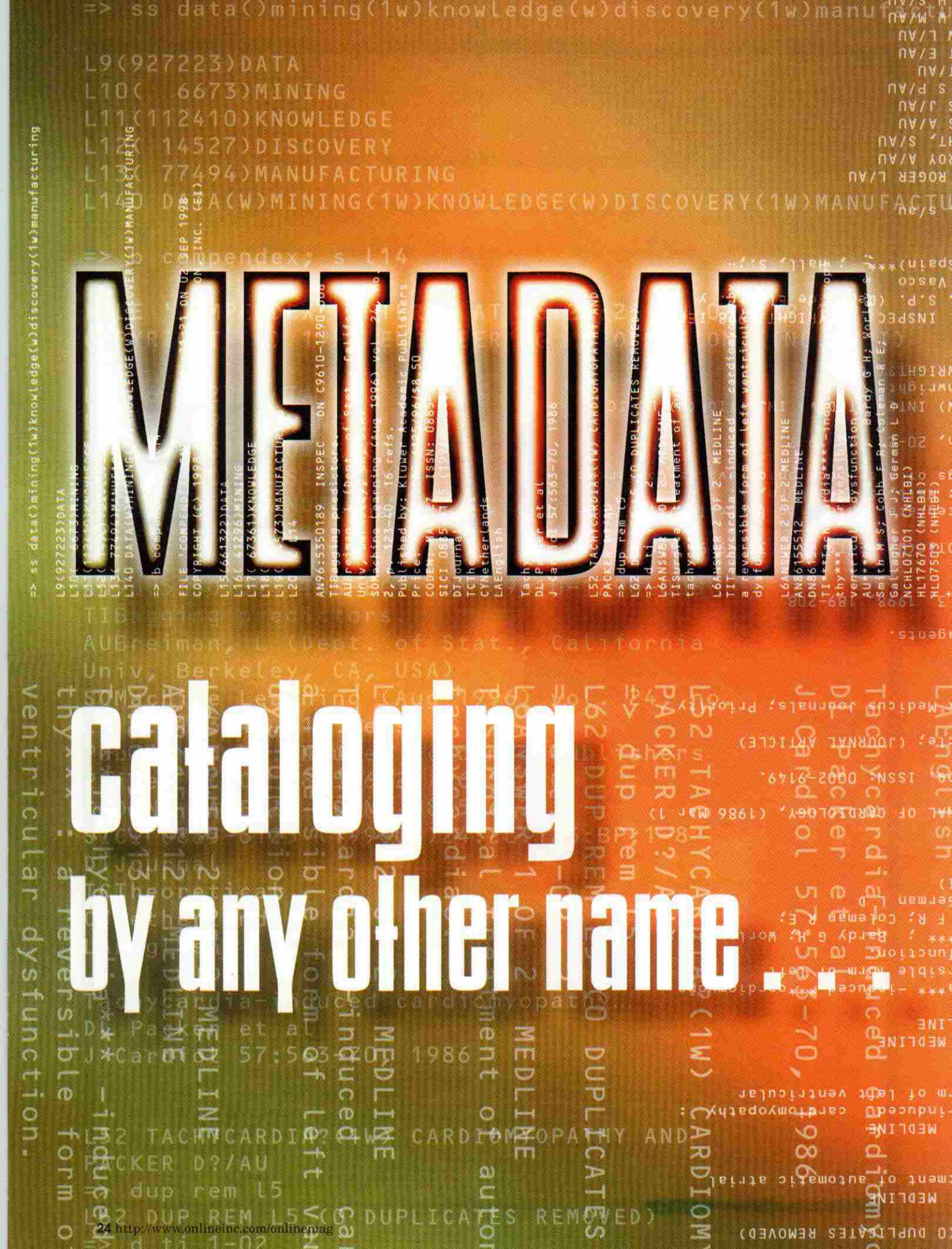
METADATA
Cataloging by any other name...

by Jessica Milstead and Susan Feldman

Editor's Note: Be sure to take a look at this article's companion piece, also by Jessica and Susan, entitled "Metadata Projects and Standards," for an overview of the variety of emerging and often conflicting projects for standardizing electronic resources. These articles are on ONLINE*'s Web site (http://www.onlineinc.com/onlinemag/metadata) and the extensive list of metadata project and resource links contained in the sidebar on page 38 is clickable—the links are also embedded in the text of the Web articles for easy reference. Also, for a look into how an organization actually applies a metadata scheme to its own corporate intranet, read Kelly Doran's piece, "Metadata for a Corporate Intranet," in this issue beginning on page 42.*

...librarians and indexers have been producing and standardizing metadata for centuries.

Whether you call it cataloging, indexing, or metadata, the concept is a familiar one for information professionals. Now the electronic world has finally discovered it. Until a few years ago, only a few philosophers had ever heard of the word "metadata." Today, it is hard to find a publication about electronic resources that ignores it.

While metadata has become a buzzword in the information business, the concept is important for both authors and seekers of electronic information. Used effectively, it makes information accessible by labeling its contents consistently. Metadata leaves a pathway for users to follow to find the information they need—all in one place. In invisible cyberspace, this is even more important than in a library where desperate users at least have shelves to browse.

Like the man who had been writing prose all his life without knowing it, librarians and indexers have been producing and standardizing metadata for centuries. Ignoring this legacy, an immense variety of other players have recently entered the field, and many of them have no idea that someone else has already "been there, done that." Different systems are being developed for different—and sometimes the same—kinds of information, resulting in a chaotic atmosphere of clashing standards.

Metadata is crucial to searching. If searching is, today, largely a matter of matching query words with words in the text of articles, then anything that makes the matching process easier or more standardized is bound to improve the process. Metadata is expected to improve matching by standardizing the structure and content of indexing or cataloging information.

→

WHAT IS METADATA?

Metadata is data about data. It describes the attributes and contents of an original document or work. The DESIRE project (http://www.ukoln.ac.uk/metadata/desire/overview/rev_ti.htm) describes metadata as "data associated with objects which relieves their potential users of having to have full advance knowledge of their existence and characteristics." In other words, standard bibliographic information, summaries, indexing terms, and abstracts are all surrogates for the original material, hence metadata.

...standard bibliographic information, summaries, indexing terms, and abstracts are all surrogates for the original material, hence metadata.

The term is generally applied to electronic resources (though it doesn't have to be) and refers to "data" in the broadest sense—datasets, textual information, graphics, music, and anything else that is likely to appear electronically. While the concept includes indexing and cataloging information (information for "resource discovery" in Webspeak), it can go far beyond conventional document representations, such as MARC records.

In fact, because most search engines are text-based, it is essential to add a text description to non-text files if anyone is to find them. Databases of images are only beginning to be searched by non-text means, such as color charts, or by matching faces or similar pictures. These non-text databases abound: spatial information, not just geographic or political names, but coordinates of latitude and longitude, altitude, or depth; data that describe the forces of a tornado; infrared images of earth resources; NASA images from Mars or the moon; databases of famous musical themes, or museum collections.

Some systems that carry information about works are a kind of metadata that is not cataloging information, but which is quite familiar. These include systems such as the Platform for Internet Content Selection (PICS; http://www.w3.org/PICS/) for rating or filtering of information. Information about authenticity, availability and accessibility, digital signatures, copyright, reproduction, etc. is also metadata. Descriptions of the data in a dataset facilitate determination of the value of the dataset for a particular purpose. Similarly, descriptions of the provenance and transformations that have been performed on original electronic works are important for certain kinds of objects.

Some metadata is designed for computers to use. For instance, it indicates the original format in which a work was created, so the computer will open both the application and the document automatically. Indications about the location of mirrored data help make the best use of bandwidth. Administrative metadata aids management of an object in a particular repository, with information about dates of creation and modification or responsibility for maintenance.

Metadata serves several functions. First, it acts as a surrogate for a larger whole. It must characterize the original work sufficiently for the user to understand its contents, as well as its purpose, source, and perhaps conditions of use. This is vital for large collections of data, such as the NASA EOSDIS data, which are enormous data sets.

In addition to its descriptive function, however, a metadata scheme, to be successful, must also establish standard structure and terminology. It benefits no one to have fields labeled "creator," "author," "sculptor," or "composer" if these fields, which all serve the same function, cannot be mapped to the same single concept. Similarly, creating an "author field," but allowing use of "Joe Smith," "Smith, Joseph R.," and "Smith, J.R." as alternative forms of the same name does not serve the purpose of grouping all the works by the same person, regardless of what form of the name is used. One form needs to be established, through either an automatic authority list or a standard controlled vocabulary, and then links need to map alternative forms to the established form. (I still remember, as an undergraduate, going away puzzled from a music card catalog that had nothing listed under "Tschaikovsky." Everything had been entered as "Chaikovskii," with no cross reference given.)

It is this larger concept of standardization that information professionals understand and use. They are concerned both with *how* to write down the descriptive information and *what* to write down. In contrast, most current non-library approaches to the problem address the structure of the data rather than its contents, and this represents a severe shortcoming. In other words, they are concerned with what fields are established, but not particularly worried about which terms are put in them. For useful overviews of metadata, see Baca [1] and Cathro [2].

THE NEED FOR METADATA

The idea of metadata as a new concept arose outside the traditional text and bibliographic arena. As files of data—especially geospatial data—were developed, it became evident that surrogates were required to provide more information about the dataset contents. Increasing numbers and types of objects were being made available digitally, but it was recognized that raw data was of little value without information about how it was collected, the purpose for which it was intended, formats, platforms for viewing and manipulation, and restrictions on reproduction and reuse, aside from more conventional identification information, such as author or producer, title, subject, and abstract.

Furthermore, the value of metadata elements is limited if there is no common agreement on what elements to use or what their content should be. They cannot be searched with any confidence; they might even be unintelligible when found.

Metadata cannot fully serve its purpose unless it is subjected to a

All of the reasons why indexing and cataloging are needed for print resources apply even more emphatically to metadata for electronic documents.

certain amount of standardization. For instance, if a hotel receives three stars from Michelin, and a check mark and two stars from the Mobil Guide, in what way are these ratings comparable? What are the two systems rating hotels on? How many stars are the maximum given by each?

Similarly, just imagine trying to search for content suitability ratings if one system rates content as a whole, another specifies type of content (sex, violence, etc.), and yet another just rates by age. All of the reasons why indexing and cataloging are needed for print resources apply even more emphatically to metadata for electronic documents. There is no hope of effectively browsing the millions of documents sitting on thousands of servers around the world without some sort of aid.

Just as in the traditional cataloging and indexing world, different levels of metadata are needed, depending on the type of object and the use for which it is intended. Resources whose value is ephemeral may warrant only minimal description, while those of permanent research or commercial value may need much fuller description.

CREATING METADATA

Metadata can be created at the time of creation of an object, either by or under the auspices of its creator. It can also be added later as part of the traditional cataloging process. The former mode of creation is expected to predominate, largely because the traditional third-party methods (a.k.a. cataloging and indexing) simply cannot cope with the massive and rapidly growing number of electronic objects in existence.

Federal agencies have taken the lead in providing Web-based forms for entry of metadata; completion of one of these forms by an object creator or owner results in the creation of a metadata record for that object. The Federal Geographic Data Committee (FGDC) and Environmental Protection Agency (EPA) both have such systems in place (http://130.11.52.178/metaover.html and http://www.epa.gov/regional/epafield.html).

The Web Developer's Virtual Library provides a guide to Web page developers on use of "Meta Tagging for Search Engines," which describes how each of the major engines processes data in the META tag (http://wdvl.internet.com/Location/Meta/Tag.html).

Even though provision by the creators will have to dominate, there are numerous efforts to "catalog the Internet," often in the form of voluntary efforts by libraries and library organizations, or by specialists in a particular area. OCLC's Internet Cataloging Project (InterCat) is a good example (http://www.purl.oclc.org/net/InterCat). Typically, such efforts are designed to cover the "best" sites, or those of most use for a particular purpose. They may include third-party evaluation in addition to the typical descriptive information.

SEARCH ENGINES AND METADATA

Of course, metadata is nothing new to professional searchers. We have been able to improve the precision of our searches through the use of controlled vocabularies, or limit the searches to the descriptor, identifier, author, title, or source fields for many years. In the Boolean world, metadata made it possible for us to find information without too much extraneous nonsense (such as false drops) creeping in.

In the statistical/probabilistic search world, however, metadata has been largely ignored. Theoretically, searching on the entire contents of a document was far superior to searching on a limited piece of that text. Moreover, the newer search engines, particularly Web search engines, have been designed to search on ill-assorted collections of unstructured text. There is no hope of cataloging the enormous array of Web pages in a systematic fashion. Applying controlled vocabularies and thesauri used by experienced, trained indexers or catalogers is much too time-consuming for the Web.

In the culture of the Web, the author of the document is the person who applies metadata, and how do we get millions of non-information professionals to understand the import of cataloging to a certain level and standard when even professionals don't always agree? However, beyond the purview of the Web, the use of metadata within statistical search engine collections, such as company intranets, has great merit. Using metadata combines the precision of Boolean searching, to a certain extent, with the necessary fuzziness of statistical or natural language searching. It introduces consistent language, but also allows new or different terminology to be included.

Statistical search engines all have complex and proprietary algorithms that assign weights, or relative importance, to various words. These are based on such elements as the rarity of the word in the collection, the frequency of the term in a document, the position of the word in the document (in the title or near the top is often given more weight), and how close the query terms are to each other within the document (closer is better). When relevance is calculated,

...beyond the purview of the Web, the use of metadata within statistical search engine collections, such as company intranets, has great merit.

these factors are added up, and the highest scoring document appears at the top of the list. If metadata were added to the equation, the weighting algorithm would have to be adjusted. Companies like Fulcrum already allow their customers to adjust the weighting algorithm, placing more or less importance on various data elements. The metadata, if well chosen, should describe the central topics of a document. Thus, it should be given a high weight, relative to the appearance of those terms in the full text of the document. Any document having the query term in its metadata should appear quite high on the ranked list of search results.

In addition, new approaches could be developed that map the metadata and query terms to a cluster of words that are related. In fact, this kind of clustering has already been tried, with varying degrees of success. Some of the Web search engines, like Excite, do "concept searches," which are based on co-occurrence of terms within the database. In other words, if one term keeps appearing near another, there should be some sort of relationship between the two. Therefore, the user should be interested in seeing documents that contain both terms, or either one.

Another use of this idea is to disambiguate terms. If ambiguous terms like "bank" occur within the context of a river and other watery words, then the "bank" in the document is probably not a financial institution. If the user is looking for erosion of river banks, then a document concerning erosion of cliffs on the shore would also be of interest. Metadata would help to pin down the kind of bank by using specific terms to describe the different meanings. The search engine would start at this core meaning and add related concepts to introduce new, but related terms to the user.

HOW DOES METADATA AFFECT SEARCHING?

Searching today is essentially a process of matching the query terms to the words in a document. Barring some innovative searching methods, if the terms don't match, then the document will not be retrieved, no matter how much it is about the subject of the query. Therefore, one of the greatest barriers to finding information is the difficulty of coming up with the right terminology.

Lists of standardized subject heading terms, structured thesauri, and fielded searching were created to remedy this problem. There are several reasons why these tools were created:

- To make sure that all the materials about the same subject were found together either on the shelf or in an online database.
- To single out important concepts from those which are merely incidental to the work.
- To ensure that the same information was found for each work, and that it was put in the same place, so that someone searching for works by an author named Fields would not find them mixed with agricultural tracts on fertilizing wheat fields.

Proper use of indexing vocabularies and field structures, both in searching and in cataloging, increases precision and minimizes the chance of false drops.

Metadata also attacks three language problems that cause poor precision:

- **Polysemy.** Most words in English—and in other languages as well—have multiple meanings. For example, if we are searching for an article that discusses types of springs and their uses, we might retrieve articles on freshwater springs or on the season of spring, as well as on leaf springs, flat springs, or coil springs. Or take the word "construct," which can be both a noun (in mathematics) and a verb.
- **Synonymy.** Many words represent the same concept, although they may do it with different shades of meaning. Take the words "ball," "sphere," and "orb," or "scuba diving" versus "skin diving." If I look for scuba diving, but the term used is skin diving, I will miss materials I might otherwise find. Good metadata should draw these materials together, despite their use of different synonyms.
- **Ambiguity.** If we return to our example of springs, we can see that what differentiates these meanings

→

If all documents carry the same fields, and also use the same controlled vocabularies, then we should be able to improve searching.

is their context. It is unlikely that an article on coil springs will also discuss water quality. The other words used in the article, and the processes described, will be entirely different. A search engine must understand the meaning, not just be able to match the spelling of a word, if it is going to differentiate between different meanings of the same word. While new natural language processing search engines can often determine these differences by the context of the text in the document, Web and Boolean engines cannot, although probabilistic search engines like those on the Web do determine relevance through co-occurrence of query terms in the document. However, the addition of controlled vocabularies in a subject field can distinguish between different kinds of springs or constructs. The goal of a controlled vocabulary, then, is at least to have each term mean one thing, and one thing alone. By using these controlled terms, we can *disambiguate* a term, as a human or a natural language processing system would.

Metadata cannot only improve precision, it can increase recall of pertinent documents by using the same standardized term for each occurrence of a subject. Thus, a document will be retrieved from properly applied metadata even if it never uses the controlled term in its text.

It is the plethora of data and documents that is driving the development of metadata standards. People are having trouble finding the right information, and they are wading through too much of the wrong information. If all documents carry the same fields, and also use the same controlled vocabularies, then we should be able to improve searching.

Those are two big IFs, however. It has taken the library profession years of innumerable conferences, committees, and proposals to establish this kind of standard. And that is only one profession, with a limited number of major institutions that act as leaders. In our case, the Library of Congress emerged as one *de facto* standardizing institution. And it had by no means the last word on controlled vocabularies. How many searchers own copies of *LC Subject Headings,* but also the *INSPEC Thesaurus,* the *Computer Database Thesaurus,* the *ASIS Thesaurus,* the *Ei Thesaurus,* the *Thesaurus of Engineering and Scientific Terms,* etc.? If we couldn't control our controlled vocabularies in pre-Web days, how can we come to agreement now, with every discipline and country in the world involved?

Also, these processes take time, and time is not a plentiful element in the electronic universe. For instance, how long did it take for the Library of Congress to admit the word "computer" to its select vocabulary in place of "electronic data processing systems?"

start up its own metadata definition effort, and creators are free to use whatever tags happen to come to mind. Conversely, any group that is developing a metadata set is free to limit its work to its narrow interests; it need not take a broader view unless it voluntarily chooses to do so.

Developers of resource discovery-oriented metadata systems have been active in developing crosswalks between their systems. The UK Office for Library and Information Networking (UKOLN) site lists about 15 of these (http://www.ukoln.ac.uk/metadata/interoperability/). For art and object identification systems, Baca [1] provides a crosswalk between Categories for the Description of Works of Art (CDWA) and eight other metadata systems. The simple fact that these crosswalks have already been found to be necessary is an indicator of the chaos in the field.

Even if common metadata elements are used, there is no guarantee that the vocabularies, the content of the elements, will be compatible. For instance, the defined list of Resource Types in the Dublin Core is strongly oriented to the needs of libraries and similar agencies, and does not fully meet the needs of other communities, such as software or geospatial data. There is a serious possibility that the situation may grow more chaotic and that metadata users will have to learn a different set of conventions for each kind of data. This is particularly likely in communities that do not have a tradition of controlled-vocabulary indexing and therefore are unlikely to understand the need for predictability in index terms. Some of the efforts are gravitating toward use of Library of Congress Subject Headings (LCSH) as a resource, even if not as an authority. Use of LCSH is probably determined more by the fact that it is there, broad in scope, and

Probably the biggest stumbling block in the way of orderly development of metadata is the sheer number of different metadata projects.

PROBLEMS AND STUMBLING BLOCKS

Probably the biggest stumbling block in the way of orderly development of metadata is the sheer number of different metadata projects. (See the accompanying article, "Metadata Projects and Standards," on page 32.

While there is a certain amount of voluntary coordination at the very top level—for instance, developers of the Dublin Core are closely involved in the World Wide Web Consortium's (W3C) RDF effort—any group may

available for use, rather than by its inherent quality or suitability for electronic searching. In fact, its use of bound terms may make it less useful for searching.

Problems of both definition and application grow out of the fact that in most cases the creators of objects are the ones most likely to apply metadata to their creations. Back-of-the book indexers have a great deal of experience with this issue; unless the author of a book is a trained indexer, she is not likely to produce a high-quality index. True, the author/creator knows her creation better than anyone else, but this does not mean that she is able to step back to see how it fits into the universe of knowledge or objects, so as best to apply useful tags.

There are other cases in which the creators of objects understand only too well how to apply tags that are useful from their point of view, even if not from that of searchers. Spamming is the bane of WWW search engines. Many Web developers are intent, for reasons of pride or because of economic incentives, on gathering as many visits to their sites as possible. They use metadata as well as other techniques to invisibly stuff inappropriate words, or multiple occurrences of the same word, into their documents.

Since search engines count the number of occurrences of a search term in a document—the higher the occurrence, the higher the ranking—inserting repetitions of the word into a document would change how it was ranked. The search engine developers have had to program filtering routines to minimize the ability of creators to use this technique. Some of the engines even ignore metadata completely, reducing the utility of adding metadata in order to increase searching accuracy.

PROGNOSTICATIONS

The only safe prediction with regard to metadata is that all other predictions, no matter how authoritative, will probably be wrong. There are too many players with too many different agendas, resulting in tremendous volatility. The capabilities of the technology are not stable; the technology we use is constantly changing, altering our priorities and the way we meet them. Having said this, we will boldly make some predictions anyway.

- For some time to come, the number of players in the field will continue to increase. More communities and sub-communities will want to make sure that their resources are covered by metadata schemes.
- As metadata schemes proliferate, so will registries of the schemes, until there are registries of registries.
- At the same time, there will be some settling toward a smaller number of "standards" in use by major groups, with a massive scattering of outliers and nonstandard or even *ad hoc* element sets.
- As fast as an element set is developed and standardized, one or more sets of guidelines and/or interpretations will also be forthcoming. The guidelines will be aimed at assuring that creators of metadata are consistent. The interpretations will be provided by major creators of metadata and will describe how *they* choose to implement the elements. Bit players will have to follow along or be out of synch.
- The situation will be similar with enumerated lists for such elements as resource type. There will be a few major "standard" lists, with various communities developing their own extensions.
- National borders will become even less relevant than they are today. The Internet itself is inherently ignorant of national borders, and users of metadata will continue to want information across borders.
- Cross-language metadata standards will be developed.

All of this is very much in the future. With no governing body, and no central profession, as well as splinter groups with large investments in having their proposal become widespread, the outlook for a single structure and thesaurus of metadata seems dim. For now, we recommend that the "meta-data" crowd at least standardize the spelling of their preoccupation to "metadata."

We are left with the idea that metadata, whatever its spelling, is immensely useful. Yet there are still questions remaining to be answered:

- Who will make a final decision about which fields to use and which not to support among competing proposals?
- Who will apply the metadata—the cataloging and indexing?
- Will we have controlled vocabularies, and how can we create one that touches every subject and idea, including the ones invented today?

Is this a hopeless endeavor? We don't think so, but it seems apparent that we will have to hope and push for some very "meta" level standards to be developed. These, if combined with clever use of new information technologies that will both disambiguate uses of terms and map synonyms to the same concept, could create clusters of related meanings not only across disciplines, but across languages. It's quite a tall order.

Is this a hopeless endeavor? We don't think so...

ACKNOWLEDGMENT

Gail Hodge was of great assistance in locating some of the metadata developers, particularly those in the federal government.

REFERENCES

[1] Baca, Martha, ed. "Introduction to Metadata: Pathways to Digital Information." Getty Information Institute, 1998.

[2] Cathro, Warwick. "Metadata: An Overview." August 1997. (http://www.nla.gov.au/nla/staffpaper/cathro3.html)

Communications to the authors should be addressed to ***Jessica Milstead,*** *Principal, The JELEM Company, P.O. Box 5063, Brookfield, CT 06804; 203/740-2433; Fax 203/740-1152; milstead@ct1.nai.net;* and/or *to* ***Susan Feldman,*** *Principal, Datasearch, 170 Lexington Drive, Ithaca, NY 14850; 607/257-0937; sef2@cornell.edu.*