HUMAN INTERVENTION IN THE NETWORKED ENVIRONMENT: 
METADATA ALTERNATIVES

David J. Carroll
Pradeep Lele
Lamar University
Mary and John Gray Library
P.O. Box 10021
Beaumont, Texas 77707 U.S.A.

ABSTRACT: This paper emphasizes (despite noises to the contrary a relatively short time ago), the increased importance of the librarian’s role as a "human" interface in the organization and retrieval of resources in the networked environment. Metadata, a 30-plus year-old term, is suddenly enjoying a heated renaissance in today’s overcrowded digital playing field. The long-established MARC format, adaptations of MARC, and several other alternative metadata are compared. The utilization of embedded META tag information in HTML documents is outlined. We describe how existing search engines find and index resources on the Web and detail the pros and cons of these tools. The implications for effective research are discussed in light of the inherent limitations of these automated indexing schemes.

Introduction

There are a variety of listservs and electronic forums which contain postings in a typical year relevant to the issue of metadata. Two years ago, about the time that librarians were beginning to grapple in earnest with the concept of organizing the World Wide Web, a wag from the New York Public Library posted the following reply to a thread in the electronic forum, Web4Lib:

Re[2]: Cataloging the Internet
atroncale@nypl.org
Tue. 17 Oct 1995 10:31:06-0400

How to catalog the internet in five easy steps:
Apply a uniform structure to the file names and their content in servers.
Catalog and index all backbone servers who distribute content. The majors first
the minors second.
SGML [sic] all metadata within.
Script an HTML/SGML browser and provide it for free.
Go home and have a beer knowing you did a good days [sic] work (Troncale 1995).
The person who posted the above was undoubtedly writing with tongue firmly planted in cheek. Still, there was probably a fair degree of seriousness in his reply, the consumption of beer notwithstanding. Librarians like to think that applying bibliographic control to the Internet is a manageable problem, no matter how daunting the scope. It is, after all, the stubborn and indomitable nature of our human spirit.

Metadata was coined as a term in the early 1960's. The terminology has been re-used recently by librarians and others, particularly in relation to the explosion of resources available in the burgeoning, seemingly untamable World Wide Web. For those librarians in the marine sciences, weary of the land-based "information highway" metaphor, a suggested replacement would be that of John Ober of the University of California, Berkeley. Ober prefers to think of Internet resources as a collective ocean, and he is not alone in that conceptualization. So much of the information, after all, is hidden beneath the surface. There are no "highway signs." In the same manner that the bottom of the ocean is always shifting and changing, so, too, does the Internet ebb and flow. Consequently, we must turn to navigational tools, one of which, it turns out, is metadata (Basch 1996).

In our profession, of course, we have been producers and purveyors of metadata for years, even if we were calling it something different – Machine-Readable Cataloging (MARC) and Anglo-American Cataloging Rules, 2nd revision (AACR2) being prime examples. The term metadata has been loosely defined in various ways, most commonly as "data about data" or "information about information." Whether metadata are descriptive or non-descriptive, whether they are embedded within a digital document or object or reside separately in a surrogate, they invariably represent a human contribution to the heavily-computerized process of information retrieval. Humans, it is generally agreed, have the discerning intuitive capabilities that machines do not yet fully possess. Automated generation of metadata, while perfectly satisfactory in some cases, fails by a wide margin in others. Research in Artificial Intelligence (AI) has been performed since the 1950's and, while there have been some breakthroughs, there have also been many disappointments. The most promising avenue of AI currently appears to lie within the area of building neural networks. However, it remains clear that certain human attributes are extraordinarily difficult to simulate – the ability to read and interpret in context, make judgments, handle non-verbal data, and separate the implied from the stated (Ward 1996).

This year happens to be the 50th anniversary of the Association for Computing Machinery (ACM). The first computer, the Electronic Numerical Integrator and Computer (ENIAC), dates back one year earlier than the formation of ACM to 1946. It is difficult to realize that computing is that elderly; so many startling developments have emerged within the last decade alone. In the heady atmosphere of those early days of computing successes, despite routinely "browning-out" large segments of surrounding neighborhood electricity with huge mainframe computation, it must have been relatively tempting to imagine that the day would soon dawn where machines would do literally almost everything for us.
There is a John Prine song entitled “Living in the Future” that speaks to the failure of technology to live up to its impossible early expectations:

We are living in the future, I’ll tell you how I know
I read it in the paper fifteen years ago
We’re all driving rocket ships and talking with our minds
Wearing turquoise jewelry and standing in soup lines (Prine 1985)

Computers, it was commonly held, would solve all of our problems and make our lives completely different. It is certainly indisputable that our lives have been vastly changed with the advent of computing. However, it is equally true that we have inherited a whole new set of problems in lieu of the ones that have dissipated with advances in automation. An overriding problem in our field is the difficulty in retrieving, with precision, digital data from an ever-increasing mass. There is admittedly little apparent problem in simply retrieving data, even from something as huge and chaotic as the Web. Unfortunately, the data we retrieve are all too often accompanied by a clutter of irrelevant or, even more annoying, repetitive output. The key to solving this problem of an unsatisfactory “signal-to-noise ratio” appears to lie at least partially within the arena of significant human intervention in tandem with powerful computing applications. This is precisely the turf that librarians have populated comfortably now for decades.

Premature burial

In the U.S. library profession, we have been somewhat lulled to sleep by our bibliographic triumphs: the initial shared-cataloging success of the Library of Congress (LC) card production program; the subsequent shared-cataloging success involved with the rise of the huge national, now international, online bibliographic utilities, of which the Online Computer Library Center (OCLC) is predominantly in the forefront; and the development and refinement of library Online Public Access Catalogs (OPACS) with their powerful Boolean/keyword searching capabilities and navigational-friendly Z39.50 gateways. All of these achievements have fostered the feeling that we have things well under control from a bibliographic standpoint.

Administrators, quite naturally seeking ways to stretch the shrinking library dollar so that they would continue to be able to purchase automation in tandem with budget-draining physical volume acquisition, have been thinking and talking about outsourcing cataloging and other technical services functions. Other administrators have moved well beyond thought and talk into action, actually phasing out cataloging positions or merging them into other areas of the library operation (Hirshon 1994). Everything of a bibliographic nature, in their view, was so tightly ordered and contained. The Internet, of course, and specifically the evolution of the popular Web with its unifying browser access to a variety of Internet protocols – HTTP (HyperText Transfer Protocol), FTP (File Transfer Protocol), Telnet, Gopher, Usenet and Mail – has inundated this seemingly
contained little bibliographic world with an enormous flood of largely unorganized digital information.

The library community has attacked the problem on several fronts. The Association for Library Collections and Technical Services (ALCTS), a major subdivision of the American Library Association (ALA), has formed the Taskforce on Meta Access to study many different aspects of the metadata/resources description issue. The OCLC Internet Cataloging Project, which resulted in the establishment of the InterCat catalog of Internet resources, was an attempt to test the traditional MARC/AACR2 methods in cataloging digital Internet objects. Bolstered by the Machine-Readable Bibliographic Information (MARBI) Committee's approval of the 856 field in MARC, the experiment tested the standard and proved that it was indeed viable (Dillon & Jul 1996). Still, many in the library community feel that the MARC format is too "flat" a structure for differing types of digital Internet objects. Some are now proposing Standard Generalized Markup Language (SGML) as a more applicable alternative (Younger 1997). What has become increasingly clear is that there will be several approaches to the metadata problem, many of them originating in organizations with parochial stakes in completely differing fields of study. The challenge will be to make all of them effectively interoperable.

Other Non-MARC metadata initiatives

The late Paul Evan Peters described the "network value chain" as a progressive link between data, information, knowledge, understanding and wisdom (Peters 1995). As is often the case, that first step can be the hardest. Data really only undergo the initial metamorphosis into information when they are (a) located, (b) determined to be relevant and useful, and (c) efficiently retrieved. Metadata initiatives are tools to accomplish all of those criteria. Metadata elements not only have the capability to describe a digital object, they can also serve as a source of content evaluation, and an outline of rights management and terms and conditions of usage. More than anything, they can greatly assist in the actual retrieval process—a process which has been largely rendered muddy and imprecise by the crowded and disorganized nature of the digital world.

There are many different metadata approaches under way at this time. The famous Dublin Core, which emanated from a March 1995 workshop between The National Center for Supercomputing Applications (NCSA) and OCLC in the Columbus suburb of Dublin, Ohio, was an important minimalist attempt to reduce descriptive metadata to thirteen core element fields (since expanded to fifteen) (DC 1997). There is a need, it has been argued, to have a common starting place. A fundamental core can then be extended and "crosswalks" can be mapped to other forms of metadata—MARC/AACR2, The Government Information Locator Service (GILS), the Text-Encoding Initiative (TEI) headers, Encoded Archival Description (EAD), Content Standards for Digital Geospatial Metadata (CSDGM) developed by the Federal Geospatial Data Committee (FGDC) to handle the special problems of spatial data, Summary Object Interchange Format (SOIF),
the Platform for Internet Content Selection (PICS) – many of which are exceedingly
more complex than the Dublin core set.

The second Dublin Core workshop, held in Warwick, England in April 1996 by the UK
Office for Library and Information Networking (UKOLN) and OCLC, dealt with, among
other things, the formation of a “container architecture” for handling and integrating
different types of metadata “packages” for the same digital object (Dempsey & Weibel
1996). The third Dublin Core workshop, a collaboration between the Coalition for
Networked Information (CNI) and OCLC in September 1996, examined metadata for
visual resources. Two more Dublin Core workshops this year – one in Canberra,
Australia in March 1997 and one in Helsinki, Finland in October 1997 – have continued
the much-accelerated momentum to reach a consensus on core problems.

The concept of persistence is a crucial issue in all of this. The average Uniform Resource
Locator (URL) of a Web resource has a life span which is typically measured in months,
not years. The URL string was never meant to serve as a permanent “call number” for
Internet documents. The Center for Networked Research Initiatives (CNRI) has been
working for several years on a system of naming that they refer to as “handles,” in order
to combat the constant shifting of URLs from one server to another. The Uniform
Resource Identifier (URI) structure includes three subsets: Uniform Resource Name
(URN), the aforementioned URL, and Uniform Resource Characteristics (URC). The
URN subset is an attempt at persistent naming. The URC, the characteristics of the
resource, actually corresponds to the metadata aspect of Internet digital objects. OCLC
has also been doing related work in the persistence area. Their Persistent Uniform
Resource Locator (PURL) research and implementation has purposely dovetailed with
the CNRI work done on handles (Ahlberg 1996).

Searching the Web

Search tools, and the indexes they create and assemble, represent the user end of the
information retrieval spectrum. Such tools should, in theory, benefit greatly from the
accumulation of systematically-provided metadata. They are, with the exception of
selected directories, a totally automated process. Essentially, a Web search tool consists
of three major elements: the robot, the index, and the search engine. Robots – also known
variously as spiders, web crawlers, worms or ants – are programs that navigate or
“crawl” throughout the Web. They connect to web pages already known to exist, either
from a previous search, Usenet postings, mailing list archives, or lists of manually-
submitted URLs. The robot typically retrieves a Web page in full and subsequently seeks
out and retrieves all of the other documents referenced by the initial page. The robot’s
findings are then parsed to assemble the index, creating, in the process, what is
essentially the lifeblood of the search tool. The search engine software serves as the
actual user interface and provides all of the pull-down box selections and modification
choices for user interaction with the search tool.

63
Indexes with search engine interfaces and selected hierarchical lists, also known as directories, are the two principle types of search tools. The latter is representative of a human contribution to the search process. Yahoo, the phenomenal success story perpetrated by two Stanford University graduate students, remains the quintessential directory – well over one-third of the people who use the Internet make a visit there. The site employs professional catalogers to create their almost gopher-like hierarchies. While Yahoo serves as an excellent source to begin a search for information on the Web, it leaves the search engine aspect to DEC’s AltaVista, with whom they have forged a commercial partnership. Munson (1996) describes how Lycos, one of the very largest of Web search tools, indexes, creates, matches, and sorts results in rank order: “The Lycos index includes the document title, headings and subheadings, the 100 most ‘weighty’ words, the first 20 lines of text, the size of the page in bytes, and the total number of words.” Most of the indexes created in this manner utilize relevancy ranking systems to scan and sort the results.

The issue of retrieving documents from a search in relevancy order is an important consideration, made even more significant by the massive size of the Web. Certainly, this is a carry-over problem that existed, and continues to exist, in large, traditional online databases like DIALOG. The more extensive Web search tools – AltaVista, InfoSeek or Lycos – can simply overwhelm the searcher with the sheer volume of retrievals, sometimes returning results numbering well into the hundreds of thousands. It is estimated that, when AltaVista was launched in 1995, the Web contained around 50 million pages on 100,000 different sites. The present estimate is anywhere from 100 to 150 million pages on 650,000 sites (Brake 1997). The large indexes attempt to overcome this problem by ranking results based on some sort of algorithm or other criteria. Search results which best meet the criteria are then displayed earliest in the hit list that is returned to the user.

Systematically embedding metadata in HTML (HyperText Markup Language) documents has been proposed as a potential method to improve the return of more relevant Web documents to the searcher. As an example, the elements of the aforementioned Dublin Core metadata scheme can be entered into the HTML source code’s META tag (see Appendix A), which is held within the HEAD section of the document. None of the Dublin Core fields are required and all are repeatable as needed. These metadata, if sought out properly by the search tool robots, would provide author-generated information about a variety of access points relevant to the document or object: its creator/author, the publisher, an email address, suitable keywords and other terms, a description or summary of the document, etc. These core elements can either stand alone or be associated with other, more complex schemes. All of this human input ensures that the document has a better chance of being retrieved and, once retrieved, displayed correctly to the user by the search engine interface. The contents of the HTML META tag can be entered by the author of the document or, alternatively, can be created by utilizing programs such as the META Tag Builder, a forms-based HTML tool that
returns a set of elements utilizing what is called DC.lite, an even more abbreviated version of the Dublin Core element set.

There are some genuine problems with the use of META tags, not the least of which is that no search tool, as of yet, uses them exclusively to create its indexes. AltaVista, InfoSeek and HotBot are three search tools that do utilize the content of META tags; however, they also exhaustively scan the entire text of Web pages for keywords. The fact that there is not yet a large body of available metadata existent in the Web is a related problem. Until that situation corrects itself, the commercial search tools will have relatively little motivation to focus specifically on the META tags.

Still another problem is the rash of "spamming" techniques rampant on personal and commercial Web pages. Spamming skews the weight assigned to terms by utilizing terms repeatedly within the META tag or even within the text itself, affecting the relevance ranking in the process. The retrieval of such Web documents invariably has little or no relation to the user's search strategy. A relatively mild example of how an unidentified company achieves this is shown (see Appendix B). By repeating relevant words such as "amortization" and "loan," and by adding other completely non-relevant but common words like "service," "disk," "diskette," "download" and "free," this company attempts to have its page displayed on the first screen of search results. A classic example of spamming outside of the META tag was the Heaven's Gate Web page. The creators of this page successfully spammed "invisibly" by utilizing black text on a black background. Search tool developers are aware of the various spamming techniques and they are becoming sophisticated enough to ignore frequently repeated words in close proximity in their determination of sorting and relevance ranking. However, the spamming phenomenon has economic implications for the search tool companies. On the one hand, there is lucrative advertising revenue to be gained every time a web page is brought up. Conversely, it is absolutely undermining the effectiveness of the search tool.

The steadily increasing number of search tool robots cruising the Web has also been known to tie up large amounts of network bandwidth. The robots import entire pages back to their own sites in order to create their indexes. In addition to tying up bandwidth in general, these frequent site visits of various robots can also make for slower access to the specific Web page they are "visiting" at any one moment (Arnett 1996). It is no small problem. AltaVista's robot, "Scooter," crawls approximately 3 million pages per day, while HotBot's robot, "Slurp the Web Hound," crawls up to 10 million pages per day (Sullivan 1997). In an effort to counteract these bandwidth problems, software called Harvest, developed at the University of Colorado at Boulder, has been created. Harvest essentially reverses the current indexing process, allowing sites to basically index themselves and ship the packaged results to the search tool sites. Unfortunately, the commercial search tools have not yet shown much interest in this approach.
Scholarship issues

The early critics of the Web argued that content, other than that found on governmental pages and a few other key selected sites, was lacking. The Web was relatively flashy from a graphical standpoint but where was the substance? That situation, as the months and years have passed, has changed considerably for the better. The content value of the Web will continue to be upgraded as the intellectual property rights question is satisfactorily resolved and as librarians discover methods to validate the accuracy and authority of Web information. As these developments progress, the inclination to provide corresponding labor-intensive forms of metadata for persistent, high-value sites will certainly strengthen.

One distressing consequence of the current technological climate has been the over-reliance on currency. In exactly the same manner that Web search engines compete furiously to access the newest, “latest and greatest” material on the Internet, more conventional electronic resources – many of the CD-ROM databases in libraries, as an example – seem oblivious to anything published prior to 1987. There are financial reasons, of course, for the failure to digitally archive non-current information. The result, however, is that the majority of students at institutions like our own are probably happily oblivious to the fact that anything of relevance to their topic was written much before this decade. All of this seems to be in direct opposition to the fundamental role of libraries and librarians. Our profession is supposed to be the touchstone to continuity. We are supposed to take the long view. Certainly, even the most myopic long view ought to extend longer than a ten year period (Crowe 1996).

Prognosis

The intent here has not been to disparage automation. After all, automation is realistically our only hope of getting a handle (no pun intended) on networked resources that occupy a “database” which is scaled to the size of the Internet. Rather, it has been the intent of this paper to acknowledge the contributions of humans in the overall automation of information retrieval. All of the work in metadata so far seems to indicate that there continues to be a place for many of us in that process. What is particularly encouraging is that a person does not necessarily have to have a name like Vannevar Bush or Ted Nelson or Doug Engelbart or Bill Gates or Tim Berners-Lee or Marc Andreessen to make a difference in this amorphous landscape. Common people – catalogers, web page developers and maintainers, editors, indexers and others – can contribute through the standardized creation and consistent application of something as basic as metadata.

How will all of the metadata initiatives sort themselves out? Certainly, we have to keep reminding ourselves that it still very early in the game. The timeline, closely examined, is eye-opening. Berners-Lee first posted his code for HTML to a fairly obscure Usenet newsgroup – alt.hypertext – in 1989. Although the Internet dates back to the Advanced Research Projects Agency Network (ARPANET) of the late 1960’s, the Web is, in
reality, only a product of this decade. It was not until 1993 that a truly effective graphical Web browser, Mosaic, first appeared. The popular Netscape browser was developed shortly thereafter, in 1994. Netscape Communications “went public” in 1995, the same year the Internet backbone structure became privatized as the NSF (National Science Foundation) stepped unobtrusively into the background. The content of the Web became dominantly commercialized with “.com” domains as late as two years ago and mass public acceptance and usage began to follow. These recent developments have led to planning discussions for Internet2 and Next Generation Internet (NGI) (Kaplan 1997). The accelerated pace of development has been simply astounding. Time has literally compressed to the point where a typical “Web year” is now conceptualized as roughly two to three months at a maximum.

It would be a critical mistake to think that the Web will remain the same, or even that the Web, in its current form, will continue to dominate the Internet. Already a compromise markup language solution between SGML and its subset HTML has been pushed in certain quarters. This compromise, called eXtensible Markup Language (XML), amounts to “SGML lite” with richer metadata possibilities than HTML without the intrinsic density of SGML (Khare & Rifkin 1997). The point is, everything mutates eventually, and, as our information sources and forms evolve, we must remain flexible in our ideas and in our approaches to the basic principles of information retrieval.

Fortunately, people have shown a great propensity for talking to each other in this globally-networked medium. Indeed, the Internet may well be the greatest human experiment in history, existing worldwide as it does without strict, formal organizational structure. There are certainly a great many players involved in the specific metadata question: librarians, computer scientists, publishers, commercial indexers, software vendors, markup language people, standards committees, and others. The semi-serious post from Roy Tennant’s Web4Lib electronic forum which began this paper was, after all, merely one voice in the middle of an extremely long and extended thread of discussion. Much is still to be said. As usual, we must keep steadily working while the conversation continues to hum.

References


Troncale, Anthony. [atroncale@nypl.org]. 1995. "REPLY: Cataloging the Internet," in Web4Lib, [web4lib@library.berkeley.edu], 17 October 1995.


APPENDIX A

<html>
<head>
<title>A LIBRARIANS LAIR</title>
<meta name="DC.title" content="A Librarians Lair - Homepage of Pradeep Lele"/>
<meta name="DC.subject" content="Librarians, Bombay, Library Science, Reference, web, Internet relay chat, IRC, University of Missouri, organ transplants, population, dictionary"/>
<meta name="DC.description" content="A personal homepage by Pradeep Lele - a librarian with links to a collection of Librarians pages on the web, some links to reference materials, bridge - the card game, links to IRC faq's, and organ transplant sites."/>
<meta name="DC.creator" content="Lele, Pradeep"/>
<meta name="DC.creator" content="(TYPE=email) Lelepu@hal.lamar.edu"/>
<meta name="keywords" content="Bombay, University of Bombay, India, Library Science, University of Missouri, Librarians Web Pages, IRC faq, Bridge, Organ Transplants, Kidney transplants"/>
</head>
<body>
The body of the document goes here
</body>
</html>
APPENDIX B

<HEAD>
<TITLE>Amortization and Coupon Book Software for Windows</TITLE>

<META name="description" content="Loan amortization and coupon/payment book software that the competition would not want you to see">

<META name="keywords" content="amortization, banking, bank, annuity, sinking fund, interest, amortization, amortizer, time value, present value, future value, loan calculator, loan, calculator, formulas, mutual funds, mutual, fund, coupon books, payment, stocks, bond, stock, bonds, free, download, money, banking, Windows, Win31, Win95, mortgage, mortgages, designer, home, homes, loan, loans, amortization, calculate, table, tables, principal, interest, equity, payment, repayment, payments, calculator, finance, refinance, cost, equity, tax, payment, payments, adjustable, ARM, variable, VRM, rate, qualification, analysis, balloon, term, compound, compounding, qualify, house, purchase, buy, buyer, consumer, percent, percentage, APR, real, estate, loan servicing, loan accounting, loan management, servicing, loan accounting software, finance company, financial, windows, loan administration, client, server, accounting, servicing, mortgage, lending, loans, micro, loan, management, loan servicing, finance, loan accounting, loans, software, loan servicing, Asset, Aviation, Bank, Bankers, Borrow, Borrower, Borrowers, Borrowing, Collect, Collection, Collecting, Commerce, Commercial, Credit, Energy, Equipment, Estate, Finance, Financing, Heavy, Lend, Lenders, Lending, Loan, Loaner, Loaners, Loans, Loaning, Market, Mortgage, Mortgager, Mortgagers, Mortgaging, Public, Real, Secure, Secured, Securing, Service, Services, Servicing, Secondary, Union, accounting, audit, trail, agings, loan, portfolio, disk, disks, diskette, diskettes, duplicate, download, free, loan receivable">

</HEAD>