

CATALOGUING THE WEB: an oxymoron?

Alaska Library Association

February 2006

***Michael Gorman
Dean of Library Services
California State University, Fresno***

Librarians of all stripes have a natural tendency to seek to bring order out of chaos, to neaten, straighten, and generally make more harmonious the unruly and the inaccessible. It is not surprising, therefore, that the idea of “cataloguing the net and the web” is almost as old as the net and the web themselves. I seek, in this paper, to describe where we are in that quest, how we got here, and what the likely future may be. Have we made progress in bringing electronic documents and resources under control? Or have we stood still, or spun our wheels, or traveled in a bewildering mess of fascinating but ultimately pointless circles and spirals? The best place to begin in answering these questions is with the purpose and the meaning of the most favored approach to the problem—metadata. Let us

be clear, despite all the verbiage and the posturing, the idea and purpose of metadata schemes were to give efficient access to online resources and documents without the expense and effort involved in “traditional” library cataloguing and subject access activities. It was, in short, an effort to invent a “third way” between time-consuming and expensive full cataloguing and the scattergun happenstance of search engines relying on free text searching.

The search for the third way arose, I believe, from a number of factors and beliefs. There was, for example the belief that, somehow, cataloguing standards and principles cannot apply to electronic documents and resources. Then there was the ignorance on the part of the early non-librarian creators of metadata schemes of those standards and principles. It was an easy, if not exactly forgivable, assumption that standards and principles developed by librarians for libraries could not be applied outside the library context. However, since libraries deal with the human record

in all its multiplicity and all its formats, the library context encompasses worlds and it is silly to think of it as some kind of “special interest.” Then there was the idea that cataloguing standards and codes were over-complex and that over-complexity made them unaffordable. The metadata people were not the first to bewail the complexity and expense of cataloguing (library administrators have been doing it for decades), but they were, I believe, the first to believe that the complexity was the result of whims of cataloguers. Here is a “reality check”—**the framework standard MARC and the content standard AACR2 are complex because the world is complex—get over it!**

The tiniest sub-sub-field in the one or sub-sub-rule in the other arises from reality and a real part of the human record—no amount of wishful thinking and naïve belief that the human record can be reduced to 15 categories alters that reality. Lastly, there was the quite understandable existential terror in the bosoms of those contemplating the unthinkably vast and inchoate nature of cyberspace—the

lurid, uncharted Internet and the infinite hall of mirrors, wreathed in blue smoke, that is the World Wide Web.

The belief that cataloguing standards and principles do not apply to electronic resources is challenged by the fact that successive editions and revisions of the *Anglo-American cataloguing rules (AACR2)*, the *International standard bibliographic description*, and the relevant MARC format have shown that they can be applied effectively to those resources, as do the existence of format neutral subject heading lists and classification schemes. That belief is also caused by a misunderstanding of the hierarchy within the bibliographic control architecture of libraries. That hierarchy, which proceeds from the most abstract to the most particular, consists of: principles, standards, rules on the content, framework applications, and interpretations.ⁱ One might believe that rules on content (such as AACR2 and LC *List of subject headings*), framework applications (such as MARC), and interpretations (such as the Library of Congress *Rule interpretations*) cannot be applied to

electronic resources (though they can), but that does not alter the fact that the principles of cataloguing and the general standards that have evolved at the international, regional, and national levels are perfectly capable of accommodating electronic resources. The next step would be to create effective and practical rules (content and structure) and interpretations based on those principles and standards. The literature of metadata is replete with references to “MARC cataloguing” and to the International Standard Bibliographic Description (ISBD) family of standards as dealing with content, when, in fact, they are frameworks within which data is presented and transmitted. If metadata were ever to have had a long-term future, it was essential that those concerned with it develop an understanding of the difference between content and the frameworks encasing that content and the difference between principles that determine the nature of the content and standards that embody those principles. Principles include coherence and control of the vocabulary of content—for example, that each

person or subject cited be cited in a standard form that is a unique identifier for that person or subject—and specificity of subject identifiers—for example, that a resource dealing with Lithuania receive the subject designation *Lithuania* and not, say, *Baltic States* or *Eastern Europe* or *Vilnius*. Other principles common to all systems of organization of surrogates for resources (created for the bibliographic architecture of libraries) include standardized order of presentation of content, standardized use of abbreviations, structures that connect related subject designators, standardized order of presentation of complex names and subject designations, use of the language used in the resource itself, and the ability to express relationships between those surrogates.

What about the terror induced by the sheer size of the problem involved in providing access to electronic resources, particularly those feral resources living often-abbreviated lives out in the Internet and the Web beyond the ken of any organization? There were and are only three responses.

The first would be to leave it all to the small mercies of search engines. The second would be to attempt to identify the proportion of feral electronic resources that are worth preserving and cataloguing and apply the principles and standards alluded to above to bring them under control.

The third is the so far ill-formulated and inchoate response—more accurately wide range of responses we lump under the name “metadata.” The problems posed by search engines are so numerous and so serious that they are known to everyone even in this age when Googlemania constitutes the biggest mass delusion since the tulip mania of 17th century Holland.ⁱⁱ Anyone with experience of Google and its less-hyped competitors knows that any search will give you a list of sources characterized by:

- a) size—a Google search on “metadata” will give you
“about 60,800,000 hits, on “metadata+libraries”,
5,400,000, and on “search engines'+metadata”,
1,120,000

- b) random order—I know Google claims to have “powerful search algorithms” that gives you the most useful hits first, but actual experience belies the claim
- c) utter lack of both precision (how many relevant documents are retrieved in a search) and recall (how many of the relevant documents are retrieved in a search)—the standard measures of the efficacy of a document retrieval system.

So, self glorification and PR aside, Google gives you far more hits than any normal person can deal with, in no useful order, and with no guarantee that the items received are relevant or that relevant items have been retrieved. No wonder that undergraduate students, and others who value the speed, ease, and glitz of Google to the exclusion of all other criteria, think that what they are doing is “research” and use whatever comes to hand happily and without a single qualm.

The second response (selecting worthwhile resources and cataloguing them fully) presents many financial

problems. Who is to do the gargantuan task of identifying the worthwhile gold amid the greater dross? Which criteria are to be used? Are such choices even possible in this culturally egalitarian age when taste and evaluations of worth are equally scorned as “elitism”? When the choice of worthwhile resources is made, who is to pay for the expensive labor-intensive task of applying cataloguing principles and standards to ensure their preservation and retrieval?

No wonder the “third way”—metadata, that attempt to combine speed and cost cutting with precision and recall in retrieval has proved to be so alluring and still provides, after the experience of a decade, the rationale for gatherings such as this on which we are engaged. It seems that the high hopes of many have been dimmed, much as the hopes of the “third way” in British politics have been dimmed and we are in a time when, in the words of the abstract of Jeffrey Beall’s presentation at the Canadian Metadata Forum, “The resulting Tower of Babel of metadata schema has made the

sharing of metadata among professional communities increasingly difficult. Although abundant resources have been devoted to the development and implementation of Dublin Core, the schema has largely failed.ⁱⁱⁱ Another session at that forum was called, and the title says it all, “Standards. It’s a jungle out there!”^{iv} Four years ago, a British commentator wrote: “One consequence of this wide range of communities having an interest in metadata is that there are a bewildering number of standards and formats in existence or under development.”^v Day’s article points out that this multiplicity of metadata schemes results from the variety of entities involved—including “library catalogues, abstracting and indexing services, archival finding aids, and museum documentation ... [and] those domains that also have an interest in metadata: e.g., software developers, publishers, the recording industry, television companies, the producers of digital educational content and those concerned with geographical and satellite-based information.” You *can* have too much of a good thing

and this multiplicity of interests involved and schemes advanced with only a bare minimum of common ground do not provide a promising basis for a genuinely different approach lying between full cataloguing and search engine free text searching. Some recent documents speak of metadata as a means of improving search engines by introducing controlled vocabularies—that is, as a means of improving the first way rather than as an alternative. Perhaps this is what the metadata movement will come to—a variety of schemes presenting keywords in internally consistent structures, each different from the other, in order to make the free text used by search engines a little more coherent and structured. More of a whimper than a bang, one would think, and certainly not cataloguing in any normal use of the word.

If the metadata movement has grander aspirations, it will profit from thinking both about the evolution of library cataloguing and consideration of the story of the Tower of Babel.

Behold the people is one and they have all one language ... and nothing will be restrained from them which they have imagined to do.^{vi}

... the LORD did there confound the language of all the earth and from thence did the LORD scatter them abroad on the face of the earth.^{vii}

I am far from an Old Testament scholar and have always been a little confused about the LORD's motivation for cursing the people of the earth with many languages, making it impossible for them to do what they imagined they can do. Be that as it may, the lesson is clear—given a common language and common ground, there was no accomplishment that was beyond the reach of the people. To generalize, people can accomplish anything if they share common aims and methods and if they apply them creatively. "United we stand, divided we fall" is as true today as it was when Aesop wrote it first.

The lesson in all this is the same as that which the history of library cataloguing and classification teaches us.

From the first part of the 19th century into the second half of the 20th century, progress in cataloguing was hamstrung by two fatal flaws. The first was that the technologies by which the results of cataloguing and classification were conveyed (print, handwriting, and typing on paper and on cardstock to create records for card, book, and sheaf catalogues, and microforms of various kinds) were unhelpful in fostering cooperative schemes. The second was the lack of uniformity in the cataloguing practices that gave rise to the records that were communicated by means of that technology. One cataloguing system would have a different set of elements from others. Even when there was agreement on the elements in a catalogue record, there was no agreement on the data to be placed in those elements or the order in which the elements appeared. There was no agreement, even within one language group, on the type and content of the access points used in catalogues to group like records and to give access to individual records. That is, a given author could appear under one form according to one

cataloguing system and under an entirely different form in another, and not appear at all in a third. The lack of uniformity was caused by the proliferation of national codes and local schemes, by the proliferation of different interpretations of cataloguing codes and classification schemes, and by the lack of agreement on the basic principles of cataloguing and the consequent absence of widely accepted standards. The first great obstacle to cooperation in library cataloguing—technology—was set on the path to solution by the creation of the MARC communication format in the late 1960s. That in turn gave rise to great international standardization movement that culminated in the UBC (Universal Bibliographic Control) effort of the 1970s and beyond. The road to an effective technological answer to cooperation in library cataloguing and effective international standardization of the structure and content of those records was long and difficult but ultimately successful.

Why am I rehearsing this old story? I suppose it is because I see the parallels between the state of metadata today and those non-standard, incoherent cataloguing systems of yore and the way in which they were made coherent by the intelligent application of technology and prolonged but ultimately successful international standardization efforts. In a paper published in 2004, Jeffrey Beall provides a threnody for the Dublin Core.^{viii} He states, among other things, that Dublin Core records, were devised to be simple and created “without the high level of training required for ... MARC records,” and that DC records have failed in their aim of facilitating more precise “searching than the search engines available at the time of its invention” precisely because of their simplicity and “lack of standardization of the data.” Mr. Beall believes that Metadata Object Description Schema (MODS)—created and run by the Library of Congress—will be much more effective than the DC and its “bloated bureaucracy.” MODS and its companion Metadata Authority Description Schema (MADS)

are much more complex than most metadata schemes.^{ix}

MODS substitutes XML coding for MARC tags and indicators but has many complex elements and sub-elements—most of which bear a remarkable resemblance to MARC fields and sub-fields. MADS is “an authority element set that may be used to provide metadata about agents (people, organizations), events, and terms (topics, geographics [*sic*], genres, etc.).”^x Perhaps I am missing something, but the substitution of XML for MARC tags and the introduction of authoritative, standard forms of name for people, organizations, subjects, geographic names, etc., would seem to me to be merely a repackaging of the basic ideas behind the MARC framework format, and such content formats as AACR2 and the Library of Congress *List of subject headings*. Beall’s elegy for the Dublin Core is, in truth, an elegy for the idea of the third way—the dream of efficient and precise searching on the cheap, the idea that one could obtain the results of labor-intensive and expensive cataloguing without the training cataloguers need and without the expense. One

of the iron rules of modern life is that you get what you pay for and it is entirely possible that the utopian dream of the third way, as with many of the utopian dreams of the digital age, is succumbing to reality. Perhaps there is and can be no third way and metadata efforts will have to choose between being a relatively cheap way to enhance search engine searching by adding relevant keywords or evolving into a system of cataloguing digital resources that is firmly rooted in the principles and standards of library cataloguing that were refined over almost 200 years and have proven adaptable to all formats. Is metadata about making Google marginally more useful or is it cataloguing in new and trendy duds? If the Dublin Core is dying or dead and multitudinous metadata other schemes are floundering around like so many gaffed fish in an uncharted ocean, then the search for the third way is over. Like the hunters of the Snark, *we sought it with thimbles, we sought it with care; we pursued it with forks and hope* only to find that *it had softly and*

*suddenly vanished away -- For the Snark was a Boojum,
you see.*^{xi}

ⁱ . I am aware of the tastelessness of citing your own works but this (Canadian) publication may be useful here. Gorman, Michael. Principles, standards, rules, and applications. *In AACR2 Seminar papers / edited by Ralph W. Manning.* Ottawa: Canadian Library Association, 1981.

ⁱⁱ . Tulip Mania. *Encyclopædia Britannica*. Retrieved September 19, 2005, from Encyclopædia Britannica Premium Service <http://www.britannica.com/eb/article-9073725>

ⁱⁱⁱ . Canadian Metadata Forum website <http://www.collectionscanada.ca/metaforum/014005-05209-e.html> Abstracts.

^{iv} . *Ibid.*

^v . Day, Michael (UKOLN). Metadata in a nutshell. *Information Europe* 6(2), Summer 2001, p. 11.

^{vi} . Bible (King James Version). O.T. Genesis, 11,6.

^{vii} . *Op. cit.* 11,9.

^{viii} Beall, Jeffrey. Dublin core: an obituary. *Library hi tech news*, no. 8 2004, pp. 40-41.

^{ix} . See <http://www.loc.gov/standards/mods/> and <http://www.loc.gov/standards/mads/>

^x . <http://www.loc.gov/standards/mads/>

^{xi} . Carroll, Lewis. *The hunting of the Snark.* 1887.