Finding Finding Aids on the World Wide Web

Helen R. Tibbo and Lokman I. Meho

Abstract

This study explored how well six popular Web search engines performed in retrieving specific electronic finding aids mounted on the World Wide Web. A random sample of on-line finding aids was selected and then searched using AltaVista, Excite, Fast Search, Google, Hotbot, and Northern Light, employing both word- and phrase-searching. As of February 2000, approximately 8 percent of repositories listed at the "Repositories of Primary Resources" web site had mounted at least four full finding aids on the Web. The most striking finding of this study was the importance of using phrase searches whenever possible, rather than word searches. Also of significance was the fact that if a finding aid were to be found using any search engine, it was generally found in the first ten or twenty items at most. The study identifies the best performers among the six chosen search engines. Combinations of search engines often produced much better results than did the search engines individually, evidence that there may be little overlap among the top hits provided by individual engines.

Introduction

"The times, they are a-changing." Not so very long ago the idea of creating MARC AMC records for archival and manuscript materials struck fear in the hearts of archivists. The idea of descriptive standards, controlled vocabularies, and specialized containers to describe archival collections alarmed, dismayed, and infuriated many archivists who believed descriptive practice had to be embodied in finding aids and registers as unique as archival collections themselves. As Steve Hensen noted in 1986, it was part of the collective folklore of archivists that there is a certain idiosyncratic (some would even say eccentric) approach to certain aspects of the practice of the archival craft. This has certainly been true in the case of descriptive standards."[1] These often hotly debated issues played out across the decade of the

A draft of this study was presented at the Society of American Archivists annual meeting in Pittsburgh in August 1999.

[1] Steven L. Hensen, "The Use of Standards in the Application of the AMC Format," *American Archivist* 49 (Winter 1986): 32.

1980s at meetings of the Society of American Archivist and at regional archives meetings, as well as in the archival literature.[2] The work of the National Information Systems Task Force (NISTF), the development of the MARC/AMC format to accommodate cataloging of archival materials, and the publication of Steve Hensen's *Archives, Personal Papers, and Manuscripts* changed everything and sent archivists, some willingly, some kicking and screaming, down the high tech road to national access for descriptive tools.[3] In 1993 Lyn Martin could write that "U.S. MARC AMC (Machine-Readable Cataloging for Archives and Manuscript Control) has 'come of age,' taking its place in the mainstream of both archival and cataloging thinking, theory, and practice."[4]

With the advent of the World Wide Web (WWW), archivists immediately saw new, previously unimagined opportunities for providing remote users with not just cataloging descriptions of collections, but with actual finding aids and even digitized collections. This archival dreaming led to the Berkeley Finding Aid Project, directed by Daniel Pitti, and the early development of what has become the Encoded Archival Description (EAD) format for encoding finding aids for the Web.[5] Despite the work involved in encoding and mounting finding aids on the Web, the idea of providing full-text finding aids to users rings true with many archivists.[6] Here is an opportunity to provide all the information contained in a finding aid to potential users anywhere in the world! Many archivists

[2] See, for example: Working Group on Standards for Archival Description, "Archival Description Standards: Establishing a Process for Their Development and Implementation," *American Archivist* 52 (Fall 1989): 448–61; Working Group on Standards for Archival Description, "Recommendations of the Working Group on Standards for Archival Description," *American Archivist* 52 (Fall 1989): 462–77. The first articles on MARC AMC were published in the *American Archivist* in the fall 1984 issue.

[3] For NISTF see David Bearman, *Towards National Information Systems for Archives and Manuscript Repositories: The National Information Systems Task Force (NISTF) Papers, 1981–1984,* (Chicago, Society of American Archivist, 1987); Richard H. Lytle, "An Analysis of the Work of the National Information Systems Task Force," *American Archivist* 47 (Fall 1984): 357–65. For MARC AMC development see David Bearman, "Archives and Manuscript Control with Bibliographic Utilities: Challenges and Opportunities," *American Archivist* 52 (Winter 1989): 26–39; Shelia H. Martell, "Use of the MARC AMC Format by Archivists for Integration of Special Collections' Holdings into Bibliographic Databases and Networks," (M.S.L.S. thesis, University of North Carolina at Chapel Hill, 1991). For standards for descriptive tools see Steven Hensen, *Archives, Personal Papers, and Manuscripts: A Cataloging Manual for Archival Repositories, Historical Societies, and Manuscript Libraries* (Washington, D.C.: Library of Congress, 1983). 2nd ed. (Chicago: Society of American Archivists, 1989).

[4] Lyn M. Martin, "Viewing the Field: A Literature Review and Survey of the Use of U.S. MARC in U.S. Academic Libraries," *American Archvist* 57 (Summer 1994): 482. Lyn Martin presented selected results of this research in June 1993 at the State University of New York Librarians Conference in Binghamton, New York.

[5] Daniel V. Pitti, "Encoded Archival Description: The Development of an Encoding Standard for Archival Finding Aids," *American Archivist* 60 (Summer 1997): 268–283; Daniel V. Pitti, "Encoded Archival Description. An Introduction and Overview," *D-Lib Magazine* 5 (Nov 1999), <http://www.dlib.org/dlib/november99/11pitti.html>.

[6] Steve Hensen discusses how EAD is becoming part of the mainstream archival standards, building on NISTF's development of the MARC AMC cataloging form and his own *Archives, Personal Papers, and Manuscripts* as a cataloging manual. Steven L. Hensen, "NISTF 2 and EAD: The Evolution of Archival Description," *American Archivist* 60 (Summer 1997): 284–96. For further discussion of EAD as a descriptive standard, see also: Kris Kiesling, "EAD as an Archival Descriptive Standard," *American Archivist* 60 (Summer 1997): 344–54.

now believe that mounting finding aids on the Web makes them instantly, constantly, and consistently available to anyone with Internet access. Indeed, this is fuel for the argument that archivists no longer need to produce MARC records for national databases such as OCLC or RLIN. If finding aids and even parts of collections are available on the Web, why would anyone search OCLC or RLIN to find a highly condensed surrogate of a finding aid, especially when relevant records are hard to locate within these databases and users often times find them difficult to interpret?[7] An interesting question and, given the cost of MARC cataloging, one that needs to be explored for economic reasons if nothing else.

The underlying premise of the above argument is that once a finding aid is mounted on the Web, users will be able to find it easily. Many factors, however, including search engine features, searcher skill, and the sheer size of the World Wide Web, influence the ease with which users may retrieve a given finding aid from the World Wide Web. What success can archivists, on average, expect users to have when they search the Web for archival materials? How easily will users discover a specific finding aid? These are the questions that motivated the exploratory research presented here.

There is already an extensive literature, both in print and online, concerning the nature of Web search engines and the challenges of locating material in this vast virtual environment.[8] Many libraries provide useful Web searching tutorials.[9] All Web search engines mount search help, tips, or FAQ pages to assist in the searching process. Yet, it is unclear who reads these pages beyond students who are in bibliographic instruction classes and we have no idea as to these pages' efficacy when they are read. Increasingly, Web search engines such as Alta Vista, Hotbot, and Northern Light are becoming more powerful and

[7] Helen R. Tibbo, "The Epic Struggle: Subject Retrieval from Large Bibliographic Databases," *American Archivist* 57 (Spring 1994): 310–26. Robert P. Spindler, "Does AMC Mean 'Archives Made Confusing"? Patron Understanding of USMARC AMC Catalog Records," *American Archivist* 52 (Spring 1993): 330–41; Susan L. Malbin, "Does AMC Really Mean 'Archives Made Confusing'? Retesting Patron Understanding," *Technical Services Quarterly* 16 (1998): 15–32.

[8] See, for example, Michael D. Gordon and Praveen Pathek, "Finding Information on the World Wide Web: The Retrieval Effectiveness of Search Engines," *Information Processing and Management* 35 (March 1999): 141–180; Steve Lawrence and C. Lee Giles, "Searching the World Wide Web," *Science* 280 (April 3, 1998): 98–100; Steve Lawrence and C. Lee Giles, "Searching the Web: General and Scientific Information Access," *IEEE Communications Magazine* 37/1 (January 1999): 116–122; H.V. Leighton and J. Srivastava, "First 20 Precision among World Wide Web Search Services," *JASIS* 50/10 (1999): 870–881; Greg R. Notess, "On the Net: Internet Search Techniques and Strategies, *Online* 21/4 (July 1997):, Greg R. Notess, "On the Net—More Internet Search Strategies," *Online* 22/5 (September 1998): 71–74, <http://www.onlineinc.com/onlinemag/OL1988/net9.html>. Greg R. Notess., "On the Net—Rising Relevance in Search Engines." 23/3 (May 1999): 84–86 <http://www.onlineinc.com/onlinemag/OLtocs/OLtocmay4.htm>. Greg R. Notess, "Internet Search Engine Update—New Search Features, Developments, and Content," *Online* 24/3 (May 2000): <http://www.onlineinc.com/onlinemag/OL2000/engine5.html>.

[9] See, for example, UCLA: <http://www.library.ucla.edu/libraries/college/instruct/instgui.htm#internet); Widener University: <http://www2.widener.edu/Wolfgram-Memorial-Library/webhome.htm>; UC Berkeley; <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfo.html>. For an on-line bibliography of instructional sites see the Library of Congress: <http://lcweb.loc.gov/global/internet/training.html>.

flexible, but in the process they are becoming more complicated to use optimally, presenting searchers with a number of options. At the same time, optimizing search strategies and techniques is often critical for useful retrieval due to the ever-increasing size of the Web.[10] Given that few individuals are willing to look at more than thirty web pages, with most stopping after ten or twenty for a particular web search, and given that many web searches retrieve hundreds of thousands of hits, the importance of optimizing searches to bring the most relevant items to the top of the retrieval set cannot be overly-stressed.

Every search engine company is working to improve its search algorithms and subsequent ranking algorithms in order to bring the most relevant items to the user at the top of the pile, but the recent improvements seem relatively minor. To add to the complexity of the situation for information seekers, each search engine employs its own proprietary algorithms and even has unique subsets of the entire Web in its index. In an article in *Nature* in 1999, Lawrence and Giles report that as of February 1999, the publicly-indexable Web contained an estimated 800 million pages, and that no engine indexed more than about 16% of this territory.[11] They also found that engines are more likely to index commercial (83%) rather than educational sites (6%), and that only 34% of sites used simple HTML keywords and description metatags, and that only 0.3% of sites used the Dublin Core metadata standard.[12] They note that "The current state of search engines can be compared to a phone book which is updated irregularly, is biased toward listing more popular information, and has most of the pages ripped out."[13]

Because each engine indexes different parts of the Web, and because they all use different retrieval and ranking algorithms, identical search protocols deployed on various search engines can produce slightly to extremely different search results, but almost never identical retrieval sets. Searchers need to know how to best search each engine for particular types of queries to optimize the relevance of retrieved items and determine which engines work best for a given type of query or a topic area. Even seemingly small changes in search query formulation, such as capitalization, truncation, or searching words as bound phrases can produce dramatically different results. It is unlikely, however, that many web searchers take such matters into account when they are looking for information.

[10] In December 1997 Lawrence and Giles, "Searching the World Wide Web," 98–100, estimated that the publicly indexable Web contained 320 million pages. In February 1999 this figure had grown to 800 million pages [Steve Lawrence and Lee Giles, "Accessibility and Distribution of Information on the Web," *Nature* 400 (July 8, 1999):107–109 (summary of article at: <www.wwwmetrics.com>), and as of February 2000 Search Engine Watch [www.searchenginewatch.com] estimated that there were one billion pages on the publicly-accessible Web.

[11] Steve Lawrence and Lee Giles, "Accessibility," 107.

[12] Dublin Core web site maintained by OCLC: <http://purl.oclc.org/dc/>.

[13] Steve Lawrence and Lee Giles, "Accessibility and Distribution of Information on the World Wide Web," available at <http://www.wwwmetrics.com>.

In designing this study, we assumed that individuals looking for archival materials on the Web would have little searching expertise. We started with known items—finding aids we had found on the Web through archival sites, we then used various search engines and simplistic search strategies to find these finding aids as a non-expert web searcher might. Because we started with the full text of the finding aids in hand, we greatly biased the results in favor of finding our finding aids. For example, we searched on exact titles as phrases and selected other very specific phrases from the finding aids, rather than searching on broad terms such as "Civil War" and "women." Thus, it is highly unlikely that real users with real subject queries, rather than searching for known items, would receive the positive results we did. Subsequent studies of Web retrieval effectiveness should involve real users and real queries. The current research sets out the baseline for whether items, given the best of circumstances, will be found.

Methodology

Because archivists have only recently been placing finding aids on the Web, and have primarily been concerned with questions of encoding and formatting, no one has yet to explore the current retrieval efficacy of the Web for archival materials. We started this research with the basic question, "How likely is it that a user would find a specific finding aid mounted on the Web, given today's most popular search engines and the most likely search strategies?"

The first step was to locate repositories with electronic finding aids. We could have done this very easily by selecting some of the largest institutions, already known for their Web presence, but decided that we wanted to have a more representative sampling of finding aids from a wider group of institutions with wider variance in descriptive and technological practices. To select such a sample, we started with the list of Repositories of Primary Resources, limiting the sublists to the United States and Canada, found on the Web site at the Special Collections and Archives Department of the University of Idaho.[14] Each repository listed in this directory has a web link, so there is evidence that the personnel at these repositories have some degree of technological expertise and could be mounting finding aids on their web sites.[15] To be as inclusive as possible, we selected all of the institutions on three lists to form our domain: Western U.S. and Canada, Eastern U.S. and Canada—States and Provinces

[14] <http://www.uidaho.edu/special-collections/Other.Repositories.html>.

[15] From the "Guidelines" for inclusion of web sites in the *Repositories of Primary Sources*: "This list of Repositories of Primary Sources is solely of Web, and a diminishing number of gopher—sites that describe collections of rare books, manuscripts, archives, historical photographs, oral histories, or other primary sources. The list focuses on actual repositories; therefore virtual collections are excluded. Each site has a separate web page or named part of a web page (i.e., it provides a direct URL to the relevant part of the page), which generally provides a description of the collection. The links are to the web sites, not email addresses or telnet access to bibliographic databases, although those can often be found on the web pages."

A–M, and Eastern U.S. and Canada—States and Provinces N–Z. Integrating these lists resulted in 1,974 items. Next, we randomly selected repositories on the list until we had accumulated twenty-five institutions that had at least four full HTML finding aids mounted on the Web. We looked at 309 institutions' web sites to find this sample.

In the process of selecting repositories, we visited each web site to see if the institution had mounted any finding aids in HTML format. It was important that the finding aids be in HTML (Hypertext Markup Language) and not in SGML/EAD (Standard Generalized Markup Language/Encoded Archival Description) format, as SGML encoding is invisible to many search engines.[16] The use of XML (Extensible Markup Language) may soon solve this problem, as all XML/EAD finding aids will be fully searchable across the Web, but, of course, we do not yet know how XML and new metadata RDF (Resource Description Framework) standards will affect retrieval, nor when Web search engines will accommodate XML.[17] Interestingly, of the approximately two thousand repositories that appear to have web pages, only 8 percent had mounted at least four full finding aids on their web sites as of February 2000.

In order for an institution's finding aids to be used in this study, they had to contain, at minimum, five of the following six elements: title, inclusive dates, extent or cubic feet of the collection, scope and contents note, biographical or historical note, and a statement about arrangement. These are the basic finding aid elements advocated in texts such as Fredric Miller's *Arranging and Describing Archives and Manuscripts*.[18] We set this minimum content because these elements are increasingly regarded as standard information for finding aids. Also, comparing retrieval results of homogeneous items gives each item a more equal chance of being retrieved than comparing the results of very different texts, such as a full finding aid and a mere abstract of a finding aid or very brief collection description.

Once we identified the twenty-five repositories, we randomly selected four full finding aids from those listed at each site. Next, we printed out the finding aids and looked for four key terms or phrases that were central to the content of each of the collections.[19] In each case we selected the title of the collection

[16] SGML/EAD finding aids are searchable within a repository's web site, but are not picked up by most of the major search engines when someone enters search terms and looks across the entire Web.

[17] There is already an enormous amount of literature concerning XML and RDF. See the OASIS SGML/XML web page by Robin Cover for extensive explanations, bibliographies, standards, and position papers at: <http://www.oasis-open.org/cover/xml.html and http://www.oasis-open.org/cover/rdf.html>. See also, Jon Bosak and Tim Bray "XML and the Second-Generation Web" *Scientific American* (May 1999): available at <http://www.sciam.com/1999/0599issue/0599bosak. html>.

[18] Fredric M. Miller, *Arranging and Describing Archives and Manuscripts* (Chicago: Society of American Archivists, 1990).

[19] Most of the "terms" were multiword "phrases." We use these two words interchangeably in this article, but do make an important distinction between "word searching" and "phrase searching."

as one of the phrases. For the remaining three terms, we selected phrases that were very specific and would have a good opportunity of being retrieved when searched as a phrase in any of the major search engines. We selected the terms from four different sections of each finding aid whenever possible. We also tried to include different types of terms from the same finding aid (e.g., personal names, titles of publications, names of organizations). When only four good terms were found, they were all selected irrespective of their location in the finding aid. The sections of the finding aids used to select terms from included:

1. "Abstract"
2. "History" or "Historical Sketch"
3. "Introduction"
4. "Biography" or "Biographical Sketch"
5. "Scope" and "Content" "Scope and Content"
6. "Description"

Many of the phrases used as search terms were personal names, institutional names, or titles of publications. Some examples include: "Los Angeles Clinical Oncology Program," "William Merritt Chase," "Navajo Folk Art," "University of the Air," and "Sunnyville Coal Mines." The final list includes 400 terms or phrases.

Once we had the list of finding aids and search terms, we selected six search engines: Alta Vista, Excite, Fast Search, Google, Hotbot, and Northern Light.[20] We selected these particular engines for several reasons, although there were several other good candidates. First, four of these engines, Alta Vista, Excite, Hotbot, and Northern Light were featured on the University of California at Berkeley Library's web site that provides an excellent "how to search the Web" tutorial, when we began this project.[21] This is one of the best web instructional sites, with an excellent reputation in the user education domain. We subsequently added Google and Fast Search (AlltheWeb.com) to our list because of great reviews at sites such as (Search Engine Watch) that provide useful analysis of the operation of each engine along with comparisons, reviews, and rankings.[22] Subsequent to our searching these engines, Berkeley's library also added them as recommended search engines to their instructional site. This, along with our findings, gives us confidence that these were, indeed, good selections. Second, all these engines are very popular and it is quite reasonable to expect that people looking for archival materials would employ one or more of them. Third, all of these engines have a history of garnering good reviews and thus should yield comparable and most likely better, results than

[20] Alta Vista <www.av.com>; Excite <www.excite.com>; Fast Search <www.alltheweb.com>; Google <www.google.com>; Hotbot <www.hotbot.com>; and Northern Light <www.nlsearch.com>.

[21] <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfo.html>.

[22] <www.searchenginewatch.com>.

other engines.[23] Finally, this array of engines includes a good diversity considering several other factors. Some, such as Alta Vista, are relatively old engines with a long indexing and retrieval history. Google, on the other hand, is a newer entry. Excite has one of the smaller indexes, while Google and Alta Vista have two of the largest. All these engines allow for phrase searching, but their retrieval and ranking algorithms are very different. While many rely on relevance ranking based on word matching, term frequency and co-occurrence, Google employs site popularity as a ranking mechanism. Each engine explains how it works to some degree, but in each case the algorithms are proprietary information.

We entered each of the terms listed in Appendix B into the six search engines, both as phrases and as words, using capitalization in all cases for proper nouns as they appeared in the finding aids and in Appendix B. Example of phrase searches included:

"Wiley Winsor Memorial Heart Research Foundation"
"Southern Californian Daughters of Charity"
"Pratt Institute Art School"

These same key terms searched as words:

Wiley Winsor Memorial Heart Research
Southern Californian Daughters of Charity
Pratt Institute Art School

In some cases, quotation marks were used to designate phrase searching, but if a search box allowed for the option of phrase, searching this option was selected. Phrase search, however it is designated, requires that multiword phrases be found with the individual words in the order listed in the query. For example, searching on "President William Jefferson Clinton" would return web pages in which those four words appeared together in the given order.

In the case of the word searching, the words were entered as a string but not with quotation marks. If a search box provided an option, "all the words," it was selected. This is equivalent to Boolean AND, and would allow for the smallest possible retrieval set with the most relevant items versus the choice of "any of the words" or Boolean OR. Searching on

President William Jefferson Clinton [all the words]

would return pages that have those four words anywhere within them. Changing the option to [any of the words] would mean that any retrieved web page would only have to have at least one of the words, although those with all four (and multiple incidences of the terms) would appear at the top of the list. When users do not add Boolean AND or OR in between search terms, most web

[23] For additional search engine review sources see: CNET: <http://www.cnet.com/internet/>. "Search Engine Showdown"<http://www.searchengineshowdown.com/reviews/>.

search engines default to AND. Alta Vista is an exception that defaults to OR, and thus retrieves much larger sets than many of the other engines. This is good for high recall but often prohibits precise searches. In each search engine the retrieval of web pages is only half of the process. Next they rank the documents with the goal of listing the most relevant ones first. Because humans will only look at a few items and because most web retrieval sets are very large, the relevance ranking algorithms are as important as the actual retrieval method.

The idea behind our search methodology was to use the most straightforward approach to searching possible, as we believed that most people searching the Web for archival materials would use such an approach. This is why we used the very simple word searches wherein the words of the search phrase were just entered as strings without the use of Boolean operators. At the same time, we wanted to give each search engine the best possible chance for success. This is why we also searched each term as a phrase. Although many naïve users would not enter the search strings as phrases (generally in quotation marks in most search engines), this search strategy dramatically improves the chance that a given web page will appear within the first few retrieved items.

Almost as important as whether an item is returned in a retrieval set is where it appears in the set. In searching on each term, both as a word string and as a phrase, we noted whether or not the finding aid was retrieved in the first thirty items of each retrieval set and, if so, where it appeared in the list of retrieved items. We decided to look at no more than thirty items, as we believed that few searchers would exhibit more persistence and continue to click on web pages after having scanned the first thirty. There is as yet no research indicating how persistent users are with regards to the number of web pages they will go through and click on, but thirty does seem to be a significant cut-off number with users of library catalogs and on-line search services such as DIALOG.[24] Indeed, many users may look at only the first ten or twenty web page citations in a search engine's retrieval set. Because of the limitations of human persistence when searching the Web, we ranked any finding aid located in our searches, 1 to 30, based on where it appeared in the first thirty items of the retrieval set.

Findings

Electronic Finding Aids

While close to 2,000 archival and manuscript repositories in North America have web pages, extrapolating from our data, approximately only 8 percent or 160, had mounted any significant number of full finding aids on

[24]Stephen E. Wiberley, Jr., Robert A. Daugherty, and James A. Danowski, "User Persistence in Scanning Postings of a Computer-Driven Information System: LCS," *Library and Information Science Research* 12 (October/December, 1990): 341–53. See also, Stephen E. Wiberley, Jr., and Robert A. Daugherty, "Users' Persistence in Scanning Lists of References," *College and Research Libraries* 49 (March 1988): 149–56. See also, Marcia J. Bates, "The Fallacy of the Perfect Thirty-Item On-line Search," *RQ* 24 (Fall 1984): 43–50.

their sites by February 2000.[25] This suggests that repositories are using their sites to describe the institution in general—providing repository holdings descriptions, access policies, and hours of operation. Certainly this is important information to have on a repository's web site, but it is clear that most institutions have yet to mount their finding aids. This is an indication that while some repositories are quite technologically advanced and have the staff or grant funding to create, tag, mount, and maintain electronic finding aids on the Web, most archival organizations have yet to move in this direction.

Retrieval Results

What are the chances that a person interested in finding a specific collection represented by a finding aid on the Web will locate it? Undoubtedly, the best way to answer this question is to run a study with a population of individuals interested in finding archival collections via the Web. That more ambitious study is not the one presented here, but archivists do need to conduct such research to discover how people are looking for their materials. This study was much more simplistic and also more controlled. Each search was conducted in a similar way, looking for very specific terms and phrases from the finding aids. The phrase searches, in particular, were constructed so as to locate as many of the finding aids as possible with the smallest possible retrieval sets. This greatly improved the chances that the desired finding aid appeared in the first thirty items of the retrieval sets.

So, how did the search engines do in finding the finding aids? How many finding aids did each of the search engines find? What were the best combinations of search engines to use to locate archival finding aids? The most important thing to note is that for each search engine, phrase searching did better than the simplistic word searching that many naïve users employ. This is probably not a matter of search recall as one might suspect, but rather, an artifact of search precision. Undoubtedly, many of the word searches, which resulted in much larger retrieval sets than did the phrase searches, retrieved the relevant items, but we will never know this, as they retrieved so many other items that the desired ones were buried under an avalanche of irrelevant information.

For all the phrase searches using either the collection title or terms from the finding aid, Fast Search located the finding aid 65% of the time; Google 59% Northern Light 56% Alta Vista 52% Excite 31%; and Hotbot a miserable 17% of the time. This is not to say that Excite and Hotbot did not find a majority of the same finding aids, but that these items did not appear in the first thirty hits for either search engine. Excite's and Hotbot's failure in these searches may be

[25] A very small number of repositories have mounted finding aids in SGML/EAD encoding. These are not counted in this study, but would not change the 8% figure significantly at this time. A small number of these sites, however, have mounted a large number of EAD finding aids.

primarily one of poor ranking rather than poor retrieval, but this is a relevant issue only for the search engine designer, not the typical end user. The word searches were generally far worse than the phrase searches on the same search engines. Google and Northern Light retrieved only 49% of the finding aids, with Fast Search dropping 31 points to 34%, followed by Alta Vista at 30%. Excite retrieved only one-fourth of all the finding aids when terms were searched word by word and Hotbot fell to 15%. These results are summarized in Table 1.

When just searching on collection titles and not any of the other terms taken from the text of the finding aids, the results improve dramatically, as can be seen in Table 2. Fast Search retrieves 87% of the collection descriptions, Google finds 76%, Northern Light 75%, Alta Vista 71% of the collections, Excite 49%, and Hotbot slightly over one-third at 34%, when using phrase searches. Even when using word searches, the results improved substantially: Google 49% to 72%; Fast Search 34% to 69%; Northern Light from 49% to 69%; Alta Vista from 30% to 53%; Excite goes from 25% to 40%, while Hotbot more than doubles its retrieval percentage to 32%. These are still not the results that most archivists would like, especially when someone is searching on a term as unique and specific as a collection title, but they are much better than keyword searches of titles and finding aid items, indicating the usefulness of title searching. Of course, many users, including school children and anyone exploring a topic at the beginning stages of research are unlikely to know which

Table 1 Number and Percentage of Items Retrieved Searching for Titles and Other Terms* (N = 400)

	Phrase Search		Keyword Search	
Search Engine	#	%	#	%
Alta Vista	206	**52**	118	**30**
Excite	125	**31**	98	**25**
Fast Search	261	**65**	135	**34**
Google	237	**59**	197	**49**
Hotbot	66	**17**	60	**15**
Northern Light	223	**56**	196	**49**

* Results are based on the top 30 retrieved items

Table 2 Percentage of Items Retrieved Using Titles Only* (N = 100)

Search Engine	Phrase Search	Keyword Search
Alta Vista	71	53
Excite	49	40
Fast Search	87	69
Google	76	72
Hotbot	34	32
Northern Light	75	69

* Results are based on the top 30 retrieved items

collection titles to use as search strings. Indeed, it is unrealistic to expect most scholars to know the specific title of collections they seek, even when they have used these titles before. This is the known item search so familiar in the library world, where users often have citations to specific titles and authors to follow. This is, however, rarely the case with archival researchers. For example, a user may wish to view the finding aid for the "Cameron Family Papers" housed at the University of North Carolina at Chapel Hill. He or she, however, may well search for "Cameron Papers" and retrieve nothing using a phrase search as the word "family" is missing from the title. Or it may be that the desired collection is the "Mary Cameron Papers," a different collection altogether.

It is important to note that while phrase searching, and especially phrase searching of specific collection titles, worked very well in this study, if the user does not enter a specific phrase (especially a lengthy title), correctly, he or she will retrieve no items. The retrieval precision the phrase search provides, demands search query precision from the user, so it stands as a double-edged sword.

Next, we compared the retrieval sets using SPSS to find which pairs of search engines worked best in combination. When searching for the titles and terms using phrases, the combinations of Alta Vista-Google, Fast Search-Google, and Alta Vista-Fast Search proved best, as seen in Table 3, retrieving 78%, 77%, and 76% of the items sought, respectively.

In summary, Table 3 shows that if an individual is conducting phrase searches on finding aids using both titles and other terms from the finding aids, he or she will find 78% of the finding aids when using Alta Vista and Google, provided that they look only among the top 30 items in the retrieval sets. If Alta Vista is used by itself, the figure drops to 52%; Google alone produces 59%.

When looking for the best combination of search engines for searching only titles in the phrase mode, the results are quite outstanding. The combinations of Google and Alta Vista or Google and Fast Search produced a near perfect 95% of all the finding aids sought. See Table 4.

Tables 5 and 6 provide the combinations for the word searching of all terms and just the titles, respectively. Again, the results, while not nearly as good

Table 3 Union of Phrase Searches Using Titles and Keywords Percentage of Finding Aids Retrieved

Search Engine	Alta Vista	Excite	Fast Search	Google	Hotbot	Northern Light
Alta Vista	52	63	**76**	**78**	56	70
Excite		31	70	68	38	65
Fast Search			65	**77**	66	71
Google				59	69	74
Hotbot					17	58
Northern Light						56

Table 4 Union of Phrase Searches Using Titles Only Percentage of Finding Aids Retrieved

Search Engine	Alta Vista	Excite	Fast Search	Google	Light	Northern Hotbot
Alta Vista	71	85	93	**95**	76	87
Excite		49	91	83	61	87
Fast Search			87	**95**	88	90
Google				76	89	93
Hotbot					34	79
Northern Light						75

Table 5 Union of Keyword Searches Using Title and Keywords Percentage of Finding Aids Retrieved

Search Engine	Alta Vista	Excite	Fast Search	Google	Hotbot	Northern Light
Alta Vista	30	43	46	59	37	57
Excite		25	42	57	33	57
Fast Search			34	57	39	57
Google				49	57	**66**
Hotbot					15	54
Northern Light						49

Table 6 Union of Keyword Searches Using Titles Only Percentage of Finding Aids Retrieved

Search Engine	Alta Vista	Excite	Fast Search	Google	Hotbot	Northern Light
Alta Vista	53	70	81	82	67	78
Excite		40	75	80	55	78
Fast Search			69	84	76	85
Google				72	85	**89**
Hotbot					32	77
Northern Light						69

as with the phrase searching, are significantly better than when each engine is used alone.

Looking at the union of only the top twenty hits from above combinations of search engines rather than the entire top thirty for each engine, we find that the percentage retrieved only drops between zero and three percent for each table.

As noted above, research has shown that most people have limited persistence with regards to the number of hits they will scan from any sort of bibliographic retrieval set. While we do not yet know how this translates to persistence

of scanning web pages returned from search engines, it is unlikely that most people will look at more web pages than they would catalog records or bibliographic entries with abstracts, simply because of the amount of clicking necessary and the time required for web pages to load. Therefore, how web search engines rank the retrieval sets they produce is very important. So, how did our engines do in terms of placing the retrieved finding aids within the top thirty of their sets?

As far as ranking is concerned, for items retrieved in the phrase searches, the results are very encouraging. Ninety-one percent of the items found in Alta Vista, 95% in Excite, 90% in Fast Search, 91% in Google, 98% in Hotbot, and 94% in Northern Light were found within the first ten items, normally the first screen or two of items presented by these search engines. Table 7 provides full details for the rank of retrieved items from the phrase searches.

Here we can see that, of the phrase searches that successfully retrieved the desired finding aid within the first thirty items, each search engine found 95% of these items within the top twenty hits. Given that, on average, approximately only 47% of the original 400 finding aids were retrieved within thirty items by the individual search engines, the results are a bit like the little girl in the rhyme; either they are very, very good, or they are horrid! Either the engines found the finding aids within the first twenty items retrieved or they didn't seem to find them at all. Table 8 provides similar data for the word searches. From

Table 7 Rank of Items Retrieved in Phrase Searches

	Top 1			Top 5			Top 10			Top 20			Top 30		
Search Engine	#	%*	%†	#	%*	%†	#	%*	%†	#	%*	%†	#	%*	%†
Alta Vista	109	27	53	175	44	85	187	47	91	198	50	96	206	52	100
Excite	85	21	68	113	28	90	119	30	95	125	31	100	125	31	100
Fast Search	167	42	64	224	56	86	235	59	90	248	62	95	261	65	100
Google	156	39	66	199	50	84	215	54	91	225	56	95	237	59	100
Hotbot	41	10	62	64	16	97	65	16	98	66	17	100	66	15	100
Northern Light	134	34	60	192	48	86	210	53	94	214	54	96	223	56	100

* Percentage of original 400 finding aids.

† Percentage of finding aids retrieved in top 30 items for search engine.

Table 8 Rank of Items Retrieved in Keyword Searches

	Top 1			Top 5			Top 10			Top 20			Top 30		
Search Engine	#	%*	%†	#	%*	%†	#	%*	%†	#	%*	%†	#	%*	%†
Alta Vista	65	16	55	89	22	75	92	23	78	105	26	89	118	30	100
Excite	60	15	61	84	21	86	86	22	88	98	25	98	98	25	100
Fast Search	73	18	54	107	27	79	116	29	86	124	31	92	135	34	100
Google	114	29	58	154	39	78	175	44	89	183	46	93	197	49	100
Hotbot	41	10	68	51	13	85	57	14	95	58	15	97	60	15	100
Northern Light	129	32	87	170	43	87	178	45	91	188	47	96	196	49	100

* Percentage of original 400 finding aids.

† Percentage of finding aids retrieved in top 30 items for search engine.

Tables 7 and 8 we can see that there is little reason for a searcher to browse through more than the first twenty items of a retrieved set.

Discussion

Although it is just the beginning of an understanding of what type of access the Web and today's search engines provide for archival materials, this small study has several implications for archivists. First, it is clear that despite all the hype and fuss, few archives are presently mounting finding aids on the Web at this time, although some very large sites are mounting a large number of finding aids and digitized collection material.[26] Despite the rather dismal retrieval findings of this study, having finding aids on the Web for users who know of an institution and its collections can be extremely useful. Researchers can now study finding aids remotely before travelling to a repository. It is surprising that by February 2000, only 8 percent of repositories had mounted at least four full finding aids on their web sites. This is particularly disappointing because many repositories appear to have the technical expertise to mount a general information homepage. Lack of funding to grow and maintain the web sites may well be the problem, but we lack data to support this hypothesis at this time. In the future we hope to explore the distribution of electronic finding aids, technical expertise, and staffing patterns a bit more closely, classifying repositories by type and size, and seeing if they are mounting electronic finding aids and what other types of materials their web sites contain. We can easily start to think of web sites from various institutions as containing pages that belong to distinct genres. An in-depth content analysis of archival, and more broadly, cultural heritage sites promises to reveal a good deal about what institutions hold to be important and what services and materials they offer their users. Such an analysis of those institutions that are leading the way in providing digital access to their collections and discovery tools, using sound web design principles, and coupled with user studies, should help other repositories better design future Web sites.[27]

Clearly, the most striking finding was the importance of using phrase searches whenever possible rather than word searches. The results of the phrase searches were always better than word searches on the same databases. This

[26] See for example: The California Digital Library's On-line Archive of California available at <http://www.oac.cdlib.org>. The OAC is a pilot project to develop a University of California-wide prototype union database of 30,000 pages of archival finding aid data encoded using the Encoded Archival Description (EAD) SGML document type definition. The University of Virginia <http://www.lib.virginia.edu/speccol/> and Duke University's <http://scriptorium.lib.duke.edu/> have also mounted significant numbers of finding aids on the Web along with the Library of Congress <http://lcweb.loc.gov/rr/mss/>.

[27] For guidance in Web design, see: Patrick J. Lynch and Sarah Horton, *Web Style Guide: Basic Design Principle for Creating Web Sites* (New Haven: Yale University Press, 1999) available at <http://info.med.yale.edu/caim/manual/> Louis Rosenfeld and Peter Morville, *Information Architecture for the World Wide Web* (Sebastopol, Calif.: O'Reilly, 1998; and Darrell Sano, *Designing Large-Scale Web Sites A Visual Design Methodology* (New York: Wiley, 1996).

makes sense, as the phrase searches produced smaller retrieval sets and there was a greater likelihood that the desired items would be located in the first thirty pages retrieved. With such large retrieval sets as are often produced on the Web, precision, practically translated into optimized ranking algorithms, is probably more important than high recall. A related finding was that searching on a collection's title (if one knows precisely what it is) as a phrase is the most likely way of finding its finding aid on the Web. Of course, many, if not most, researchers will not know the exact title of a collection.

Also of significance was the fact that if a finding aid was to be found using any search engine, it was generally found in the first ten or twenty items, at most. The clear implication is that if you do not find relevant web pages in the first ten or twenty hits, abandon that search and try a different more focused search query, using phrases if possible. After looking at twenty web pages listings, there is a clear point of diminishing returns. Related to this, we found that the combination of search engines, at least a few selected engines (right now, Fast Search, Google, Northern Light, and Alta Vista), often produced much better results than did the search engines individually. This is evidence that there is frequently little overlap between the top hits provided by individual engines, and that searching more than one engine for a query greatly improves the chance of finding a specific finding aid.

We do not advocate metasearch engines as much as doing the same search sequentially in two or three search engines until a reasonable number of items are retrieved. With metasearch engines, you frequently cannot control the specific engines being searched, and because of their speed, many metasearch engines retrieve only the few highest ranked hits from each individual search site they harvest. Also, most metasearch tools will time-out search engines or directories that take too long to respond to queries. Finally, metasearch engines do not allow users, nor do they themselves, optimize a query for a given search engine and its unique characteristics. It is far better for users to master the ins and outs of a few engines that work well for the types of material they are trying to find, and to employ them one at a time, using the best possible search query in each system.

The next interesting finding is that Google and Fast Search each produced better retrieval results than did Alta Vista and Northern Light, and much better results than did Excite and Hotbot. When this study was first reported in August 1999 at the SAA annual meeting, both Google and Fast Search were lesser known and not evaluated in our data at that time. Thus, archivists need to keep current with new search engines as they appear and test their usefulness for retrieving archival materials. While Google and Fast Search do well today, they may be tomorrow's runners-up.

The reasons for particular search engine success are not clear at this point, but we will be exploring this further as this line of research continues.

Explanations include the possibility that more archivists have registered their institutional pages with particular engines or that the specific retrieval algorithms employed work better for these materials. Despite the claims of most search engines that their spiders or robots look in every nook and cranny on the Web to find pages to include in their indexes, registering a page is the best way to facilitate its appearance in a web search. Another explanation could be the relative sizes of the databases of indexed web pages that feed each search engine. Alta Vista and Northern Light have two of the largest databases, but Hotbot is also extremely large and it had the poorest retrieval results, so this may not be a significant factor in explaining retrieval engine success with the finding aids. Undoubtedly the specific retrieval and ranking algorithms each search engine uses are key to explaining these results. These are, of course, beyond the control of the archivist mounting finding aids on the Web or the searcher, but knowledge of these retrieval mechanisms should lead to application of more effective indexing metadata and the development of better search strategies. Interestingly, only one institution of the twenty-five from which we took finding aids for this study supplied any type of metatags to the electronic versions of these finding aids.

Related to the retrieval effectiveness of the various search engines is the fact that even the poor performers retrieved some finding aids that none of the other systems retrieved. None of the engines retrieved all of the finding aids, while each engine retrieved some unique items. This indicates the efficacy of searching multiple engines for any given query.

Finally, it is important to realize that the search queries were optimized for the very best retrieval results in each search engine. What we see here is as good at it gets! We had such relatively high retrieval results because we started with the finding aids in our hands and searched for the titles and specific key terms using phrases. We looked for key terms that would not only locate a given finding aid, but distinguish it from the mass of materials on the Web. It is extremely unlikely that real users looking for archival materials would ever have this much information at the beginning of a search (why would they need to search if they already had the finding aid?) or would ever have this much success looking for particular collections.

So what is an archivist, or archives user to do? When searching the Web today for finding aids, probably the best thing to do would be to use Google or Fast Search, learning the searching protocol for each engine very, very well. The next thing would be to use phrase searching whenever possible. Finally, it is important to be willing to search a number of engines if particular materials are desired. At the same time, archivists must start to place more finding aids and other materials on the Web, providing users with remote access, and become expert searchers so that they can provide searching instruction and aid to their users.