ED 454 856 IR 058 147

| | |
|---|---|
| AUTHOR | Dillon, Martin |
| TITLE | Metadata for Web Resources: How Metadata Works on the Web. |
| PUB DATE | 2000-11-00 |
| NOTE | 22p.; In: Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web (Washington, DC, November 15-17, 2000); see IR 058 144. |
| AVAILABLE FROM | For full text: http://lcweb.loc.gov/catdir/bibcontrol/dillon_paper.html. |
| PUB TYPE | Opinion Papers (120) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150) |
| EDRS PRICE | MF01/PC01 Plus Postage. |
| DESCRIPTORS | *Access to Information; *Cataloging; Knowledge Representation; Library Role; *Metadata; Standards; *World Wide Web |
| IDENTIFIERS | *Dublin Core; Web Sites |

ABSTRACT

This paper discusses bibliographic control of knowledge resources on the World Wide Web. The first section sets the context of the inquiry. The second section covers the following topics related to metadata: (1) definitions of metadata, including metadata as tags and as descriptors; (2) metadata on the Web, including general metadata systems, resource description, PICS (Platform for Internet Content Selection) and other content controllers, the BizTalk and SOAP (Simple Object Access Protocol) frameworks, and rights management; and (3) the resource description framework, including the future of XML (eXtensible Markup Language). The third section addresses issues related to the Dublin Core metadata standard, including degree of completeness, institutional support, implementation, extensibility rules, and difficulties with the creator and relation elements. This section also considers difficulties with the object-attribute model. The fourth section discusses the role of libraries in Web resource description, including reasons why searching alone will not replace the need for human cataloging in the near future. The fifth section presents recommendations related to the following three options for libraries to provide access to knowledge resources on the Web: use or adapt MARC/AACR2; create a library metadata system with the same aims as the Dublin Core; or use or adapt the Dublin Core. Several relevant Web sites are listed. (MES)

# Metadata for Web Resources: How Metadata Works on the Web

by

Martin Dillon

Final version [1]

## 1 Context of our Inquiry

First a brief, blunt statement of the context for our current activities. We are living through a revolution in knowledge representation. After a long and various evolution, knowledge representation settled into paper products for most of its output. Now we are shifting to digital forms for representing knowledge and to the Web as the primary distribution channel. This change will have profound consequences. There is little question, for example, that paper products will gradually be replaced by Web-accessible digital products. Is the Web here to stay? A premise of this paper is that the Web, or its evolutionary successor, will define the shape of our world for decades.

We are addressing questions concerning the cataloging function in this new world, a task that is complicated by uncertainties surrounding the future functioning of the library. Of necessity, one is closely tied to the other. Cataloging, after all, served libraries in a two-fold way: as a means of providing patron access to a collection of knowledge resources, and as a means of managing an inventory of such resources. Both of these were defined primarily as local functions applied to a local collection of paper products, which now will virtually disappear. How will this shift to digital knowledge change cataloging?

Addressing cataloging from the vantage offered above is a question that is central to the inquiry of the conference, but not of this paper. Even so, I want to make one point before I proceed here:

***The library has to be reconceived as a unified cooperative, and cataloging has to be redefined as a function within that cooperative.***

This fact seems painfully obvious but may still be worth stating, since the consequences that flow from it have never been worked out in any detail. Also, I regret to say, few of our colleagues have internalized this fact. Issues arising from managing Web resources from the collective viewpoint are not receiving the attention they deserve. Regrettably, most library activities directed toward providing Web access do so in isolation, acting to control an ocean tide with a teaspoon.

By contrast, cataloging in the paper world has benefited very much from the need to share work products among libraries, even though the results, from the individual library's perspective, were often viewed as primarily benefiting itself. One symptom of this local perspective, which can serve as an example of the split mentality of libraries, occurs when modifications are made to bibliographic records to conform to some local practice. These have long been a source of tension from the perspective of the global cooperative. I would argue, and have argued, that libraries would better serve their constituencies if they universally abandoned local variations in records in favor of record creation to serve a broader community.

In other words, where the bibliographic task in the paper world was defined primarily as the need to fit records into a local catalog, the new task we are designing our systems for is fitting surrogate descriptive records into a universal catalog for Web knowledge resources, with the added need, at least for the foreseeable future of having this catalog work congruently and seamlessly with the bibliography of the paper world.

That brings us to the task of this paper: how do we gain (bibliographic) control over knowledge resources on the Web? We have a new terminology to help us: resource description (or resource discovery) using metadata. I will address the reasons why I distinguish discovery from description below, when we get to the Dublin Core, but first I want to discuss the concept of metadata.

## 2 Definitions of Metadata

Metadata is a recent coinage though not a recent concept. In today's jargon, metadata is data about data: information that communicates the meaning of other information. As nearly as I can tell, the term has come to prominence in our context only with the Web, dating from the early 90's, where it surfaced in the face of a newly recognized need: resource discovery on the Web. (See below in the Notes section, METADATA, the trademark)

We find the first oblique reference to metadata in the "HyperText Markup Language Specification Version 2.0," which discusses "meta-information" in the header section of a HTML document:

*Meta-information has two main functions:*

- *to provide a means to discover that the data set exists and how it might be obtained or accessed; and*
- *to document the content, quality, and features of a data set, indicating its fitness for use.*

(/http://www.w3.org/MarkUp/html-spec/html-spec_toc.html)

The first of these bullets targets resource discovery; the second targets resource description. The first mention I can locate for the term "metadata" used in this sense occurs in the Geospatial community and its efforts to define resource description systems for geospatial data: "Content Standards for Digital Geospatial Metadata Federal Geographic Data Committee," dated June 8, 1994.

At the risk of adding to the confusion surrounding this term, I would like to expand the concept of metadata to include a second type: data labeling. Indeed, this type of metadata can be viewed as primary, as more basic than resource description. I would like to elaborate briefly both forms of metadata.

### Metadata as tags

The most common form of this type of metadata arises from the use of tags to characterize the content of fields. This kind of metadata has a great variety of uses. It is found in all information forms: survey instruments, purchase forms of all sorts, and yes, tax forms. What all of these forms have in common is that they contain labeled fields: a text definition followed by a blank space. The different fields are meant to be filled in and later processed. Labeled fields of this sort are also found in all commercial record keeping, most particularly in the world of electronic data processing, where such standards as EDI have been promulgated to allow information exchange among cooperating commercial firms.

Our focus is exclusively on fields defined by the tagging that occurs in markup languages. SGML was the first of a series of standards that were initiated in the late 80's and has recently culminated in XML. The tags in these systems occur in pairs; each pair defines and delimits a field, with the contents of the field occurring between the two tags. All markup languages (SGML, HTML, XML) make use of this kind of metadata. A simple example:

```
<title> Any title </title>
<publisher> Amazon.com </publisher>
<price> $12.50 </price>
```

Each field (or element in the terminology of markup languages) has a start-tag (<...>) and an end-tag (</...>). The character string within the brackets identifies the field; the area between the start-tag and the end-tag contains a character string that is the value of the field. In the above example, the pairs of bracketed names: <title>, </title>; <publisher> ,</publisher>; and <price>, </price> are the metadata; these metadata convey information about the character strings within each of the pairs. The data thus described are 'Any title', 'Amazon.com', '$12.50'.

This kind of metadata has the advantages of simplicity, machine and human readability, and great expressive power, as HTML has demonstrated in the Web environment. Until recently, HTML tagging has been used to "mark up" all Web content, promiscuously conveying information about formatting, linkages and descriptors.

### Metadata as descriptors

But here's the kicker: In our example above, the strings occurring between each start-tag and end-tag are also data about data: they are also metadata. In the example, they are about a publication and are therefore bibliographic in nature.

When discussed in a Web context, the term "metadata" can refer to either type: the tagging system that defines a set of fields and its contents, or the contents of certain fields that act as descriptors for other resources. This duality can create confusion and it doesn't help that the same string of characters can act as metadata on one level, and data on another, depending on the perspective being used.

## 2.2 Metadata on the Web

In tackling the problem of providing descriptive surrogates for library-related Web resources, we have to be concerned about both kinds of metadata for the following reason: the tagging systems for Web pages, and the conventions and standards for processing them, create the context within which library practices reside; the infrastructure of the Web is driven by them and creates the opportunity for us to build within it a means to achieve our own ends. Since it is the crucial underpinning for our own efforts, before we focus on resource description, we need to discuss briefly the general use of metadata tagging in the Web environment. Such tagging has had a wide variety of applications on the Web independent of libraries. Each application has had its metadata standard proposed, debated, implemented and sometimes abandoned. We will consider some as preparation for our library applications.

### General Metadata Systems

By general metadata system, I mean a methodology for fully characterizing all of the data for an application. The two primary examples of such general systems are:

- *"The Meta Content Framework Using XML," a proposal submitted to the World Wide Web Consortium (W3C) in June 1997, Netscape's major contribution to the metadata initiative.*
- *The "Channel Definition Format," submitted in March 1997, is Microsoft's major contribution to the metadata initiative. It "extends XML and Web Collection work that the W3C" has worked on. CDF is the "industry's first" channel framework for push technology on the Web.*

It will not benefit us here to do more than mention general metadata systems other than to state that their primary aim is to enable the precise mark up of data streams for system interoperability.

### Resource description

Problems of resource description have pervaded the Web since its beginnings. Not surprisingly, however, metadata for resource description have not always been provided explicitly in Web pages. The "Head" section of the HTML Standard was introduced in version 2.0 (early 1994) when the Web was 2 years old. It included the "Meta" element for the first time with such attributes as "title". Metadata in this form proved very popular, with its use growing very rapidly. By 1998, 70 % of public Web sites made use of them, with an average of 2.75 meta fields for each site that used them. ("Web Characterization Project: An Analysis of Metadata Usage on the Web," Edward T. O'Neill, et al) (www.oclc.org/oclc/research/publications/review98/oneill_etal/metadata.htm)

This form of resource description, our primary topic here, engages virtually all Web users, and ranges from search engines and directories of all types to the identification and discovery of special interest communities.

### PICS and other content controllers

The Platform for Internet Content Selection (PICS), an activity related to resource description, both historically and practically, is based on the desire to filter or restrict access to materials of certain types. The most obvious is pornography and the filter or restriction is with respect to juvenile access; but there are many cultures that wish to restrict access to other materials, mostly of a political nature. How to do this within a Web context is the primary question, and the answer is through characterizing the content of resources from this vantage. The O'Neill study noted above does not find much use of PICS tagging. See (www.w3.org/TR/REC-DSig-label/#DSig_1_0_Overview) or (www.w3.org/PICS/) for further information on PICS.

### Commerce - BizTalk and SOAP

From a Microsoft June, 1999 press release, "the BizTalk Framework is an open specification for XML-based data routing and exchange. The BizTalk Framework makes it easy to exchange information between software applications and conduct business with trading partners and customers over the Internet." SOAP, the "Simple Object Access Protocol" developed by Microsoft, "is a lightweight protocol for exchange of information in a decentralized, distributed environment. It is an XML based protocol that consists of three parts: an envelope that defines a framework for describing what is in a message and how to process it, a set of encoding rules for expressing instances of application-defined datatypes, and a convention for representing remote procedure calls and responses." (Taken from the document submitted to the W3C recommending the formation of a working group for Web protocols (Simple Object Access Protocol (SOAP) 1.1, W3C Note 08 May 2000) See (www.microsoft.com/biztalk/) for details.)

Depending on whom you talk to, BizTalk and Soap are either an alternative to the Resource Description Framework (discussed in the next section) or a complement to it. In either case, the existence of both, with neither giving any evidence they are aware of the other, is indicative of the diffuse effort that reigns in the Web arena over how to solve the need for interoperability and data exchange among distributed applications that are the norm on the Web.

### Rights Management

And one such distributed application is the management of intellectual property rights on the Web. The need is to protect intellectual property rights on the Web and enable commercial publishers to control effectively the electronic transfer of such rights. The International DOI Foundation, in collaboration with commercial publishers, is responsible for advancing the definition and uses of the Digital Object

Identifier (DOI(r) ), and is among the leaders in the endeavor to manage property rights. The DOI is "an identification system for intellectual property in the digital environment." Its principle objective is "to develop automated means of processing routine transactions such as document retrieval, clearinghouse payments, and licensing." (http://www.doi.org/index.html) Metadata arises in this context as a means to identify, describe, and allow the tracking of all manner of intellectual property on the Web, to protect it from misuse, and to enable its creators to be properly remunerated.

Although part of the objective of DOI Foundation is to provide a basic resource description to accompany the DOI identifier, much like the elements of the Dublin Core provides, it is noteworthy that no mention of the Dublin Core occurs on their site.

## 2.3 RDF: the Resource Description Framework

Before concluding this section on general issues dealing with metadata on the Web, and before turning to the metadata of resource description, I would like to discuss briefly the relevance of the Resource Description Framework, henceforward referred to as RDF. The best overview of what RDF is and what it is to be used for remains Eric Miller's "An Introduction to the Resource Description Framework" appearing in D-Lib Magazine (http://www.dlib.org/dlib/may98/miller/05miller.html). From the abstract, "The Resource Description Framework (RDF) is an infrastructure that enables the encoding, exchange and reuse of structured metadata."

From the W3C RDF FAQ:

> *RDF emphasizes facilities to enable automated processing of Web resources. RDF metadata can be used in a variety of application areas; for example: in resource discovery to provide better search engine capabilities; in cataloging for describing the content and content relationships available at a particular Web site, page, or digital library; by intelligent software agents to facilitate knowledge sharing and exchange; in content rating; in describing collections of pages that represent a single logical "document"; for describing intellectual property rights of Web pages, and in many others. RDF with digital signatures will be key to building the "Web of Trust" for electronic commerce, collaboration, and other applications.*
> *(http://www.w3.org/RDF/FAQ)*

It is not clear as yet what relevance RDF has to the library world; more broadly, and perhaps causative, it is not clear as yet what relevance RDF will have in the Web. The attitude of Web practitioners toward RDF varies greatly. At one end of this spectrum is the W3C community, which maintains that RDF will provide the mechanisms to solve many of the interoperability problems in the Web. At the other end is Microsoft, which, so far at least, has exhibited a deafening indifference to RDF. The latter attitude is manifested by a total avoidance of its use within Microsoft's product line, and is an almost reflexive corporate reaction to any standard not created by Microsoft itself. If the Microsoft reaction is indicative of the low rate of adoption generally, then RDF is in trouble.

What does the success or failure of RDF matter to the library community? From the perspective of the library world acting within the boundaries of its own community, successful resource description standards and methods are possible without an RDF. Moreover, as with many other Web developments, RDF will succeed or fail based on the practices of the larger world outside libraries. As is so often the case with emerging standards, watchful waiting is probably the best approach.

### The Future of XML

The future of RDF is tied closely to the emergence of XML. What is the future of XML? First and foremost, it appears clear as of this writing that HTML as the markup language of choice for the Web will eventually give way to XML. XHTML, a recent variant of HTML, was designed to provide a bridge between the two. I have heard numerous optimistic predictions about the pace of this evolution, all of them wrong so far: installed systems are always slower to give way than one would wish. Two milestones will be worth watching for: when half of all new Web pages being written are in XML; and second, when half of all the pages on the Web are in XML. Neither will occur any time soon, certainly not in one year, very probably not in two.

Below, I discuss the impact of this change on library issues. The primary issue, however, remains that we are at the mercy of the general Web community in these areas. Progress will occur at a pace dictated by the needs of large movers on the Web, influenced to some degree by the general problem of resource discovery experienced by all Web users, and also by all of those other applications awaiting an effective solution. If this brief consideration of metadata uses on the Web accomplishes anything, I hope it communicates the diversity of communities engaged in providing standards and also the lack of cohesive efforts and results that have been achieved thus far.

## 3 Metadata Standards for Resource Description

Now that we have gotten through the preliminaries, we can turn to our major topic: metadata used for resource description on the Web. It may help clarify Web efforts to touch first on standards that fall under the general topic of resource discovery but were not designed specifically for Web resources. They include such standards as those developed by the Consortium for the Computer Interchange of Museum Information (CIMI), those standards whose development is funded or directed by the Federal Geographic Data Committee, mentioned above in relation to the term "metadata"; and the Government Information Locator Service (GILS), now used to provide access to government documents. These three standards were developed outside the library community. Examples of metadata standards developed within the library community would include the Text Encoding Initiative (TEI) and the Encoded Archival Description (EAD), which were created using SGML and pre-date the Web, but which have since been converted to XML for use within the Web. Links to all of these are provided in the "Resource Section" below. *None of these can be said to have arisen because of the Web, nor was their initial focus on Web resources*. Rather, they use metadata to provide finding tools for patrons in their respective applications. They are more or less parallel to systems of MARC bibliographic records: they are systems constructed to provide descriptions for various classes of objects in the areas of application, ranging from the contents of museums to archived papers. As with almost everything in today's world, the Web is increasingly important as a mechanism for meeting the needs of users by connecting them to resources, whether those resources are available for use on the Web, or only described through the Web and require further action in the non-web world. Items purchasable through the Web fall into the latter category.

The major Web mechanism for connecting user to resource is the search or directory service. Both make use of resource descriptions either to allow the user to perform a search or allow browsing. Typical and relevant is an OPAC search to locate a book, or a similar search on Amazon.com. In neither case is the book itself available on the Web, at least not yet.

To the extent that the standards referred to above deal with objects not directly usable through the Web, they fall outside my concern here because I would like to focus exclusively on Web resources.

One final point: This distinction between Web resource and objects outside the Web may appear somewhat arbitrary. While deploying metadata systems, there is often an overlap between the two. CIMI, for example, has been and is a very active participant in the Dublin Core community, which is responsible for creating the Dublin Core, the preeminent resource description standard in the Web environment. CIMI participates in the Dublin Core at least in part because so much of its resource description activity is manifested in some form on the Web. Increasingly, it is possible to link to images of museum objects on the Web; these images are Web resources par excellence, and thus very much a target of the Dublin Core community. The same can be said for archival information covered by the EAD community: one day all of these materials may be accessible on the Web.

The needs of these various communities for resource description capabilities create a challenge for standards bodies seeking to create tools that can accommodate them. In their complex combinations, they raise questions about the nature of surrogate records. The Web is so universal, so all-encompassing, that we look toward a time when everything will require its Web surrogate to find its user. This aim implies a need for surrogate languages with great expressivity. The ambition of standards such as XML, RDF and the Dublin Core is to achieve this level of expressivity.

We can now turn to the Dublin Core and assess its attempt to accomplish the lofty aims set forth here. And we will encounter a regrettable limitation on the human condition: when we try for too much, we often deliver too little.

## 3.1 The Dublin Core Metadata Standard

The standard central to our purposes is the Dublin Core, which arose within the diverse standards creation activities of the mid-90's. From the outset the Dublin Core had as its focus resource discovery on the Web. As stated in a 1998 IETF document, "The Dublin Core Metadata Workshop Series began in 1995 with an invitational workshop which brought together librarians, digital library researchers, content experts, and text-markup experts to promote better discovery standards for electronic resources."

([RFC2413] Dublin Core Metadata for Resource Discovery. Internet RFC 2413. (http://www.ietf.org/rfc/rfc2413.txt))

"Discovery standards for electronic resources" - as noted earlier, I have used the phrase "resource description" instead of "resource discovery" because description is more general, and in my view more accurately characterizes what is required. One may claim that an effort is restricted to resource description, but if one does not deal with user needs effectively, no justification will satisfy. Resource discovery is impossible without resource description; adequate resource description assures effective discovery. The difference is as basic as the difference between a keyword search and an adequate display of results. The former allows discovery; the latter, based on resource description, allows effective selection from an extended list. I will elaborate this more fully below when we discuss alternatives to cataloging.

In library terms, the Dublin Core is a simple system for cataloging Web resources, no more, no less. And it should be judged from that perspective.

## 3.1.1 Issues with the Dublin Core

Many issues surround the primary question of the effectiveness of the Dublin Core, and I would like to list and discuss them briefly.

### Degree of completeness

Unfinished - the most serious problem of the Dublin Core to date. The first official version of Simple Dublin Core was available in 1997 after 2 years of discussion and debate. The first published version of a qualified Dublin Core was made available in July of this year. It is obviously incomplete, with no qualifiers being offered for the Creator, Contributor, Publisher elements. As yet no one has been able to provide documentation, extensibility rules or implementation guidelines for a qualified Dublin Core. What this has caused in the intervening years is the development of various community versions of qualified Dublin Core's. What this has also caused in the intervening years in every community attempting to apply the Dublin Core to a collection is endless debate over what the various elements mean and how they are to be used. What this has also caused in the intervening years is very slow adoption of the Dublin Core as a standard for resource description for the Web. (Again, see O'Neill's report cited above for statistics.)

### Institutional support

Lack of institutional support is not surprising given the degree of incompleteness of the Dublin Core. CORC (Cooperative Online Resource Catalog), a new service from OCLC introduced in July of this year, which incorporates the newly published qualified Dublin Core, is a strong step in the right direction, but much more is needed, including a standards body and procedures for evolving and changing the Dublin Core.

### Documentation

Documentation, of course, must follow on a published standard and can't precede it. After the recent release of a qualified Dublin Core, it may be possible now to provide at least some usable documentation.

### Implementation guidelines

As yet there is no direction on how to implement the qualified Dublin Core in HTML or XML, though this may change at any time.

### Extensibility rules

There is as yet no precise direction on what counts as an allowable extension to simple Dublin Core, or what syntax extensions must conform to. The absence of a clear definition of the syntax of qualifiers continues to make implementation guidelines difficult if not impossible to achieve. Sufficient for this purpose may be the Dublin Core Metadata Initiative (DCMI) publication prepared by the DCMI Usage Committee, which "describes the principles governing Dublin Core qualifiers, the two categories of qualifiers, and lists instances of qualifiers approved by the Dublin Core Usage Committee." ("Dublin Core Qualifiers," July 2000) (http://purl.org/dc/documents/rec/dcmes-qualifiers-20000711.htm)

In that document, two kinds of modifier for elements are recognized: Element refinement and Encoding scheme. The first is characterized by such modifiers as "created" for the date element; the second by "LCSH" for the Subject element, and "URI" for the Identifier element. For explanations and further examples, please refer to the official publication cited above where all qualifiers defined for the current version are presented in a table. I have gone into this level of detail here concerning acceptable qualifiers for Dublin Core because I explore a problem with respect to them in the next section.

## 3.2 Other Issues

The above list of Dublin Core issues may be transitory. Indeed, it is possible that some of them will be removed or at least alleviated by the DCMI July, 2000 publication cited above. What if they were all fixed? Would our need for a resource discovery standard for the Web be satisfied? There are two general areas of concern that I can see. First, if we generously assume that the Dublin Core in its current form is approximately finished, and that its major focus is on "document-like objects", how close is it to an acceptable standard? Will tweaking over time and through experience in its use gradually provide us with a standard we can live with? Or are there major fissures that must be bridged? Second, does the architecture of the Web require a standard that goes beyond an object-attribute model for resource discovery? I would like to discuss each of these briefly.

## 3.2.1 Difficulties with the current form of the Dublin Core

The current structure of the Dublin Core limits its usefulness in critical ways. As outlined above, Qualified Dublin Core currently allows an element to have two modifiers: the first is considered to be a refinement of the element; and the second, the encoding scheme, is considered to modify the value of an element. The distinction between these two types of modifiers, and others that might be used, have been the source of much discourse within the Dublin Core community, one cause of the delay in completing a draft of a qualified Dublin Core. My problem is fundamental and practical and can be expressed by citing, as examples, what I consider to be serious weaknesses in two Dublin Core elements: the Creator and the Relation elements.

### Creator element (and Contributor and Publisher as well)

What is needed for a Creator element (or what I would like to see it have!) is a structure that provides for the name of the creator as its value, a modifier that states whether the name is corporate, personal, or geographic, and a further modifier that is a URI pointing to an authority record for the name. (All modifiers, like all elements in the Dublin Core, are optional.) The capability of attaching a URI to a Creator element would not only obviate the need to include supplemental Creator information such as an email address (which many have recommended, and which I consider to be highly undesirable), but it would also allow, and thus encourage, a far more effective means of authority control in the Web environment. The fundamental Web mechanism is the link; a Creator field should link directly to the authority record . What could be more natural, desirable, powerful? My understanding is that a group is investigating how to handle authority linkages with the Dublin Core; I hope this solution is still a possibility.

### Relation element

The Relation element poses a similar problem arising from the same structural cause: more modifiers are required to give the Relation element what it needs for effective use. A relation element contains information about a "related item". Three pieces of information are required for this element to be a useful Web construct: the name of the relation ("Is part of", "Is version of", etc.), the name of the item (in the simplest case, a title), and, when available, a URI to get to the item.

Under the current structure, we can provide either a name or a URI, but not both.

There is a solution to both of these problems and one in accord with the essence of the Web: define as part of any Dublin Core element a pointer element for "additional information."

## 3.2.2 Difficulties with the Object-attribute model

### Web Resources: the medium is the message

Marshall McLuhan's famous dictum, "the medium is the message", recommends caution in how we understand the workings of a new medium. Our new medium is the Web and what McLuhan meant, I suppose, and what has application here, is that the characteristics of the medium often have greater impact or influence than the actual content. We are moving from a print culture to an online culture. In the present context, the characteristics that are most at issue involve the change from "collections" and "objects" to ... pages and pointers? Resources? To what? And why do we care?

We care for a number of important reasons. It can be argued that AACR2 cataloging is, by its very nature, tied to physical objects, and when we move into a world without physical objects, the target of the cataloging effort becomes fuzzy or without boundaries. This lack of definition may create insurmountable obstacles to the effective application of cataloging principles and practice. I subscribe to this view without understanding it fully, and I will attempt in what follows to explain why.

### Objects vs. resources vs. whatever

Back in 1992 when we undertook to examine "access to Internet resources", ( a project reported in "Assessing Information on the Internet: Toward Providing Library Services for Computer-Mediated Communication," (Spring 1993), Internet Research, 3(1) 54-69) we played a simple trick on ourselves to sidestep the issue I want to discuss here. The trick was tactical and was necessary at the time for us to make progress: *we restricted our investigation to "document-like objects" on the Internet*. We chose this route to make progress because our first meetings had become bogged down in discussions about what sorts of things were on the Internet, how they differed from documents, and what the implications for cataloging were. After a few rounds of profitless discussion and no progress, by fiat we restricted our focus.

What is the essence of the problem? I believe it is in the notion of object-hood and how that notion does not translate very well to the Web. Consider first one of the basic principles of Anglo-American cataloging: the item in hand. Much depends on this concept including a well-defined boundary for the cataloger in the cataloging process. Of course, even in our workaday world where the cataloging target is a discrete physical package, there are severe problems that must be overcome. Many of these arise because of the differences between the class of objects related to what is referred to as the work and the classes of objects in the work's various manifestations. Questions concerning differences between one class of manifestations and another are legitimate and deserve the attention they receive; how they are resolved determines, among other things, when a new record is required for an item in hand, and when an existing record will suffice. Though important, discussions of these issues have often been unsatisfying. It may be that the problems they pose are fundamentally intractable, that cataloging offers a means for creating round holes into which through various compromises we force a collection of square pegs.

In the world of physical objects, part of the problem certainly is the oversimplification encouraged by the illusion that the ground is solid beneath our objects. One example, long a favorite with me, has to suffice. A trivial pursuit question:

*Category: cataloging. What is the smallest difference between two books that will lead to the creation of two different bibliographic records?*

In more general terms, how big does a difference have to be between two objects to justify the creation of a second bibliographic record? We are touching on the question for which the Dublin Core "1:1" rule offers the answer. And the answer may be unwise, wrong-headed or otherwise misguided, but it assumes object-hood: one object generates one record.

The problem I want to address is the following: is object-hood an effective metaphor for successful resource description in the Web? Please remember that we are not dealing with absolutes, either all or none. In the print world, object-hood has its limitations: the concept of serial was invented to deal with one of them and the discussion above exposed a more subtle definitional problem in dealing with monographs. On a scale of 1 to 10, we could say that for monographs, item-in-hand object-hood is 9.8 successful. What degree of success are we likely to achieve using object-hood as the basis of cataloging on the Web?

The "1:1" rule assumes objects as a given. Its primary purpose is to deal with problems arising when more than one manifestation of the same work exists. Simple examples will suffice: differences in format, say PDF and RTF; or different representations of some object, say image or Html. This oversimplifies but does no harm here, because the very notion of recognizable objects is undermined in the Web.

From the perspective of managing those Web resources that are of interest to the library community, the question becomes: how many conform comfortably to the notion of an object; conversely, how often will an assumed object-hood get us into trouble? Is the use of an object as the underlying metaphor a useful fiction? Or is it more apt to get us into a heap of trouble?

It is always useful to bring forward examples from the print world when they are available to shed light on difficulties like the current one. Two occur to me. The first is the practice of faculty creating a collection of readings gathered from disparate sources as a quasi text book for a course. I have never heard of anyone advocating that libraries catalog such an object. But why not? Surely, surrogates for such objects would be useful if the table of contents were included. Would not others teaching similar courses benefit from having access to the description of the book?

Perhaps a more apt example, certainly a more recent one, is the possibility of anyone creating his or her own book by gathering pieces and parts from a large database of books, whose contents are themselves stored and accessible in parts. Not only chapters and sections could be extracted, but pictures and tables and any other pieces at the whim of the purchaser. As depicted by Lisa Guernsey, "Under this model, books have not only turned into streams of electronic bits that are downloaded to hand-held devices or printed on demand. They have also turned into databases -- pools of digital information that people can extract and combine on their own terms." (From "Books by the Chapter or Verse Arrive on the Internet This Fall," NY Times, July 18, 2000)

Clearly, the results of this process are outside the scope of the cataloger.

I would argue that a Web resource is often much more like a fluid, multi-dimensional, multi-layered, constantly changing complex of things and relationships than it is like a simple object. Web resources do not have tidy boundaries.

### Web Resources

It is necessary to probe this issue further. Web resources are different from monographic objects in ways that profoundly change the cataloging problem; this difference is growing: more of the Web can be thus characterized and the distance between such resources and the monographic object is growing.

Most simply, the problematic characteristic of the Web resource is one of extent: it is difficult, if not impossible, to define the extent of a Web resource, to state where it begins and where it leaves off. Try defining these terms: Web page or Web site. They are used ambiguously on the Web and in the literature. Moreover, what relation do they have to the terms: file, directory, or server? The vagueness of the terminology in this area is symptomatic of the vagueness, in physical terms as well as conceptual terms, of the underlying concepts.

Before we can catalog something, we have to know what we are talking about.

## 4 The Role of Libraries in Web Resource Description

We also have to know what we want to accomplish. Barbara Baruth, in a recent article in American Libraries ("Is Your Catalog Big Enough to Handle the Web," August, 2000, pp. 56-60) explores the question of the library's role in resource discovery on the Web. She asks, "Will the impressive second-generation search engines out now or third-generation engines now incubating make the idea of quality-based services such as CORC obsolete?" Future search engines, she continues, may be able to do a fine job, "scouring the net and bringing back tailored results." And finally she asks the sixty-four dollar question, "Is it possible that manual efforts to explore, evaluate, and catalog the vast reaches of the Internet just can't compete [with these advanced search engines]?"

What is the library responsibility with respect to providing access to Web resources? What is its role, and how should it carry out this role? Until we provide credible answers to these questions, it is not possible to chart the future course of libraries, and secondarily, cataloging. Even if we agree with Barbara Baruth's assessment that search technology will improve sufficiently to eliminate the need of human resource description, how long will this take? I am always suspicious, and I recommend this scepticism to all, when delivery is promised of technologies that are not yet in beta test. Experience tells us that the promised date almost invariably stretches into the future.

Let me state my own view: I see no hope that searching alone will replace the need for human cataloging in the forseeable future, that is, the next 5-10 years. Here are some reasons for my view:

### Wrong, obscure or missing information

Searching is similar to automated cataloging in that neither can overcome the absence of data inferable from a resource, and Web resources will not evolve stable self-describing mechanisms for a long time, if ever; such mechanisms are not yet even being broadly discussed. Desired characteristics such as creation date, revision date, and expiration date, just are not easily available from most Web resources. Inappropriate titling, weak or absent content descriptors - we can go on and on. The absence of these descriptors, or their presence in corrupt or unrecognizable form, within a Web resource corrupts the results of any searching; and we can expect such problems to grow for a long time rather than abate.

### Authority control

The problem of coordinating and differentiating names, a modest source of difficulty within the controlled environments of the library catalog and the commercial publishing world, becomes a nightmare on the Web. All of the usual suspects are involved: personal names, corporate names, geographic names, subject descriptors; all now compounded by language and character set confusion on an immense scale.

### Selection

Finally there is the issue of selection. The Web now has over a billion pages, whatever that means. The task of culling from this huge morass the population of stuff that we want to search is almost overwhelming. It can only be accomplished by an equally huge, ongoing effort of thousands of people, effectively coordinated by well-designed online systems.

## 5 Conclusions and Recommendations

Let me take a final quote from Barbara Baruth's article cited above: "The future of library systems architecture rests in the development of umbrella software that digests search results from rapid, coordinated searches of a variety of disparate databases." That is, the job of resource discovery will be accomplished primarily through software directly acting on Web resources without benefit of human intervention, particularly of the cataloging sort. I disagree with this position on a number of grounds, not least that I believe that searching alone will reach a point of diminishing return (may have already). A second, library-centric reason is based on the assertion that if the library role can be encapsulated by such search engines, we can dispense with libraries forthwith: this functionality can be provided by software firms and distributed directly to patrons either as clients or by glitzy Web portals.

I would argue that it is the responsibility of the library to provide effective access to knowledge resources on the Web. If the various commercial services can adequately accomplish this library goal, let's get on with other worthwhile knowledge management tasks required by our patrons. Barbara Baruth is certainly not alone in the belief that such services are rapidly succeeding in this goal. A parallel here is the dependence of libraries on abstracting and indexing services, which provide tools for accessing the journal literature. Nothern Light and Google are Web versions of the same idea.

Let us assume that library intervention is required for successful access to Web resources of interest to patrons. For those resources that are roughly equivalent to documents in the physical world - self-contained, more or less static - the cataloging task emerges in much like its historic form. No small task because there are a great many such objects. Let us continue to ignore that other class of resources, those whose object-hood is in question.

How should libraries provide access to document-like knowledge resources on the Web? If the library community decides that it is necessary to establish a form of bibliographic control for such objects, three paths are open:

1. Use or adapt MARC/AACR2
2. Start fresh creating a library metadata system with the same aims as the Dublin Core
3. Use or adapt the Dublin Core

I will discuss each of these briefly.

### Use or Adapt MARC/AACR2

There may have been a time when this was a useful direction to take but it is long past. The result of such an exercise would have many of the attractive attributes of the Dublin Core, particularly its simplicity and flexibility.

### Start Fresh

A fresh start, guided by the lessons learned from the long parturition of the Dublin Core is an intriguing idea. But is it realistic? Can the library profession manage the rapid creation and deployment of such a standard? Nothing in our history encourages optimism.

### Use or Adapt the Dublin Core

We are left with this final option. It is more likely that we can make progress by either using whatever version of the Dublin Core is current, or, far better in my view, attack the problem of creating a library-specific variant of the Dublin Core that suits the aims of the library. The criticisms of the Dublin Core offered above provide at least a starting point for what such a variant might look like.

As a final point, I would only strongly recommend that at least one action be taken fothwith: that a MARC version of the Dublin Core be developed, with appropriate instructions and examples. The work products of such a MARC include at least the following:

- The list of fields and sub-fields defining the MARC Dublin Core record, including an indicator that the record is a Dublin Core record.
- Necessary documentation with appropriate examples.
- A definition for a MARC input screen to guide local system vendors and utilities.
- A plan to urge cataloging utilities to incorporate this style of record into their editors.

I am not suggesting a multi-year project; my guess is that this work effort could be accomplished satisfactorily in a matter of a very few months.

This MARC version and its accompanying documentation would be suitable for use in library OPACs, if desired, and would be directly convertible to and from any database of Dublin Core records. The advantages of doing this are obvious. It would immediately communicate to thousands of catalogers the essential nature of the Dublin Core and equip them to make use of existing systems and software to create resource descriptions for Web resources. Would this be a solution to our problems? No, but it would put us in the game as it is defined in today's Web world. Consider where we would be today if a library-defined version of the Dublin Core existed 3 years ago. If the MARC Dublin Core was adopted and vigorously applied by thousands of libraries, we would be far better positioned to serve the Web needs of library patrons and Web knowledge access would be far different and far better.

# 6 Notes and Sources

## 6.1 METADATA, the trademark

Thanks to Rick Pearsall, FGDC Metadata Coordinator, I learned that the term "Metadata" was trademarked in 1986 by The Metadata Company (The Metadata Company, http://www.metadata.com). Its invention is credited to Jack E. Myers who is said to have coined the term in early summer of 1969. The trademark should be written with capital letters and should be distinguished from both "meta data" and "meta-data".

## 6.2 Metadata System Examples

### 6.2.1 Content Standard for Digital Geospatial Metadata (CSDGM)

http://www.fgdc.gov/metadata/contstan.html

An outstanding example of metadata definition is that developed for Geospatial data and mandated by the Federal Government.

*The standard was developed from the perspective of defining the information required by a prospective user to determine the availability of a set of geospatial data, to determine the fitness the set of geospatial data for an intended use, to determine the means of accessing the set of geospatial data, and to successfully transfer the set of geospatial data. As such, the standard establishes the names of data* *elements and compound elements to be used for these purposes, the definitions of these data elements and compound elements, and information about the values that are to be provided for the data elements.*

As stated in the documentation for the standard, "The first impression of the CSDGM is its apparent complexity; in printed form it is about 75 pages long. This is necessary to convey the definitions of the 334 different metadata elements and their production rules. Do not let the length dismay you." (http://www.lic.wisc.edu/metadata/metaprim.htm, 'Metadata Primer -- A "How To" Guide on Metadata Implementation') If you are dismayed by its length and complexity, join the crowd!

### 6.2.2 U.S. Geological Survey. Government Information Locator Service.

URL: http://www.gils.net/

A useful source document is available through the U.S. National Archives and Records Administration (NARA). Guidelines for the Preparation of GILS Core Entries.

URL: http://www.ifla.org/documentslibraries/cataloging/metadata/naragils.txt

### 6.2.3 The Consortium for Interchange of Museum Information (CIMI)

From the introduction at the site: CIMI (the Consortium for the Computer Interchange of Museum Information) is committed to bringing museum information to the largest possible audience. We are a group of institutions and organizations that encourages an open standards-based approach to the management and delivery of digital museum information.

http://www.cimi.org/

A useful overview is provided in, "The use of XML as a transfer syntax for museum records during the CIMI Dublin Core test bed : some practical experiences."

http://www.cimi.org/documents/XML_for_DC_testbed_rev.doc

## 6.3 Other Sources

### 6.3.1 INDECS: interoperability of data in e-commerce systems

An international initiative of rights owners creating metadata standards for e-commerce - "putting metadata to rights" . INDECS provided the metadata model for the DOI. The site has links to background information on the INDECS project and its results.

http://www.indecs.org/index.htm

## 6.3.2 Digital Library: Metadata Resources -

The single best source for all aspects of resource discovery metadata

http://www.ifla.org/II/metadata.htm

## 6.3.3 The Resource Description Framework

Dave Beckett's Resource Description Framework (RDF) Resource Guide

http://www.ilrt.bris.ac.uk/discovery/rdf/resources/

The offical source document for RDF defines it as

*Resource Description Framework (RDF) is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. RDF emphasizes facilities to enable automated processing of Web resources. RDF can be used in a variety of application areas; for example: in resource discovery to provide better search engine capabilities, in cataloging for describing the content and content relationships available at a particular Web site, page, or digital library, by intelligent software agents to facilitate knowledge sharing and exchange, in content rating, in describing collections of pages that represent a single logical "document", for describing intellectual property rights of Web pages, and for expressing the privacy preferences of a user as well as the privacy policies of a Web site. RDF with digital signatures will be key to building the "Web of Trust" for electronic commerce, collaboration, and other applications.*

http://www.w3.org/TR/PR-rdf-syntax/

---

Library of Congress
January 23, 2001
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

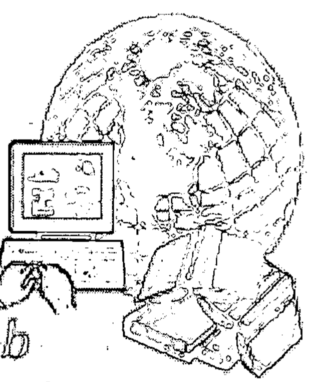
Conference discussion list

Logistical information for conference participants

Conference Organizing Team

# Bicentennial Conference on Bibliographic Control for the New Millennium

*Confronting the Challenges of Networked Resources and the Web*



***sponsored by the Library of Congress Cataloging Directorate***

## Martin Dillon

Former Executive Director of the OCLC Institute
Adjunct Faculty, OCLC Institute

## Metadata for Web Resources: How Metadata Works on the Web

## About the presenter:

From 1970 to 1985, Martin Dillon served on the faculty of the School of Information and Library Science at the University of North Carolina at Chapel Hill, where his research and teaching focused on topics in library automation and information retrieval. He came to OCLC as Visiting Distinguished Scholar in 1985. In 1986, he assumed the position of Director of the Office of Research, where he guided a staff of 30 in research supporting OCLC's mission of improving access to information. From June1993 until he became executive director of the OCLC Institute in January 1997, he served as director of OCLC's Library Resources Management Division, which is responsible for managing OCLC's Cataloging and Resource Sharing services.

As the inaugural director of the OCLC Institute, he led the Institute in forging new ways to facilitate the evolution of libraries through advanced educational opportunities.

## Full text of paper is available

## Summary:

This paper begins by discussing the various meanings of metadata both on

Cataloging Directorate Home Page

Library of Congress Home Page

and off the Web, and the various uses to which metadata has been put. The body of the paper focuses on the Web and the roles that metadata has in that environment. More specifically, the primary concern here is for metadata used in resource discovery, broadly considered. Metadata for resource discovery is on an evolutionary path with bibliographic description as an immediate predecessor. Its chief exemplar is the Dublin Core and its origins, nature and current status will be briefly discussed. From this starting point, the paper then considers the uses of such metadata in the Web context, both currently and those that are planned for. The critical issues that need addressing are its weaknesses for achieving its purposes and alternatives. Finally, the role of libraries in creating systems for resource discovery is considered, from the perspective of the gains made to date with the Dublin Core, the difficulties of merging this effort with traditional bibliographic description (aka MARC and AACRII), and what can be done about the gap between the two.

Library of Congress
June 27, 2000
Comments: lcweb@loc.gov