AN ASSESSMENT OF THE EFFECTIVENESS OF METADATA AS A TOOL FOR ELECTRONIC RESOURCE DISCOVERY

by
Shauna L. Stephenson

A Master's paper submitted to the faculty
of the School and Information and Library Science of
the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science

Chapel Hill, North Carolina
April, 1999

Approved by:

_____
Advisor

Shauna Stephenson. An Assessment of the Effectiveness of Metadata as a Tool for Electronic Resource Discovery. A Master's paper for the M.S. in L.S. degree. April, 1999. 72 pages. Advisor: Jerry D. Saye.

This study assesses the effectiveness of the United States Environmental Protection Agency's (EPA) metadata system. Since information seeking over the Internet has become a complex and sometimes arduous task, many organizations are looking toward the use of metadata to enhance the discovery of electronic resources. Metadata are data elements, which are used to describe electronic resources to facilitate their later discovery over the Internet. Although there has been great interest among information professionals and academics about metadata, limited focus has been placed on its actual effectiveness. Twenty-four real reference questions for known items were reformulated into search queries which were run in the EPA's Public Access Website and Web Inventory (metadata) databases. Eight out of the 24 queries retrieved responsive metadata records. The results indicate the need for certain improvements to the system, including specificity in keyword assignment and more consistent generation of metadata by data owners.

Headings:

Information retrieval -- Evaluation

Information systems -- Cataloging

Information systems -- Evaluation

Metadata

Online searching

World-Wide-Web

Acknowledgments

I am extremely grateful to Dr. Jerry D. Saye for his suggestions, insight, and patience during the entirety of this project. It is rare to find a professor so personally committed to the success of his students. I also extend gracious thanks to Dr. Jane Greenberg and Kelly MagLaughlin, whose contributions related to metadata and information retrieval proved invaluable.

Finally, I extend a continued thanks to Shawn and Shirley, whose empathy, laughter, and support are always there when I need it the most.

Table of Contents

Chapter 1

INTRODUCTION

Statement of the Problem

The rapid growth of the Internet since the creation of the World Wide Web (WWW) and the graphical web browser has made searching the Internet a complex and sometimes arduous task. The lack of consistent searching mechanisms and quality control across today's search engines often makes searching the Internet an unreliable and unmanageable tool for resource discovery. Furthermore, the Internet's sheer size can make finding even the most ordinary document an overwhelming process, especially when a searcher is faced with sifting through hundreds or thousands of search results in response to a single search query. In the last several years, there have been numerous efforts to organize and facilitate the discovery of information contained on the Internet. Most of these efforts have been directed at solving some of the problems involved with the automated classification methods used by web crawlers, robots, and spiders to index the Internet. Because search engines perform their functions without much human intervention or systemization, their task of finding the best documents responsive to a specific query is often unwieldy and imprecise. In the mean time and until search engines can be restructured to provide consistent and reliable search results, most circles acknowledge the immediate need for an appropriate mechanism to enhance the discovery

of electronic resources over the Internet. Metadata has been proposed as one of the primary means to make this happen.

Metadata Defined

Metadata are descriptive sets of data elements (such as title, author, date, etc.), either embedded directly in the HEAD element of an HTML document or maintained in a separate database, which are used by organizations to 1) identify and describe their electronic information resources, and 2) to facilitate discovery and improved access to those information resources over the Internet. Metadata is designed to allow searchers to discriminate among similar electronic resources and determine the authenticity of the information they really want during an online search. Most importantly, metadata is designed to allow organizations to leverage their visibility on the Internet by making it easier for the general public to find information contained within their sites. Librarians and other information professionals are hoping that the use of metadata on organizational websites will increase the likelihood that individuals searching those sites will be able to locate the information they need without having to sift through extraneous and lengthy search results.

Purpose of the Study

There has been significant activity in the last few years dedicated to defining the semantic and syntactic aspects of metadata for the use of describing and facilitating access to Internet resources. A number of metadata initiatives have been proposed

including technological frameworks to support varying metadata schemes. Although there has been a great deal of interest in the development and deployment of metadata for organizational websites, with many systems operational and functioning, there has been limited focus placed on assessing the actual effectiveness of metadata as a means for resource discovery. In other words, although metadata is perceived by many to be a panacea for electronic resource discovery, classification, and organization, there has been little, if any, conclusive research on the ability of metadata to perform these functions. The paucity of such research is no doubt attributable to the burgeoning status of metadata and the few environments where metadata systems have actually been deployed. Moreover, a review of the literature has made clear that the effectiveness of metadata is often dependent upon the organization using it and the resource types being described.

This study will therefore assesses the effectiveness of metadata by examining the United States Environmental Protection Agency's (EPA) use of metadata as a tool for facilitating public access to Agency documents and information over the Internet. The Environmental Protection Agency was chosen as a venue for this research because of the author's affiliation with the Agency and because of the nature and comprehensiveness of metadata system.

Chapter 2

LITERATURE REVIEW

There is a consensus among librarians and information professionals that the current state of information seeking and retrieval over the Internet is in need of dire improvement. Although the Internet is easily accessible to millions of people, Ianella and Waugh (1997) have found that "the ability of those people to find relevant material has decreased dramatically as the quantity of information on the Internet grows." Neuss and Kent (1995) believe that "Because of its decentralized architecture, the user experiences the Web as a large information repository without an underlying structure." And although ranking and relevance feedback are available as features of commercial search engines, Hahn (1998) believes that "Few actual users of Web search engines understand how to manipulate and control a query to maximize the quality of their retrieval... Thus, despite the vast amount of information that is, in theory at least, accessible via the World Wide Web, most users still retrieve documents that have little or nothing to do with the topic of interest and fail to find the material most pertinent to them."

Neuss and Kent (1995) found that locating accurate and relevant information on the Internet cannot be accomplished by browsing with search engines alone. This is because search engines have very few means of distinguishing between relevant and

incidental words in document texts (Cathro, 1997). Dempsey and Heery (1998) note that, “The web crawlers currently operate at a very fine-grained level: they see a world of pages.” Because these automated tools categorize information differently than people do, relying on this machine-generated metadata often leads to imprecise descriptions of Internet resources, producing poor search results (Lynch, 1997). With respect to search engines, Lancaster (1998) and other individuals who have worked in the field of information retrieval for thirty years or more, garner little enthusiasm “over tools that routinely retrieve thousands of items for even highly specific searches… Presumably many of these will be completely irrelevant, of very low quality, or redundant.” Looking toward a solution, Warwick Cathro (1997) asserts that

> If we could target our searches onto words which are used as significant terms, we could achieve an enormous improvement in precision... we could retrieve just those resources where "Green" is the name of the author, without retrieving resources about green peas or environmental issues.

Finally, MacLennan (1998) is convinced that in terms of finding information over the Internet, “The best hook we can put in, at the moment, and probably for the foreseeable

There are many ways to define metadata as it can serve many functions and can be generated and maintained in an array of distributed environments. According to Lange and Winkler (1997), “metadata are data about data, or data elements used to describe or represent electronic resources… The primary function of metadata is to aid a user in locating desired and relevant data... it should be simple, but expansive, and should assist searchers in locating and accessing a resource.” Madsen (1994) believes the use of

metadata allows a person "to identify data which may satisfy the requirements of the user, and to store information about its location, content, and quality relative to the interests and situation of the user." Milstead and Feldman (1999) astutely recognize that searching today is largely a matter of matching query words to the text of desired documents and that "metadata is crucial to searching" since it can standardize indexing to greatly improve the matching process. Efthimiadis (1997) follows with the idea that

> Metadata can enhance the probability that a pertinent resource will be retrieved, provide a clearer overview of a subject area and improve the user's ability to discriminate among similar resources... it provides a user (human or machine) with a means to discover that the resource exists and how it might be obtained or accessed.

Many consider metadata to have grown out of the traditional catalog card with the technical information necessary to describe electronic resources. Iannella and Waugh (1997) believe that the traditional library catalog is, in effect, metadata that is used to find books and journals. And, oddly enough, although document description has generally been within the realm of catalogers, Larsgaard (1996) finds that "It is ironic that information derived by cataloging had to be called something else - metadata - before noncatalogers dealt with it." Milstead and Feldman (1999) further agree that "All of the reasons why indexing and cataloging are needed for print resources apply even more emphatically to metadata for electronic documents."

Metadata can take the form of an index or template, and is not limited to describing just documents; any resource, e.g., video, images, and audio, may also be described with metadata (Hudgins-Bonafield, 1995; Iannella & Waugh, 1997). And,

according to Lide (1995), “The bottom line is that metadata... is crucial to the use of almost every data set and must be included in any archiving plan.”

Metadata is not a new concept to information organizations or libraries. The term “metadata” was first used in relation to database management systems in the early 1980’s (Lange & Winkler, 1997). According to Inmon (1996), “Metadata has been a part of the information processing milieu for as long as there have been programs and data.” In fact, because the development of varying metadata typologies are still in a very nascent stage of development and the growth of the Internet so rapid, no single metadata standard has emerged to describe and manage all electronic resources across all platforms. Instead, several competing metadata schemes have been developed, with different levels of complexity and richness, which correspond to the types of resources they are describing. Milstead and Feldman (1999) recognize that just as different levels of cataloging are used, “different levels of metadata are needed, depending on the type of object and the use for which it is intended.” Finally, although the creators and proponents of individual metadata schemes are quick to show their allegiance to one metadata typology versus another, according to Dempsey and Heery (1998), “it is inevitable that many of the diverse approaches [to metadata] will continue to exist, and new formats will be created to respond to new user communities and market opportunities.”

Dempsey and Heery (1998) have identified three distinct categories of metadata formats, which they refer to as Band one, Band two, and Band three. Band one includes simple, unstructured proprietary data automatically extracted from resources by search

engines and web crawlers such as Yahoo, Lycos, and Alta Vista. The metadata created by these services is limited, and generally does not allow users to make relevance judgments in advance of actually retrieving the resource. Band two is based on emerging standards such as the Dublin Core, and contains structured data and descriptive attributes to support fielded searching. Typically, Band two metadata is generated manually by non-specialist users. Finally, Band three includes the rich and more elaborate formats such as MARC and the Encoding Archive Description (EAD), which are primarily used for scholarly or research oriented collections. Band three requires a specialist to create and maintain the data.

The most popular and well-known metadata scheme used today is no doubt the Dublin Core metadata set. The Dublin Core was created in 1995 at a workshop convened by OCLC and the National Center for Supercomputing Applications (Oder, 1998). The current metadata set, finalized in December 1996, consists of 15 elements (such as title, creator, subject, etc.) which can be embedded in the HEAD section of an HTML document. The Dublin Core was designed to be a simple and flexible data element set that could be created by non-catalogers to facilitate discovery and access to electronic resources in a networked environment (Caplan & Guenther). Although the Dublin Core looks at one aspect of metadata - simple description - the element set can also be extended to “enable more complex description for particular specialist domains, as well as to extend the types of resources described” (Dempsey & Heery, 1998). Many organizations have developed in-house variations of the Dublin Core. Milstead and Feldman (1999), however, agree that metadata cannot fully serve its purpose nor be of

any real value unless a common agreement or standard is reached on what elements to use and what content they should contain.

Metadata can be deployed in electronic documents in two ways. The first, and easiest way, is to embed the metadata descriptions into the HEAD portion of an HTML document by using the META tags. According to senior OCLC research scientist Stuart Weibel (1997), "The advantage of embedded metadata is that no additional system must be in place to use it; the metadata is integral to the resource and can be harvested by Web indexing agents." There is however, a downside to using embedded metadata that is often overlooked in the literature. If Weibel is referring to search engines and web crawlers as "Web indexing agents," than he and other researchers are assuming that these "agents" index the information contained in meta tags and metadata during the search and indexing process. According to Sullivan (1998) of Search Engine Watch, "Many believe that all search engines acknowledge keywords and descriptions placed in meta tags. In reality, only some do." As of this writing only three commercial search engines (Alta Vista, HotBot, and Infoseek) support and index meta elements contained within HTML documents (Sullivan, 1998). Much of the literature ignores this fact, and the fact that without the support of the commercial search engines, the use of meta tags and metadata will be ineffective as a means of improving access to electronic resources. And contrary to what is often stated in the literature, users are not yet "able to find material tagged with metadata by using their favorite Web search engine" (Griffen and Wason, 1997) nor has metadata yet to "increase the level of precision and recall for WWW search engines" (Iannella & Waugh, 1997). Many proponents of metadata are hinging its success upon

the ability of search engines to recognize metadata elements. It seems imprudent, however, to extol the virtues of embedded metadata when the support of major search engines has not yet been established.

A review of the literature has revealed that a second, and perhaps more manageable way of deploying metadata is to create a database that collects and manages metadata records (Weibel, 1997). Here, the metadata is not embedded in the resource it describes, but is instead generated by the document owners themselves and stored separately in a web database system, separately from the resource it describes. This concept is sometimes referred to as "data warehousing." This deployment method is often used to support more complex, domain specific document collections, such as those comprising Dempsey and Heery's (1998) Band three format. Finally, these metadata systems are often used in conjunction with a customized search engine to optimize resource discovery over the Internet.

Although online searching vis à vis metadata will be evaluated at greater length in the Methodology and Results chapters of this study, the literature has a definite opinion about the use of metadata in conjunction with online searching. Similar to descriptive indexing, metadata, if used, should be well chosen and flexible enough to accurately describe the central idea or topic of a document (Milstead & Feldman, 1999). Just as in any indexing exercise, omitting key words, specific topics, or concepts from a descriptive metadata record will most likely result in that document not being retrieved during an online search. According to Lancaster (1994), "no variations in searching strategy will ever be able to compensate for lack of specificity in indexing." And although controlled

vocabularies can increase consistency when searching across a set of documents, creating metadata by using a pre-defined list of controlled vocabulary terms, which do not accurately capture the essence of the document being described, can result in search results exhibiting poor precision during online searching and retrieval. Ideally, metadata should increase the probability that a document containing descriptive metadata responsive to a particular information need, will receive a higher ranking than a record not containing metadata during an online search. Milstead and Feldman (1999) amplify this fact by observing that, "The metadata, if well chosen, should describe the central topics of a document. Thus it should be given a high weight, relative to the appearance of those terms in the full text of the document. Any document having the query term in its metadata should appear quite high on the ranked list of search results." These statements recognize the fact that metadata must be "good metadata" in order for it to enhance resource discovery during online search and retrieval activities.

Trial implementations of the Dublin Core and other metadata schemes are currently underway in many library and information centers around the world. According to Qin and Wesley (1998), as of April 1998, there were over forty projects in more than ten countries that are using either the Dublin Core proposed standard or a similar scheme based on it. As previously stated, however, a review of the literature has revealed that few organizations have conducted research to assess the actual effectiveness of metadata as a method for facilitating improved access to electronic information.

The Consortium for the Computer Interchange of Museum Information (CIMI) is one organization that is exploring certain assumptions that have been made about

metadata and, more specifically, the Dublin Core within a museum environment. CIMI has developed a Dublin Core Metadata Testbed Project, now in Phase II, to test certain fundamental assumptions about the Dublin Core against a null hypothesis, which would suggest that the Dublin Core is unfit for the purpose of facilitating the discovery and retrieval of resources in a networked museum environment ("CIMI metadata testbed," 1998). The parameters for defining "fitness of purpose" will be defined according to CIMI's needs, because as the CIMI researchers point out, "the purpose for which Dublin Core may or may not be a fit is likely to vary from institution to institution" (id.). The researchers at CIMI believe that the success or failure of the Dublin Core rests on a number of preconceived assumptions, which lie at the heart of the Dublin Core, and that "have largely been accepted rather than questioned or tested in any

id.). The CIMI researchers are hoping that the CIMI Dublin Core Metadata Testbed Project will provide them with an opportunity to explore the effectiveness of the Dublin Core as a means for improving access to electronic resources within a museum environment. They are also hoping the Project will enable them to prepare a "Guide to Best Practice," which will provide recommendations for implementing the Dublin Core across other networked museum environments ("CIMI Dublin Core metadata testbed phase II," 1999).

The Nordic Metadata Project, Stage I of which was completed in June of 1998, also explored the effectiveness of Dublin Core metadata as means of improving electronic resource discovery of Nordic collections. According to Juha Hakala (1998), the Project's manager, "The emergence of the Internet as an important IR tool has also

fostered general awareness of serious problems associated with Internet information retrieval, of which massive recall, coupled with an equal lack of precision, is arguably the worst one." Thus, the Project was an attempt to improve the discovery, indexing, and retrieval of digital resources within Scandinavia through the use of metadata, a *metadata aware search service*, and the creation of an enhanced Nordic Web Index database to recognize, extract, and index metadata embedded in the HEAD portion of HTML documents (id.). One of the main project goals was to successfully enhance existing Dublin Core metadata specifications to create structured resource descriptions for Nordic classification. According to Hakala (1998), "The Nordic Web Index is… still the only major web index in the world, which is fully metadata aware and compliant to Dublin Core."

Finally, a model metadata system was developed as part of the Leicester University (U.K.) Metadata Project to test whether the design functionality of the system's Integrated Metadata Processor (IMP) is fit for the purpose of identifying stored electronic data which meets the requirements of a user's query (Madsen, et al., 1994). Discovering whether the IMP contains any topics relevant to a particular query was accomplished by presenting the IMP with a query to which related topics are known to exist within the IMP system. The researchers intended to use a variety of methodologies for matching topics to metadata queries including word-counting, string-matching, as well as nonparametric statistical methods such as cluster analysis. The most important features of the IMP system would be performance, fast indexing, and the ability of the system to establish an appropriate context for the user's query while then locating the

proper response within that context. From this project, the researchers at Leicester University hope to provide the framework for the construction of global, integrated, metadata information systems, which are simply designed and flexible in use.

Chapter 3

METADATA USE BY THE ENVIRONMENTAL PROTECTION AGENCY

Although the literature makes clear that the use of metadata can provide a viable means for enhancing and facilitating access to electronic information resources over the Internet, it seems logical that one must first assess the actual effectiveness of such a system before posing recommendations for its use. The purpose of this study is to assess the use and effectiveness of the United States Environmental Protection Agency's (EPA) metadata system.

Although it is part of a future plan, the EPA has not yet conducted any research on the effectiveness of its metadata system as a means of improving public access to its information resources. The databases are still under development and the overall system too new to justify a comprehensive study of this type at this point in time.

General Background

The development of metadata at the EPA was begun in 1995 by EPA's chief administrator, Carol Browner, who desired a system that would make it easier for the general public to find information on the Agency's Public Access Website. She

recognized that in 1995, the searching and browsing tools put into place to support the Agency's initial web efforts, would not be able to keep up with EPA's rapidly expanding information system. Thus, in the summer of 1997, an Agency workgroup developed a set of metadata elements to describe the information resources that the EPA makes available through its Public Access Website. The EPA's goal for implementing metadata, according to an Agency employee, is that it will allow the Agency to "read the public's mind and provide them with what they want… useful and relevant documents, with the most important documents retrieved from a search to be returned first" (L. Smith, personal communication, November 12, 1998).

Managing information contained on EPA's Public Access Website, is the joint responsibility of the data owners, or authors, of that information as well as the Office of Information Resources Management (OIRM), which is located in Research Triangle Park, North Carolina. Although the OIRM maintains primary responsibility for overseeing the metadata development process, it shares this responsibility with EPA's Environmental Information Management Division (EIMD), which handles the content and interface issues associated with metadata development. To provide sufficient background on the development and deployment of metadata at the EPA, an interview was conducted with an EIMD employee on November 12, 1998. From the interview, it was learned that metadata was implemented at the EPA to serve two functions: 1) to manage all of the environmental information resources contained on EPA's Public Access Website, and 2) to facilitate public retrieval of relevant information in a distributed environment (L. Smith, personal communication, November 12, 1998).

Following the initial interview, e-mail correspondence with the EIMD employee was maintained for the purpose of clarifying subsequent issues and findings raised during the course of the research. The follow-up questions clarified several issues regarding the mechanics of the Verity search engine and its performance in searching certain fields, the Agency's use of a keyword controlled vocabulary, and issues surrounding the structure of the Web Inventory Database.

The Web Inventory Application and Database

EPA began its implementation of metadata by creating two tools, a "Web Inventory Application" and a "Web Inventory Database," which offer a disciplined and automated approach to website management and metadata generation. The Web Inventory Application allows data owners to generate their own metadata records to describe items being added to the Agency web environment. The Web Inventory Application is flexible enough to support the creation of general metadata records as well as specialized metadata to support more elaborate and complex documents.

Once generated, the metadata records are stored in a series of relational tables inside an Oracle RDBMS database, which is known as the "Web Inventory Database." In addition to the metadata records, the Web Inventory Database includes all other content that is made available to the public through EPA's cluster of web servers.

EPA's Web Inventory Application and Database are only available on the Agency's Intranet, to which the general public does not have access. Furthermore, since EPA's metadata resides separately from the resources it describes, no metadata is evident

or viewable in the source code of documents retrieved by the general public through EPA's Public Access Website.

Generating Metadata

Owners and authors of Agency information are responsible for generating the metadata records for documents destined for EPA's Public Access Website. Each EPA metadata record template contains 28 descriptive metadata elements consisting of three key elements: 1) the "Metadata Core List of Fields," 2) "Optional Fields," and 3) keyword and geographical keywords to further enhance document description. A copy of an EPA metadata record template is included as Appendix A.

The number of metadata fields a data owner opts to include to describe a particular document depends upon the complexity of the document being described, its ultimate destination, and the personal wishes of that data owner. Of the 28 metadata fields available to describe a document resource, 13 fields are considered mandatory, or "core," for documents that are to be made available to the public through EPA's Public Access website. The 13 mandatory fields include:

- Title
- Description
- Organizational Author (Level 1)
- Document Date
- Entry Type
- URL
- Approving Manager
- Internet Contact
- Legal Authority (if applicable)
- General Keyword Entry (Broadest, More Specific, Most Specific, Open Keywords)
- Geographic Keyword Entry fields

In theory, these core fields are designed to allow data owners to highlight the most important and/or significant aspects of their documents to enhance their subsequent retrieval by the public over the Internet.

Keyword Selection

A critical aspect of metadata generation at the EPA is the data owner's decision to include keyword data in their metadata records. Data owners can include keywords by either choosing words from an official EPA Keyword List of controlled vocabulary terms or by choosing their own "additional" keywords to describe concepts not captured by the controlled vocabulary. EPA's controlled vocabulary is designed to include those topics that are most representative of the content of the documents data owners will most likely be describing while also promoting consistency in keyword assignment to enhance retrieval during online searching. A copy of the 3-Level Hierarchy of Keywords is included as Appendix B.

Data owners choosing to use EPA's official keyword list of controlled vocabulary, must select terms from three hierarchical levels of keyword specificity, "Broadest," "More Specific," or "Most Specific," to describe their documents. The "Broadest" category contains the concepts which allow data owners to broadly describe their documents. The "More Specific" category contains "terms" which provide more detail about the main concepts expressed in a document. Finally, the "Most Specific" category contains "keywords," which provide the highest level of specificity available to data owners to describe their documents. When generating keyword metadata terms,

data owners must adhere to the strict hierarchy of specificity and cannot access the lower, more specific keyword levels until terms are selected from the “Broadest” level first.

Finally, in addition to the EPA controlled vocabulary keywords, data owners may enter open or “additional keywords” to further describe their documents. This is appropriate for those documents that cannot adequately be described by the general EPA keyword controlled vocabulary alone. In particular, the inclusion of additional keywords should allow data owners to differentiate their documents from similarly described documents by assisting in the natural language or free-text retrieval of their documents over the Internet.

File Diversity

Unlike organizations that use an embedded metadata scheme, the EPA has chosen to store its metadata in a separate Web Inventory Database, apart from the resources actually described. This decision was made to facilitate public access to the diverse file types held by the EPA. For organizations with distributed file types, embedded metadata is considered too restrictive because its use is limited to collections only containing HTML file types. In addition to its numerous HTML files, EPA’s website contains 15 other file types available for public downloading that, by design, cannot support embedded metadata. These files include Zipped (compressed) files, Dbase Files, Word Perfect files, ASCII Text files, PDF (Adobe Acrobat) files, PC Executable files, Computer Graphics Metafiles, PC Executable modules, Graphics Interchange Format

(gif) files, Microsoft Access and Excel files, Microsoft Power Point and Word files, Freelance Graphics files, and Lotus 1-2-3 Worksheet files. A description of these file types is included as Appendix C.

The Verity-97 Search Engine

Because of the diverse file types existing within EPA's web environment, the EPA has implemented a metadata-aware Verity97 search engine (Verity), which runs on Digital UNIX. Verity was designed to provide free-text searching and indexing of every word in the Web Inventory Database in an effort to produce more relevant search results. Verity does this by indexing and searching the relational tables comprising EPA's Web Inventory Database, which include the free text and <title> tag of EPA's HTML files, which are stored with the metadata records, albeit in separate structures, inside EPA's Web Inventory Database.

The only metatag supported by the Verity search engine is the <title> tag found in HTML files. It is extremely important to clarify here that although Verity searches the HTML <title> tag, which is embedded in all HTML documents, Verity's prioritization algorithm currently does not place additional weight on words found in the <title> tags or in the early sentences of a document. According to the EIMD, this weakens Verity's ability to effectively and correctly prioritize documents that are searched by known title (L. Smith, personal communication, January 14, 1999). Because of this feature, the searching conducted for this study was performed in free-text mode to avoid missing

potentially responsive documents, which would otherwise not be retrieved during a search by title alone.

According to the EIMD, one of the Agency's biggest problems with Verity is its ability to search the Agency's PDF files, which comprise approximately 50% of EPA's total files (L. Smith, personal communication, January 14, 1999). Verity stores information inside its index in "zones" and "fields." When Verity searches its "zones," the prioritization algorithm is executed quickly and, at least, theoretically, the most responsive documents are pushed to the top of the list of documents retrieved. HTML metatags are contained in Verity's search "zones," and are generally the first to be indexed and returned by Verity during a search query. Unfortunately, Verity stores PDF files in "fields," where the prioritization algorithm is slow and no title ranking mechanism exists. Since PDF title searching can only be performed by utilizing a "field" search, PDF files are often returned in random order with no priority or ranking mechanism ascribed to the retrieved document set. This has left the EPA with two choices when searching for Agency documents, including the PDF files: 1) to either using a title search and miss the PDF files or 2) to not use a title search and have documents with less prioritization returned (L. Smith, personal communication, January 14, 1999). The EIMD does acknowledge however, that overall, Verity's searching capability has improved with the inclusion of metadata in the Web Inventory database.

Support

Optimally, every EPA employee should be generating metadata records for any information they create, which is destined for EPA's Public Access Website. The metadata records are designed to provide adequate descriptions of available documents that are sufficient to facilitate their later discovery over the Internet. Thus, the metadata records contained in the Web Inventory Database should coincide with their electronic counterparts made available through EPA's Public Access website. This however, requires the full support of Agency staff. According to the EIMD employee, creating metadata for their documents is a "big cultural change" for EPA employees, and there is currently a great difficulty in building enough consensus among employees to assure that everyone generates a metadata record for items destined for EPA's Public Access Website (L. Smith, personal communication, February 7, 1999). Because OIRM and EIMD have not yet gained the full support of Agency employees in this effort, for the time being, library staff are generating metadata for the top level navigational pages so as to assure public access to the most important documents appearing on the EPA's Public Access Website.

Chapter 4

METHODOLOGY

The public searches EPA's Public Access Website to satisfy various information needs. Primarily they search for "known" titles or "known" subjects, which are those resources that the requester knows to be responsive at the time of his/her request and that comprise those titles and/or subjects which exist in electronic form on EPA's Public Access Website. It has been this author's experience that public patrons searching EPA's Public Access Website often have trouble locating known items, and are often left with search results that are wholly unresponsive to their initial request. If it is EPA's goal that metadata will help the public find what they are looking for on EPA's Public Access Website, it is logical, therefore, to assess the effectiveness of EPA's metadata system by evaluating whether or not it is successful in satisfying real information needs.

Test Construction

To conduct this research, twenty-four search queries were formulated from real reference questions posed by public patrons to the staff of the EPA's Air Information Center library, located in Research Triangle Park, North Carolina, during October through December 1998. These particular reference questions were chosen for use in this

study because they 1) represented known items, 2) were representative of the diverse types of reference questions commonly posed by public patrons, and 3) would provide enough data to complete a meaningful assessment of EPA's metadata system. Finally, it was felt that using "actual" reference requests submitted by public patrons would remove the artificiality of using hypothetical search requests and would add to the "reality" of the research since each request represented an actual information need.

Procedure

The reference questions were reformulated into 24 search queries that reasonably approximated the patrons' original information requests. The search queries were run in two databases, the Public Access Website and the Web Inventory Database, which contains the Agency's metadata records, during the time period of January 15 through February 15, 1999. To optimize the free-text searching capabilities of EPA's search engine and to promote consistency across search queries, the search queries were constructed by selecting terms from the title or the subject matter of the request, separating those terms with commas, and selecting the search option, "and," which requires that all words in the search query must be contained in the documents retrieved. This method of searching EPA's Public Access Website and the Web Inventory Database has proven to be very successful in the past, mainly due to the limitations Verity places on searches that are field delimited. A copy of a chart illustrating the 24 reference questions and search queries is included as Appendix D.

Although the Public Access Website and Web Inventory database allow for more complex searching by specifying a combination of field names, comparison operators, and search strings, this was avoided to reduce searching bias and to preserve the integrity of each search query across both databases.

Additionally, although consideration was given to searching the queries by delimiting them to particular fields, it was learned from EIMD that the Verity search engine's prioritization algorithm currently does not give weight to terms found in the title field, nor does it have the ability to search the Agency's PDF files by title. Title searching by field was therefore not a viable options since over 50% of the files on EPA's Public Access website are PDF files, which cannot be retrieved by a search delimited to the title field alone.

Finally, the first 30 records displayed from the Public Access Website and the first 20 records from the Web Inventory Database were examined for responsive documents, with the hit position and Verity generated ranking of the *first* responsive document noted. If no responsive documents were retrieved within the first 20 or 30 records, an examination of the succeeding records, through records 200, was conducted.

Evaluation Criteria

The goal of patrons searching EPA's Public Access Website is to effectively locate useful and pertinent items while minimizing the possibility of retrieving items that are useless. Since there continues to be much debate and disagreement in the literature as to what the terms "pertinent," "useful," and "relevant" really mean, for this research this

author will use Lancaster's (1994, 1998) definitions of pertinent, relevant, and useful to assign a uniform meaning to these terms. In terms of the satisfaction of some information need, Lancaster considers the expressions "useful," "pertinent," and "relevant" as synonymous in that "a pertinent (useful) item is one that contributes to the satisfaction of some information need." In other words, a useful, pertinent, or relevant document is one that is "responsive" to a user's particular information need. When evaluating the performance of an entire retrieval system, Lancaster (1994) believes that a "relevant document is nothing more nor less than a document of some value to the user in relation to the information need that prompted his request." The problem according to Lancaster "is to retrieve as many as possible of the useful items and as few as possible of the useless ones." Therefore, to avoid any further ambiguity as to what constitutes relevance, the word "responsive" will be used in this research to describe all documents that are pertinent, useful or relevant to a user's information need.

Following Lancaster, it is appropriate for this research that one of the parameters used to assess the effectiveness of metadata be the calculation of a *relative* precision ratio for the responsive items retrieved as a result of a search query. According to Lancaster (1998), "the ratio of useful items to total items retrieved is usually referred to as a precision ratio," which illustrates a system's ability to hold back useless or nonresponsive documents and retrieve useful ones. For this research, a *relative* precision ration was calculated since only the first 30 website records and the first 20 metadata records were reviewed. It was hoped in this research that the existence of metadata would increase the precision ratio of documents retrieved during an online search.

A determination of recall, the other commonly used criterion for evaluating information retrieval performance, was omitted from this research since it is virtually impossible to determine how many potentially relevant items could exist in EPA's database. Another factor prohibiting a determination of recall is the fact that EPA's public access website is configured to show only the first 200 hits retrieved for any search query. In other words, a search query retrieving 714 hits would only permit the first 200 hits to be evaluated, with the remaining 514 unavailable for review and evaluation.

The relative precision ratio for each search query was calculated by dividing the number of responsive items retrieved by the total number evaluated. Therefore a search of the Metadata Web Inventory producing 7 useful documents would be divided by 20 (the total number of hits evaluated) to yield a relative precision ratio of 7/20 or 35%.

Four additional categories of information were examined during this research, which include an evaluation of the hit position and Verity generated rankings for the first responsive documents for a given search query, an evaluation of the use of keywords by data owners, an evaluation of the artificial extracts generated by Verity for the Agency's PDF files, and an evaluation of the system's retrieval of documents responsive to *other* search queries.

Limitations

There have been few studies to date, which have assessed the effectiveness of metadata as a tool for resource discovery. This can be attributed to the fact that most organizations that have implemented metadata, have done so recently, and are still

working out the technical problems commonly associated with new information systems. The Environmental Protection Agency is no exception. One of the most challenging aspects of this research was trying to observe and assess a novel metadata system that was continuously subjected to change and modification during the study by its designers. Even seeking clarification from Agency employees about the nuances of the system was a formidable task, since few employees were actually able to keep up with the daily changes themselves. Finally, the complexities of a constantly changing system made data analysis difficult since all of the data collected had to be considered against the system as it was when the data was collected, and not as it was when the data was actually analyzed.

Therefore, given the time limitations for the completion of this research, this study only assesses those aspects of the EPA's metadata system, which were operational during the time period of January and February 1999. The author acknowledges that the EPA's metadata system has undergone numerous changes and enhancements since this study was originally conducted and that additional changes will continued to be made by the Agency for quite some time into the future. Therefore, although unique and contributory, the findings and conclusions expressed here are limited to the time period as stated and may not reflect present-day applicability to certain augmented features of EPA's metadata system.

Chapter 5

RESULTS

The ranked search results for each of the 24 search queries set forth in Appendix D were assessed. For the search queries run in the Web Inventory Database, individual metadata records for responsive documents were downloaded in addition to the ranked search results.

Overview of Results

Overall, the search results were rather lackluster as only seven search queries retrieved responsive documents from both the Public Access Website and the Web Inventory Database for the 24 given queries. In other words, only seven search queries retrieving responsive documents from the Public Access Website had a corresponding metadata record in the Web Inventory Database. As shown in Table 1, these seven search queries were search queries #1, 10, 16, 20, 22, 23, and 24. Fifteen queries retrieved responsive documents from the Public Access Website search, but failed to retrieve corresponding metadata records from the Web Inventory Database search.

Table 1
Overview of Search Results

Search Query#	Retrieved Responsive Public & Metadata Records	Retrieved No Responsive Records	Retrieved Responsive Metadata Records Only	Retrieved Responsive Public Records Only	Metadata Records Containing "Additional Keywords"	Search Queries that Retrieved Public Records Responsive to "Other" Search Queries
1	X				X	X
2				X		
3			X		X	
4				X		
5				X		
6				X		X
7				X		X
8				X		
9				X		
10	X					X
11				X		X
12				X		X
13		X				X
14				X		
15				X		
16	X					
17				X		
18				X		
19				X		
20	X					
21				X		
22	X					
23	X					
24	X					
Totals:	7	1	1	15	2	7

One query (search query #3) retrieved a responsive metadata record from the Web Inventory Database, but failed to retrieve a corresponding responsive document from the Public Access Website. Finally, one search query (search query #13) failed to retrieve

any responsive documents from either database. This search failure could be attributed to the complexity of the information sought by the reference request as well as the author's inability to accurately reformulate the patron's original reference request into an adequate search strategy. However, a review of the EPA's 3-Level Hierarchy of Keywords (attached hereto as Appendix B) indicates that the query term "furnace" exists as a keyword under the broad categories of Air – Indoor Air Pollution – Furnaces. If this keyword were selected during metadata generation for this document, it is likely that a document responsive to search query #13 would have been retrieved.

Relative Precision

Table 2 illustrates the relative precision ratios calculated for responsive items retrieved from both the Public Access Website search and the Web Inventory Database search. In terms of the systems' ability to hold back useless or non-responsive documents from the items displayed, a mean relative precision ratio of 17.2% was calculated for the Public Access Website search and 5.62% for the Web Inventory Database. The range of the relative precision ratios calculated for the Public Access Website searches ranged from a high 66.6% (20 out of 30 documents were deemed responsive) to 0.0% (no documents were deemed responsive). For the Web Inventory Database, the range of relative precision ratios ranged from a high 50% (10 out of 20 documents were deemed responsive) to 0.0% (no documents were deemed responsive). These results indicate, that in terms of the Public Access Database, the system retrieved many more documents than just responsive items, thereby forcing users to sift through

extraneous and non-responsive search results. This is the very thing that metadata was designed to abate.

Table 2
Relative Precision Ratio/Verity Ranking and Hit Number

Search Query#	Relative Precision Ratio (%)		Hit Number and Verity Ranking (0.00) of First Responsive Document	
	Public Access Website (30 documents)	Web Inventory Database (20 documents)	Public Access Website	Web Inventory Database
1	66.6 (20/30)	50.0 (10/20)	1 (1.00)	1 (0.92)
2	6.6 (2/30)	0.0 (0/20)	1 (1.00)	0 (0.0)
3	0.0 (0/30)	0.0 (0/20)	0 (0.0)	21 (0.40)
4	13.3 (4/30)	0.0 (0/20)	1 (1.00)	0 (0.0)
5	16.6 (6/30)	0.0 (0/20)	7 (1.00)	0 (0.0)
6	26.6 (8/30)	0.0 (0/20)	1 (1.00)	0 (0.0)
7	6.6 (2/30)	0.0 (0/20)	7 (1.00)	0 (0.0)
8	0.0 (0/30)	0.0 (0/20)	151 (0.98)	0 (0.0)
9	8.3 (2/24)[a]	0.0 (0/20)	1 (0.93)	0 (0.0)
10	33.3 (10/30)	20.0 (4/20)	1 (1.00)	1 (0.88)
11	13.3 (4/30)	0.0 (0/20)	19 (1.00)	0 (0.0)
12	0.0 (0/30)	0.0 (0/20)	39 (0.91)	0 (0.0)
13	0.0 (0/30)	0.0 (0/20)	0 (0.0)	0 (0.0)
14	26.6 (8/30)	0.0 (0/20)	1 (0.93)	0 (0.0)
15	13.3 (4/30)	0.0 (0/20)	1 (1.00)	0 (0.0)
16	23.3 (7/30)	5.0 (1/20)	1 (1.00)	1 (0.96)
17	56.6 (17/30)	0.0 (0/20)	1 (1.00)	0 (0.0)
18	3.3 (1/30)	0.0 (0/20)	27 (0.79)	0 (0.0)
19	16.6 (5/30)	0.0 (0/20)	1 (1.00)	0 (0.0)
20	0.0 (0/30)	5.0 (1/20)	36 (0.82)	2 (0.88)
21	20.0 (6/30)	0.0 (0/20)	1 (1.00)	0 (0.0)
22	33.3 (10/30)	25.0 (5/20)	2 (0.98)	1 (0.95)
23	26.6 (8/30)	15.0 (3/20)	1 (1.00)	1 (0.78)
24	3.33 (1/30)	15.0 (3/20)	1 (1.00)	1 (0.94)
Mean Relative Precision Ratio	17.2 (414.3/24)	5.62 (24/135)	N/A	N/A

[a] Search Query #9 retrieved a total of 24 records.

In terms of the Web Inventory Database, the results indicate an inherent lack of responsive documents indexed within the database itself. In other words, the documents sought simply were not there. Without the inclusion of responsive documents, no search strategy, however well formulated, would ever retrieve responsive documents from this database at the current time. Table 2 also illustrates the hit number and Verity generated ranking for the first responsive documents retrieved for *each* of the 24 search queries.

Hit Position and Verity Generated Rankings

Although the mean relative precision ratios for both databases are somewhat disappointing, the hit positions and Verity generated rankings for each search query tell quite a different story. A mean hit position of 4 and a mean Verity generated ranking of 0.98 was calculated for the first responsive documents retrieved from the Public Access Website. A mean hit position of 4, when the first 30 hits are displayed, is very good. This mean hit position can be attributed to the large number of responsive HTML files that appeared as the #1 hit for 12 out of the 24 search queries. When the mean hit position includes those results beyond the first 30 items displayed, the mean hit position of the first responsive document drops to 13, which is still not too bad. This drop is attributed to search queries #8, 12, and 20, where the first responsive document occurred at hits #151, 39, and 36 respectively. Likewise, of those 12 HTML files appearing as the #1 hit, 11 contained relevancy rankings of 1.00, thus boosting the overall rankings across the responsive documents retrieved from the Public Access Website search. These high rankings can be attributed to the fact that Verity can create a search zone from any

HTML metatag. Verity's most precise searching is conducted in zones, where the prioritization algorithm is configured to return the most relevant documents, if any, first. In addition, the HTML files are queried separately from the metadata files inside the Web Inventory Database, and are theoretically supposed to be the first files returned as a result of a search query. Table 3 illustrates the mean hit position and Verity generated ranking across all search queries in both databases.

Table 3
Mean Ranking and Hit Position
Across all Search Queries in Both Databases

Categories	**Mean Hit Position (n=1)**	**Mean Verity Generated Ranking (n=1.00)**
Public Access Website (Across Hits 1-30)	4	0.98
Public Access Website (Across all Hits)	13	0.89
Web Inventory Database (Across Hits 1-20)	1.14	0.90
Web Inventory Database (Across all Hits)	1.20	0.28

The calculation of the mean hit position and mean ranking across all hits for the Web Inventory Database search includes the 16 search queries for which no responsive documents were retrieved. This explains why the mean hit position, as illustrated by Table 3, remains a high 1.20 while the mean ranking receives a low 0.28. The high mean hit position of 1.20 is attributable to the fact that six out of eight search queries retrieving

responsive documents during the Web Inventory Database search, produced results where the first responsive document was also the number one hit. The low mean ranking of 0.28 across all hits can be attributed to the fact that 16 search queries retrieved no responsive documents at all.

Ranking Variation

As illustrated by Table 4, unlike their counterparts retrieved from the Public Access Website search, the seven search queries retrieving responsive metadata records from the Web Inventory Database search, received lower individual rankings and a lower mean average ranking of 0.90 compared to a mean ranking of 0.97 for the same 7 documents retrieved in the Public Access Website search.

Table 4
Variation in Verity Generated Rankings for Like Responsive Documents Retrieved from Web Inventory Database and Public Access Website

Search Query#	Public Access Website Ranking	Web Inventory Database Ranking
1	1.00	0.92
10	1.00	0.88
16	1.00	0.96
20	0.82	0.88
22	0.98	0.95
23	1.00	0.78
24	1.00	0.94
Average Ranking:	**0.97**	**0.90**

Although search query #3 retrieved responsive metadata records, it was not included in the calculations for Table 4 because this query did not retrieve any corresponding responsive documents during the Public Access Website search. Overall, Table 4 illustrates the weaknesses in Verity's ability to generate consistent rankings for like documents across databases.

Keyword Use

As indicated in Table 5, only two search queries retrieved responsive documents containing metadata records, which included open or "additional keywords."

Table 5
Open or "Additional Keywords" Appearing in Responsive Metadata Records

Search Query & Search Request	**Open or Additional Keywords**
#1 Wood Furniture Manufacturing NESHAP	"Wood Furniture Manufacturing"
#3 Cities in Non-Attainment Status	"Nonattainment"

Although many of EPA's documents are of the same subject matter or genre, an assessment of the metadata associated with the search queries for this study, revealed that few authors utilized the "additional keywords" field to differentiate and/or distinguish their documents from other like documents in the collection. Instead they relied upon EPA's Hierarchical keyword controlled vocabulary to describe their documents. This

occurred even though it is well known that the more specificity used to describe like categories of documents, the greater the probability that responsive documents would be located during an online search. Greater specificity also works to reduce recall while improving precision. According to Lancaster (1998), the best discriminators for an indexed collection of electronic resources "are those that are unexpected and rare in a collection." In terms of the EPA's collection, indexing an "air" document under the term "air" or "air emissions" in a database that contains primarily air documents is not as helpful as using a more specific term, in addition to "air" and "air emissions," to differentiate the document from the other "air" documents in the collection. Table 5 indicates how the very specific terms of "nonattainment" and "wood furniture manufacturing" were included by data owners as "additional keywords" to differentiate their documents from other air-related documents.

Keywords assigned during the metadata generation process can either aid or hinder a document's subsequent retrieval. Lancaster (1994) recognizes two types of indexing errors: "1) omission of a term necessary to describe an important topic discussed in an article, and 2) use of a term that appears inappropriate to the subject matter of the article." Omitting key topical terms will generally lead to recall failure while the use of inappropriate or incorrect terms will lead to precision failure (Lancaster, 1994). In terms of keyword assignment at the EPA, it appears that the use of broad descriptive terms leads to precision failures while increasing the overall recall of nonresponsive documents indexed with the identical keywords as the responsive documents.

Hierarchical Keyword Data

The EPA's hierarchical keyword data is designed to allow data owners to accurately and consistently describe their resources for their subsequent retrieval over the Internet. It is also designed to include those terms that are broadly representative of the content of the types of documents data owners are most likely to be describing.

Unfortunately, use of a controlled vocabulary, such as EPA's hierarchical keyword data can also reduce the specificity of indexing and/or not accurately describe the concepts contained in a document. Table 6 illustrates the hierarchical keyword data entered by data owners for the eight search queries which retrieved responsive documents from the Web Inventory Database search.

To demonstrate the lack of specificity practiced by data owners during metadata generation, Table 6 illustrates that five out of eight responsive documents were assigned the term "Air" as a top-level ("topic") keyword. In addition, five out of eight responsive documents were assigned the term "Air Pollutants" as a mid-level ("term") keyword. Finally, 2 out of 8 responsive documents were not assigned any most specific ("keyword") keywords at all.

Table 6
Hierarchical "EPA Keyword Data" Appearing in
Metadata Records Responsive to the Search Query

Search Query & Search Request	EPA Hierarchical Keyword Data Entered		
	Topic ("Broadest")	**Term ("More Specific")**	**Keyword ("most specific")**
#1 Wood Furniture Manufacturing NESHAP	Compliance and Enforcement Air Legislation Business and Industry	Settlements Air Pollutants Clean Air Act (CAA) Industries	NESHAPs Hazardous Air Pollutants MACTs
#3 Cities in Non-Attainment Status	Air Air	Air Pollutants Air Quality	Emission
#10 New Regulations for National VOC Emissions Standards for Architectural Coatings	Air Air Compliance and Enforcement Legislation	Air Pollutants Air Pollutants Clean Air Act (CAA)	Volatile Organic Compounds Ozone
#16 National Water Quality Inventory: 1996 Report to Congress	Water	Water Quality	Monitoring
#20 Safe Drinking Water Act – Reauthorization of 1996	Water Legislation Human Health	Drinking Water Safe Drinking Water Act	
#22 National VOC Emission Standards for Consumer Products – Automotive Refinishing	Air Air Automobile Repair Industry Ozone Trucks and Buses Volatile Organic Compounds (VOCs)	Air Environmental Protection Agency Ozone VOC Air Quality Automobiles and other Vehicles	
#23 1998 Interstate Ozone Transport Report	Air Air Air Government Legislation	Air Pollutants Air Quality Air Pollutants State Government Clean Air Act (CAA)	Nitrogen Oxides Emission
#24 Enabling Document for the New Source Performance Standards for Municipal Solid Waste Landfills	Air Wastes Air Air Wastes Government	Clean Air Act (CAA) Solid Wastes Air Pollutants Air Quality Landfills State Government	Municipal Solid Wastes Volatile Organic Compounds Emission

Several interesting things can be gleaned from Table 6. First of all, although the term "ozone" is included in the keyword data for search queries #10 and 22, it is not included as a term for search query #23, "1998 Interstate Ozone Transport Report."

Instead of ozone, the term “nitrogen oxides” is included as keyword data for search query #23. Perhaps the data owner felt that because the term “ozone” was already contained in the document’s title, no additional description was necessary. This contradicts, however, the responsive documents retrieved by search queries #16, 20, 22, and 24, where words from the title are included as keywords, e.g. search queries #16 and 20 contain “water” in the titles and “water” in the keyword data. It seems logical that specificity in indexing when like documents are dispersed throughout a collection is absolutely essential to enhancing their later discovery over the Internet.

Documents Retrieved Responsive to Other Search Queries

Search queries #1, 6-7, and 10-13, retrieved documents specifically responsive to other search queries within the first 30 items displayed. In addition, they were usually retrieved with a higher rate of precision than the more exact search query. For example, Hit #25 from search query #1 which sought documents related to the “Wood Furniture Manufacturing Operations NESHAP,” retrieved a PDF file with a relevancy ranking of 0.88, which was responsive to search query #12, which sought the document, “Handbook for Air Toxics.” This is extremely odd, since the more specific search query of “handbook, air, toxics” formulated for search query #12, did not retrieve responsive documents within the first 30 items displayed. Instead, a user would have to have kept scrolling through search results until he/she reached items 39 and 40, which consisted of two HTML files, with relevancy rankings of 0.91, responsive to the query.

In addition, search query #6 (which sought the "Quality Assurance Handbook for Air Pollution Measurement Systems") retrieved five PDF files responsive to search query #7, which sought the document "List of Designated Reference and Equivalent Methods." These files comprised hits #15, 19, 21, 23, and 24 of search query #6 (although responsive to search query #7) and were ranked 0.94, 0.94, 0.94, 0.94, and 0.93, respectively. Search query #7, on the other hand, retrieved only two documents responsive to the initial search query.

A partial explanation for these occurrences could be that the assignment of the same hierarchical keywords in the metadata records for each of the retrieved items increased the overall recall of items retrieved. In other words, if a document is indexed with the same keywords as another document, more items will be retrieved as potentially responsive to the query. Unfortunately, only one query, search query#1, also retrieved corresponding metadata records. A review of the metadata associated with search query #1 revealed several keywords related to "air" and "air pollutants." There were also responsive metadata records containing the term "Air Toxics" as part of the title terms.

It is difficult to ascertain the exact reason why these documents are being retrieved in the manner in which they are without being privy to the algorithms set for the Verity search engine. Perhaps future refinements to the search engine will resolve this problem.

Verity Generated Extracts for PDF Files

A major flaw in searching the EPA's Public Access Website, independent of any metadata issues, occurs during a user's screening of retrieved items. A severe

shortcoming of the Verity search engine is the artificial extracts it generates for the PDF files it retrieves. During the course of this research, over 450 document extracts were examined. For a majority of the PDF files retrieved, the document descriptions or "extracts" generated by Verity did not present an accurate representation of the true content of any given document. In other words, the extract itself does not allow users to sufficiently predict the responsiveness of a particular document based on a reading of the extract alone. Without adequate descriptions, users will most certainly fail to recognize topics that are of potential use and/or responsive to their initial query.

For example, search query #17 which seeks the "1997 Mercury Report to Congress" retrieved the following PDF and HTML file extracts, which are responsive to the search request and refer to the same document. Based on the content of the extracts, a reading of the two documents would cause most users to think the documents had little, if anything in common:

Hit #1	1.00 HTML	Mercury Study Report to Congress Summary: EPA's Report to Congress on Mercury is an eight-volume document. The eight volumes of the December 1997 report (EPA 452/R-97-0003) are also available for download below in Portable Document Format (PDF). The Report provides an assessment of the magnitud....
Hit #16	1.00 PDF	List of Tables (continues) Page xi 4-26 Daily Intake of Sportfish and Total Fish Summary: List of Tables (continued) page xi 4-26 Daily Intake of Sportfish and Total Fish for the Fish-consuming Portion of the Population studied by Fiore et al. (1985) in Support of Analyses of Freshwater Fish.....4-69 4-53 Fr

The HTML file was retrieved first as the #1 hit, with a relevancy ranking of 1.00. In fact, hits #1-13 were all HTML files with a 1.00 relevancy ranking, and provided

succinct and clear extracts of the resources described. Strangely enough, however, the hardly intelligible PDF file also received a relevancy ranking of 1.00, even though the extract would make any user question its potential responsiveness. Additionally, hits #15, 18, and 19 for this search query also contained PDF files, which were responsive to the query. Although they received rankings of 1.00 for responsiveness, they contained poorly generated extracts, which hindered their discovery.

Similarly, search query #2 also retrieved a poorly generated extract for the PDF file responsive to the search request. It is contrasted here with the responsive HTML file also retrieved:

Hit #1	1.00	HTML	Gasoline Distribution Industry (Stage 1) Background Information for Promulgated Standards, Final Summary: Gasoline Distribution Industry (Stage 1) Background Information for Promulgated Standards, Final Department of Commerce National Technical Information Service 5285 Port Royal Rd Springfield, VA 22151 Phone Number: 800-553-6847
Hit #13	0.97	PDF	C:ATOXUWEBTESTFI~1GDIBID.PDF Summary: ii This report has been reviewed by the Emission Standards Division of the Office of Air Quality Planning and Standards, EPA, and approved for publication. There were 48 comment letters (see Table 1-1) submitted by facility owners and operators, trade as

A review of the first 30 records displayed revealed only these two documents as responsive to the search query. The HTML file was the #1 hit in the displayed results while the PDF file was retrieved as hit #13. As previously stated, Verity searches the server containing the HTML files very quickly and can create a search zone from any HTML metatag. Verity has much more difficulty searching the Agency's PDF files. It is

a much slower process and since Verity is unable to search the PDF files by title, it instead must generate an extract by pulling random text from the document. Since there is currently no rhyme or reason to Verity's extract generation tools, the public must sift through unintelligible extracts until the problem can be resolved.

As long as responsive HTML files are also retrieved in addition to the PDF files, users are better able to discriminate between documents from the list of items retrieved. Unfortunately, PDF files were the only responsive documents retrieved for search queries #5, 7, 8, 11, 14, and 23. Without clear extracts or the experience of a trained searcher, the general public would have found it extremely difficult to locate responsive documents from the lists of items retrieved from these searches.

Chapter 6

Summary & Conclusions

Metadata has been proposed as one of many solutions to improve public access to varying electronic resources over the Internet and has captivated many information organizations seeking to increase the visibility and accessibility of their online resources. Although metadata has been used for various applications over the years, using metadata to describe and manage Internet resources is a relatively new endeavor, thus making it the neonate of electronic resource description. Although the potential for metadata to improve electronic resource description and discovery is indeed very great, there is still much to be learned about its performance and effectiveness.

Summary

As was revealed in the literature, much of the work conducted to date concerning metadata has been scholarly in nature and has focused upon the perceived effectiveness of metadata, rather than its actual performance. This study provided an excellent opportunity to assess one organization's early use and implementation of metadata in a real-life setting. The use of real patron queries for known items added credence to the study since satisfying real information needs was one of the primary reasons for the

EPA's implementation of metadata in the first place. The study was designed to be evaluative in nature and much information was gained about metadata generation and its functionality as a tool for resource discovery in a large organizational environment.

Conclusions:

The objectives of this study were accomplished and the results provide a unique overview of the performance of metadata in its early development. A primary goal of this study was to identify and assess those factors contributing to the effectiveness of metadata. On the basis of the evaluation of the results from the 24 search queries against the performance of the two search systems, the following conclusions and recommendations can be drawn.

First, greater emphasis must be placed on the generation of metadata by data owners for those items destined for EPA's Public Access Website. This research revealed that only 8 out of 24 search queries, or 33% of the queries retrieved responsive metadata records from the Web Inventory Database. If the EPA views metadata as a tool for improving the public's access to its informational resources, then the success of EPA's metadata system hinges upon a commitment by data owners to generate metadata for their documents. Perhaps an interface of some type could be created that prompts data owners to generate metadata records before documents can be published on the Web. Or, a more stringent policy, such as preventing any resources from reaching the Web unless they contain a corresponding metadata record, could also be instituted to assure compliance.

Testing the system with real-life search queries was a an effective way to learn about and discover flaws in the system. The EPA might find this practice advantageous as they continue to test the system in the future. Use of real-life search queries is good because developers and data owners are challenged to see the system as the users do, thereby catching a glimpse of what works in terms of resource description, and what does not.

The Verity search engine and its algorithms should be reevaluated and perhaps, adjusted. The results of this research indicate that Verity's ability to consistently rank like documents is lacking. In addition, since recall is currently so high for the Public Access Website, the rankings assigned to individual documents retrieved are not always realistic. For example, the first 81 documents retrieved for search query #8 were ranked 1.00, even though they were wholly non-responsive to the search request. A responsive item was finally retrieved at hit #151, which was still given a high 0.98 ranking by Verity. It is highly unlikely that public patrons would have the patience to scroll through 151 records before resorting to another search strategy. It is also very difficult to discriminate between and determine the responsiveness of numerous records, which have been assigned the same ranking. Verity's ranking ability is definitely a weakness to the system and is deserving of further attention by developers.

Greater specificity of indexing, in terms of keyword assignment, is also critical to the success of the EPA's system. Although necessary for consistency and good for describing more generic titles, the EPA's hierarchical controlled vocabulary could be limiting data owners' choice of keywords when describing their documents. In fact, the

results of this research indicate that data owners generally limit themselves to use of the EPA's hierarchical keywords, and do not bother to include additional, or more specific, keywords to describe their documents. As previously stated, only two out of the eight search queries retrieving responsive documents, contained additional keyword terms. Perhaps making the "additional keyword terms" field mandatory, where data owners would be required to enter one or two specific keywords to describe their documents, would increase precision during online retrieval.

Finally, the artificial extracts generated by Verity for the PDF files it retrieves do not allow users to sufficiently predict the responsiveness of a document based on a reading of the extract alone. In fact, the content of the extracts are rarely intelligible and hinder the public's ability to discriminate between like documents. What is worse is the fact that many responsive documents, containing extracts from both PDF and HTML files, receive the same Verity assigned ranking, making it even more difficult to choose between like documents.

Many organizations are beginning to implement metadata schemes in an effort to increase the visibility of their networked resources over the Internet. Although metadata can be extremely useful for improving access to electronic resources, it may not be the answer for every organization, and may have to be adjusted to fit the domain specific parameters of others still. It is hoped that the information contained in this study will be of use to those organizations interested in implementing metadata, as it provides an analysis of what factors contribute to the effectiveness of a good metadata system and which do not. This research also provides an overview of a very complicated, albeit

successfully operational, metadata system that provides distributed information resources to millions of people all over the world. This is rare as most metadata systems implemented to date have been for smaller organizations and thus created on a much smaller scale.

Finally, this research contributes to a body of research just beginning to appear in the literature and it is the hope of this author that this research encourages more organizations to explore the effectiveness of their own metadata schemes. Although great progress continues to be made in worldwide cooperative efforts directed toward the development of a single metadata framework, it is hoped that the evaluation of individual metadata systems, such as the EPA's system, will act as a guide toward the development and acceptance of an operational and effective metadata framework that can be used across domains around the world.

Appendices A-C are not available

in the PDF Formatted Version of this Document

Appendix D

EPA Reference Questions

Known Title or Known Subject	Search Queries
1. Wood Furniture Manufacturing NESHAP	"<and>(wood, furniture, manufacturing, operations,
2. Gasoline Distribution Industry (Stage I) Background Information for Promulgated Standards – Final	"<and>(gasoline, distribution, industry, background, information)"
3. Cities in Non-Attainment Status	"<and>(nonattainment, cities, status)"
4. OAQPS Cost Control Manual	"<and>(oaqps, cost, control, manual)"
5. Traffic Concerns - Air Quality Impacts of Traffic and Transportation	"<and>(traffic, transportation, air)"
6. Quality Assurance Handbook for Air Pollution Measurement Systems	"<and>(quality, assurance, air, pollution, measurement, systems, handbook)"
7. List of Designated Reference and Equivalent Methods	"<and>(list, designated, equivalent, reference, methods)"
8. EPA Test Method 21	"<and>(test, method, air, 21)"
9. CD-ROM Emissions Trends Viewer/Net Viewer	"<and>(emissions, trends, viewer,
10. New Regulations for National VOC Emissions Standards for Architectural Coatings	"<and>(architectural, coatings, national, standards)"
11. 1997 National Air Quality Trends Report	"<and>(national, air, quality, emissions, trends, report)"
12. Handbook for Air Toxics	"<and>(handbook, air, toxics)"
13. Emissions from heating and cooling technologies in residential units from oil or gas furnaces	"<and>(heating, cooling, residential, furnace)"
14. Benefits and Costs of Clean Air Act, 1970-1990	"<and>(benefits, costs, clean, air, act, 1970-1990)"
15. United States EPA, Environmental Investments: The Cost of a Clean Environment	"<and>(cost, clean, environment, environmental, investments)"
16. National Water Quality Inventory: 1996 Report to Congress	"<and>(national, water, quality, inventory, report, congress)"
17. 1997 Mercury Report to Congress	"<and>(mercury, report, congress)"
18. User's Guide for the Industrial Source Complex (ISC) Dispersion Models	isc, dispersion, models, user's, guide)"
19. Analysis of Composting as an Environmental Remediation Technology	"<and>(composting, analysis, remediation, technology, environmental)"
20. Safe Drinking Water Act – Reauthorization of 1996	"<and>(safe, drinking, water, act, reauthorization, 1996)"
21. Choosing Where You Live	"<and>(choosing, where, you, live)"
22. National VOC Emission Standards for Consumer Products - Automotive Refinishing	"<and>(national, volatile, organic, compound, emissions, standards, automotive,

EPA Reference Questions

Known Title or Known Subject	Search Queries
23. 1998 Interstate Ozone Transport Report (overview document preferred)	"<and>(1998, interstate, ozone, transport, report)"
24. Enabling Document for the New Source Performance Standards for Municipal Solid Waste Landfills	"<and>(municipal, solid, waste, landfills, enabling, document)"

References

Caplan, Pricilla, & Guenther, Rebecca. (1996). Metadata for internet resources: the Dublin Core metadata elements set and its mapping to USMARC. Cataloguing & Classification Quarterly, 22 (3/4), pp. 43-58.

Cathro, Warwick. (1997, August). Metadata: an overview. Paper from Standards Australia Seminar, "Matching Discovery and Recovery". [On-line]. Available: http://www.nla.gov.au/nla/staffpaper/cathro3.html.

CIMI metadata testbed project: implicit assumptions behind the Dublin Core and proposals for their exploration within the testbed. (1998). [On-line]. Available: http://www.cimi.org/documents/DC_hypothoses.html.

CIMI Dublin Core metadata testbed phase II. (1999, January 18). [On-line]. Available: http://www.cimi.org/documents/meta_011899_pd11_final.html.

Dempsey, Lorcan, & Heery, Rachel. (1998, March). Metadata: a current view of practice and issues. Journal of Documentation, 54 (2), pp. 145-172.

Efthimiadis, Efthimis N., & Carlyle, Allyson. (1997, October/December). Organizing internet resources: metadata and the web. Bulletin of the American Society for Information Science, 24 (1), p. 5.

Griffen, Steve, & Wason, Tom. (1997, November/December). The year of metadata. Educom Review [On-line serial], 32 (6). Available: http://www.educom.edu/web/pubs/review/reviewArticles/32656.html.

Hahn, Trudi B. (1998, April/May). Text retrieval online: historical perspective on web search engines. Bulletin of the American Society for Information Science, 24 (4), pp. 7-10.

Hakala, Juha. (1998, July). The Nordic Metadata Project: final report. Available: http://linnea.helsinki.fi/meta/nmfinal.htm.

Hudgins-Bonafield, C. (1995). Who will master metadata? Network Computing, 6 (8), pp. 102-108.

Iannella, Renato, & Waugh, Andrew. (1997). Enabling the internet [On-line]. Available: http://www.dtsc.edu.au/RDU/reports/CAUSE97.

Inmon, W.H. (1996). Building the data warehouse (2nd ed.). New York: John Wiley & Sons, Inc.

Lancaster, F.W. (1994). MEDLARS: report on the evaluation of its operating efficiency. In Karen S. Jones & Peter Willett (Eds.), Readings in information retrieval (pp. 223-246). San Francisco: Morgan Kaufman.

Lancaster, F.W. (1998). Indexing and abstracting in theory and practice (2nd ed.). Champaign, Illinois: University Of Illinois, Graduate School of Library and Information Science.

Lange, Holley R., & Winkler, B.J. (1997). Taming the internet: a work in progress. Advances in Librarianship, 21, pp. 47-72.

Larsgaard, M.L. (1996). Cataloguing planetospatial data in digital form: old wine, new bottles-new wine, old bottles. In Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information (L.C. Smith & M. Gluck, eds.), pp. 17-30. Clinic on Library Applications of Data Processing, 1995, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Lide, David R. (1995). Metadata: a description. Library Hi-Tech Issues, 49-50 (13:1-2), pp.33-34.

Lynch, Clifford. (1997, March). Searching the Internet. Scientific American, pp. 52-56.

MacClennan, Alan. (1998). Interesting times. Library Review, 47 (2), pp. 106-109.

Madsen, Mark S., Fogg, Ian, & Ruggles, Clive. (1994). Metadata systems: integrative information technologies. Libri, 44 (3), pp. 237-257.

Milstead, Jessica, & Feldman, Susan. (1999, January/February). Metadata: cataloging by any other name. Online, 23 (1), pp. 24-31.

Neuss, Christian, & Kent, Robert. (1995). Conceptual analysis of resource meta-information. Computer Networks and ISDN Systems, 27, pp. 973-984.

Oder, Norman. (1998, October 1). Cataloguing the net: can we do it? Library Journal, 123 (16), pp. 47-51.

Qin, Jian, & Wesley, Kathryn. (1998, September). Web indexing with meta fields: a survey of web objects in polymer chemistry. Information Technology and Libraries, 17 (3), pp. 149-156.

Sullivan, Danny. (1998, August 4). Search engine features chart. Search Engine Watch [On-line]. Available: http://searchenginewatch.com/webmasters/features.html.

Weibel, Stuart. (1997, October/November). The Dublin Core: a simple content description model for electronic resources. Bulletin of the American Society for Information Science, 24 (1), pp. 9-11.