PEASE AWARD

Retrieval of Archival Finding Aids Using World-Wide-Web Search Engines

Kathleen Feeney

Abstract

This article describes a study of the retrieval of on-line archival finding aids using two Internet search engines. The study was conducted to assess the value of electronic full-text finding aids as tools for locating archival collections. Topical subject headings and personal name headings were chosen from on-line inventories produced by the Southern Historical Collection (SHC) at the University of North Carolina at Chapel Hill. Keyword, phrase, and Boolean searches were developed using the topical subject headings; keyword and phrase searches were conducted for each personal name. The first one hundred documents retrieved by each search were examined to determine the number of SHC finding aids retrieved. The study found that searches often retrieved unmanageably large result sets and that the majority of SHC finding aids containing the search terms were not among the first one hundred documents retrieved. These results suggest that on-line archival descriptions are often not accessible through common methods of Internet searching and that archivists must continue to develop means to help researchers locate their collections.

Kathleen Feeney was awarded the 1999 Theodore Calvin Pease Award for this paper, which was written for a course taught by Helen Tibbo in the School of Information and Library Science at the University of North Carolina at Chapel Hill. The Pease Award is named for the first editor of the American Archivist *and is given to the best student paper as judged by the Pease Award Committee chaired by the current editor of the* American Archivist. *The 1999 selection committee consisted of Philip B. Eppard, Kathleen Roe, Robert Shuster, and Ken Thibodeau. The award was presented on August 26, 1999, in Pittsburgh at the 63rd annual meeting of the Society of American Archivists.*

Introduction

Archivists have long sought to increase the visibility and accessibility of their collections, to make information about their holdings available to those physically distant from a repository, and to ensure that potential researchers are aware of relevant archival resources. Traditionally, these goals have been met by publication of subject and collection guides, inclusion in directories and bibliographies, and reliance on researchers' efforts to locate collections through provenance-based "detective work." While these tactics have often been quite successful, they suffer from many liabilities; printed guides are expensive to produce and to update, and researchers needing materials scattered across multiple collections or held in unexpected places face a daunting task. In recent decades, however, developments in computer and network technology have offered archivists many new means of making their holdings information accessible to a broad array of remote users and of making this information electronically searchable.

The first major step in reaching this goal was the introduction of the USMARC Format for Archival and Manuscripts Control (MARC AMC), allowing archival collections to be cataloged in a format suitable for inclusion in bibliographic utilities such as OCLC.[1] Publication of Steven Hensen's *Archives, Personal Papers, and Manuscripts (APPM)* in 1983 further encouraged archival adaptation of electronic cataloging tools. Revised and promoted as a standard in 1989 by the Society of American Archivists,[2] *APPM* addresses inadequacies in standard library cataloging rules, developed for classification and description of published materials such as books rather than for unique materials such as archival documents and manuscripts. Development of computer network technology and archivists' desire to produce electronic descriptive documents that can extend beyond the information that can be contained in MARC AMC records has led to the adaptation of traditional archival finding aids, such as inventories, registers, indexes, and repository guides, to electronic formats, mounted locally or available worldwide via the Internet.

Significant effort has been devoted in recent years to the development and propagation of a standard for encoding electronic archival finding aids. The resulting Encoded Archival Description (EAD) is intended to provide repositories with a means of establishing an effective, accessible, and stable presence for their holdings information. EAD accommodates variations in the length and content of finding aids within and among repositories, and preserves in electronic form the complex, hierarchically structured descriptive information

[1] Throughout this paper, the term "archival collections" will be used to refer to both archival and manuscript collections.

[2] Steven L. Hensen, *Archives, Personal Papers, and Manuscripts: A Cataloging Manual for Archival Repositories, Historical Societies and Manuscript Libraries* (Washington, D. C.: Manuscripts Division, Library of Congress, 1983); 2d ed. (Chicago: Society of American Archivists, 1989).

found in archival repositories and registers, while also enabling the documents to be navigated and searched in ways that their printed counterparts cannot.[3]

The term "finding aid" is used by archivists to describe a wide range of descriptive resources, ranging from subject cards in a card catalog to guides surveying the whole of a large repository's holdings. The finding aid most central to archival description and access, however, is the inventory, a detailed description of a collection and its creating agency. Inventories differ among repositories, but most include an introduction or abstract; a history or biography of the collection's originating agency or individuals; a scope note detailing the size, contents, media, and arrangement of the collection; descriptions of the series subdivisions within the collection; container-level (box or folder) listings of the materials in each series; and an index or list of subject headings used to describe the collection. Inventories vary in length from a few paragraphs to hundreds of pages, depending upon the size and complexity of a collection. According to *APPM,* catalog entries, whether in the form of MARC AMC records or catalog cards, are created from the inventory and can ideally be described as an abstract of a complete inventory. An inventory offers a significantly more complete description of a collection than does a catalog record and, in a searchable electronic format, offers a greater number of access points to the collection. The two formats also differ in that subject terms in a catalog record are generally drawn from a predefined, controlled vocabulary, while an inventory or register may contain a much broader array of descriptive text. Finding aids were developed to meet the unique needs of archival collections, which may include numerous materials in diverse media, created and used for many different purposes. Unlike a book, which may generally be succinctly described within a traditional catalog record, an archival collection may encompass a broad array of divergent and unrelated subjects.

Electronic finding aids range from text files scanned from a repository's paper inventories to documents created electronically and encoded according to EAD standards. They may be made available within an institution in an internal database and/or worldwide via the Internet and may be linked to a MARC AMC record or to other resources describing the collection.

Archival repositories of all sizes and affiliations currently devote significant and increasing quantities of personnel and financial resources to the creation of electronic records of their holdings information. These archivists have benefitted from advances in network, database and searching technology and from the development of electronic formats and standards such as EAD. Yet, many questions regarding the accessibility and effectiveness of electronic archival finding aids remain largely unexplored. What are the goals of archivists in creating

[3] The EAD homepage, <http://lcweb.loc.gov/ead>. EAD is drawn from Standard Generalized Markup Language (SGML), a "set of rules for defining and expressing the logical structure of documents which enables software products to control the searching, retrieval and structured display of those documents" <http://lcweb.loc.gov/ead/eadback.html>.

these resources? Are they intended to supplement or to replace MARC records and other descriptive materials? Are they expected to serve as pointers to a repository or as detailed resources for researchers who have already identified the repository through other means? Are finding aids available on the Internet intended to make a repository's resources available to users unfamiliar with traditional means of archival searching? And perhaps most importantly, how will researchers locate an archives' on-line finding aids? Will commonly used tools for searching the Internet lead users to archival resources? This project is an effort to address these final questions by assessing the ease with which on-line archival finding aids may be found by searchers using the most popular and accessible search engines for the World Wide Web.

Literature Review

Research related to electronic archival holdings information has focused mostly on the creation of MARC AMC records and on researchers' efforts to retrieve these records from databases. These studies reveal both the limitations of MARC records as descriptive tools and the difficulties inherent in searching large, heterogeneous collections of electronic records. Avra Michelson's 1986 study of descriptive practices by repositories inputting records into the Research Library Information Network's (RLIN) AMC file found "massive inconsistency" both within and among repositories in their choice and construction of subject headings and their treatment of out-of-scope materials and a dearth of authority control in the creation of personal and corporate name terms.[4] In Michelson's survey of twenty-one repositories on their choice of access points for a collection of family papers, survey respondents collectively chose 162 different descriptive terms; no term was chosen by all of the repositories.[5] Michelson advocates further study of user queries in an effort to determine the sorts of access points most useful to archival researchers, the direction of resources to the development of an "archival science of searching" suited to the specifications of archival description, and efforts to develop a more consistent descriptive vocabulary designed for inclusion in automated systems.[6] Michelson further argues that because the "heterogeneous" nature of archival collections leads them to require a much greater number of index terms than books, indexing and searching methods and technologies developed for libraries might be unsuitable for the unique needs of archival holdings information.

[4] Avra Michelson, "Description and Reference in the Age of Automation," *American Archivist* 50 (Spring 1987): 192–208.

[5] Michelson, "Description and Reference in the Age of Automation," 194.

[6] Michelson, "Description and Reference in the Age of Automation," 195–99.

Two articles written in the wake of Michelson's study differ in their interpretations of the implications of the inconsistencies she uncovered. David Bearman concludes that authority control cannot be effectively established over topical subject index terms, and that archivists may best enhance the retrievability of holdings information by use of increased numbers of access points, supplementing subject headings with geographical, chronological, and form-of-material terms.[7] Bearman argues further that the subject content of archival materials cannot be accurately analyzed because, unlike books, archival documents were created to serve a function, rather than to discuss a subject. Jackie M. Dooley disagrees with Bearman's conclusion that consistency in subject indexing is "unattainable," arguing instead that subject headings are a valuable, if imperfect, means of access to collections and that archivists should strive for the greatest consistency possible in the use of terminology in catalog records and finding aids.[8]

Traditionally, subject access to archival collections has been obtained through published subject or repository guides. Dooley claims that proper electronic description can recreate this sort of access in an easily updatable format, she further argues that user queries often require integrated access to archival, bibliographic, and other sorts of informational materials, a need best met by inclusion of holdings information in on-line finding aids or bibliographic databases. Consistency and predictability in indexing practices is particularly necessary for Internet-based searches, writes Dooley, because these searches are very often "unmediated by reference archivists."

Helen Tibbo's study of the retrieval of archival records entered in the OCLC Online Union Catalog by use of subject terms also focuses on searches of MARC AMC records rather than on-line finding aids, but is relevant to this study because of its discussion of the problem of extremely large retrieval sets in free-text searches.[9] Tibbo found that archival records are often digital needles buried in an electronic haystack because the terms commonly used to index and search for them appear in bibliographic databases in far greater numbers than a researcher can browse effectively. She suggests that indexing practices suitable for records entered in local databases, such as the catalog for an individual library, are often insufficiently "specific, appropriate and exhaustive" for effective retrieval in large databases like OCLC.[10]

Among the few published efforts to explore the respective roles of MARC AMC records and full-text electronic finding aids in describing and directing researchers toward archival collections is Rita Czeck's comparison of the contents

[7] David Bearman, "Authority Control Issues and Prospects," *American Archivist* 52 (Summer 1989): 286–99.

[8] Jackie M. Dooley, "Subject Indexing in Context," *American Archivist* 55 (Spring 1992): 344–54.

[9] Helen R. Tibbo, "The Epic Struggle: Subject Retrieval From Large Bibliographic Databases," *American Archivist* 57 (Spring 1994): 310–26.

[10] Tibbo, "The Epic Struggle," 322.

of the two sorts of descriptive records for a sample of twenty Southern Historical Collection collections.[11] Czeck evaluated the extent to which MARC records included the geographical, chronological, personal name, and corporate name subject headings found in the corresponding full-text finding aid for each collection. She found the MARC records to be inconsistent in their inclusion of the most significant subject terms from full-text finding aids. She argues that users will benefit greatly from the ability to access collections by searching electronic inventories, giving them access to a greater number of subject terms and to information from the portions of finding aids generally not included in MARC records.

Little research has been done into identification of the intentions and expectations of archivists who create electronic finding aids, or into the experiences of archivists in their development and implementation, or of researchers in their use. Much of the literature published on full-text electronic description has focused on the technical aspects of such an undertaking, such as the development of standards for Encoded Archival Description (EAD). Articles focusing on repositories' experiences with on-line description, such as Stuart Glogoff and Kristin A. Antelman's discussion of a project centering on a collection of congressional papers, suggest that Internet-based finding aids are expected to provide "global access" to the collection and to enable the repository to "overcome some of the long-standing problems associated with bibliographic access to manuscript and archival materials,"[12] including the fact that many archival collections are not cataloged in any sort of on-line catalog and that catalog records generally do not contain sufficient information to fully describe archival materials. Glogoff and Antelman also discuss ways in which technology can increase the usefulness of traditional finding aids by making them electronically searchable, linking them to digitized documents or multimedia presentations, and displaying the finding aid in a format accessible and attractive to both novice and experienced researchers. Nevertheless, the article also stresses the importance of linking finding aids to MARC records in a library's OPAC and providing a means of searching the finding aids locally.

Project Description and Methodology

This study is based on finding aids describing the collections of the Southern Historical Collection (SHC) in the Manuscripts Department of the University of North Carolina at Chapel Hill. At the time the study was completed, inventories of 1,671 of the Southern Historical Collection's 4,471 collections were available on-line, mostly in the form of text files scanned from

[11] Rita L. H. Czeck, "Archival MARC Records and Finding Aids in the Context of End-User Subject Access to Archival Collections," *American Archivist* 61 (Fall 1998): 426–40.

[12] Stuart Glogoff and Kristin A. Antelman, "Relieving Archival Gridlock: Congressional Archives on the Web," *Internet Reference Services Quarterly* 2, no. 1 (1997): 39–50.

paper documents. The inventories are accessible through a title index on the SHC's home page or via links from MARC records in the university's Internet-mounted library catalog. The finding aids were not publicly searchable from the SHC's site at the time that this study was completed.[13] The inventories follow a relatively, though not entirely, standardized format, beginning with an abstract giving broad details of the collection's provenance, contents, and structure; a list of terms under which the collection is included in the University's on-line catalog; a biographical or historical note; descriptions of the series in the collection; and a container list.

While the Southern Historical Collection contains materials documenting almost all periods and aspects of Southern history, this study focused on the most well-known and highly developed segment of the collection—the papers of families and individuals dating back to the antebellum plantation, the Civil War, and the Reconstruction-era South. A representative sampling of these collections was obtained by a keyword search in the University Library's on-line catalog.[14] The search retrieved 119 MARC records, 90 of which were linked to on-line finding aids. The twenty-five topical subject terms and personal names used for searches in this study were chosen from the list of "Online Catalog Terms" in these ninety finding aids. The terms were selected by the researcher, rather than being chosen randomly, because it was deemed desirable to select searches that varied in length and structure and to avoid the inclusion of many similar terms, such as "Slavery—North Carolina" and "Slavery—South Carolina."

Although Library of Congress Subject Headings are unlikely to be used by a large number of Internet searchers, they were suitable for this study for several reasons. First, they are likely to be common to many finding aids in the Southern Historical Collection. Second, they are included in the MARC record for each collection, allowing the researcher to use a search of the library catalog to determine the number of finding aids containing each term. Third, they may be adapted to more likely user searches through use of Boolean search logic. Finally, a search on a complex inverted subject heading exactly as it is entered in a catalog record is likely to produce a small result set consisting primarily of archival and library materials, allowing the researcher to determine if a particular finding aid has been indexed by a particular search engine.

Two search engines were used in this study. Multiple search engines were used because no single engine indexes more than a portion of the total number of documents available on the Internet. Search results using identical search terms often vary widely among search engines. HotBot claims to be the largest search engine on the Internet, indexing approximately 110 million web pages. AltaVista indexes about 100 million pages and is probably the most

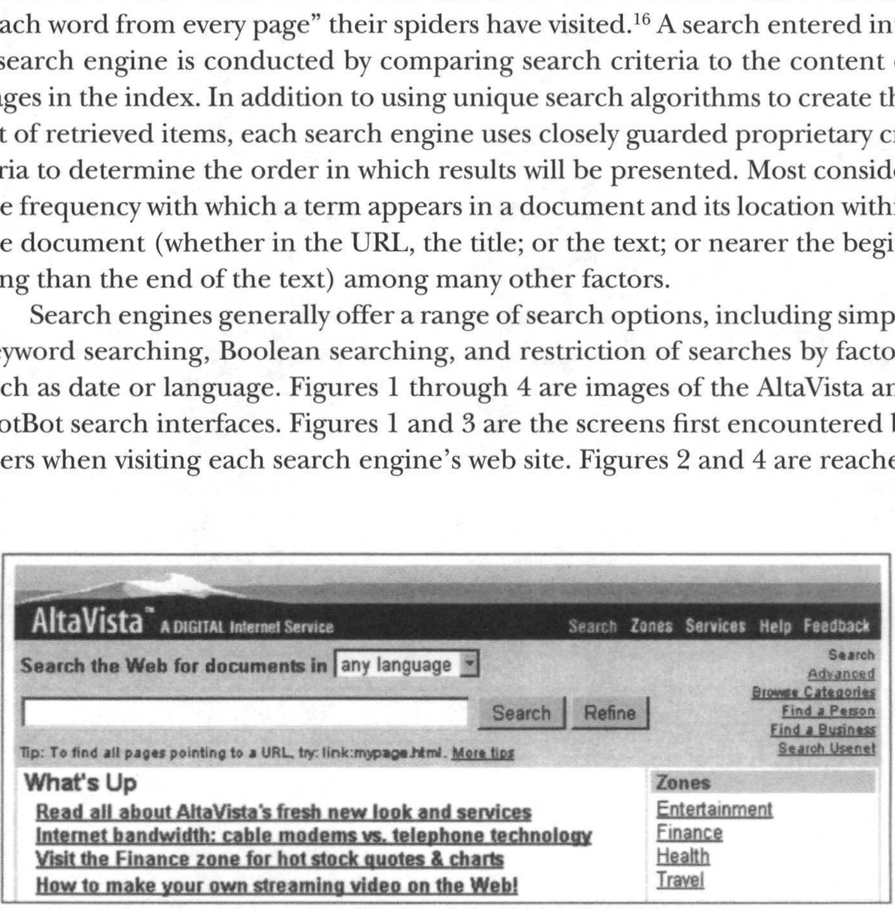
[13] The Manuscripts Collection, University of North Carolina at Chapel Hill, <http://www.lib.unc.edu/mss/inv.html> (accessed April 25, 1998).

[14] (((PLANTATION) AND (MANUSCRIP?)) AND (SLAV?)) AND (SOUTHERN HISTORICAL COLLECTION)) NOT (SERIES)

widely recognized search engine among Internet users.[15] A third search engine, Excite, was eliminated from the study after a pilot study revealed that it did not index any of the Southern Historical Collection's Internet publications.

Although the precise means by which search engines index pages and rank search results are considered by their creators to be confidential and proprietary, most use fairly similar programs and technology. Resources such as HotBot and AltaVista are created by programs called "robots" or "spiders" that constantly "crawl" the Internet, following links and adding copies of web pages to the search engine's index. Spiders regularly visit previously crawled pages, re-indexing those that have been altered and deleting from their records those that have been eliminated. Indexing and updating of a page may occur over the course of several months. Both AltaVista and HotBot claim to have indexed "each word from every page" their spiders have visited.[16] A search entered into a search engine is conducted by comparing search criteria to the content of pages in the index. In addition to using unique search algorithms to create the set of retrieved items, each search engine uses closely guarded proprietary criteria to determine the order in which results will be presented. Most consider the frequency with which a term appears in a document and its location within the document (whether in the URL, the title; or the text; or nearer the beginning than the end of the text) among many other factors.

Search engines generally offer a range of search options, including simple keyword searching, Boolean searching, and restriction of searches by factors such as date or language. Figures 1 through 4 are images of the AltaVista and HotBot search interfaces. Figures 1 and 3 are the screens first encountered by users when visiting each search engine's web site. Figures 2 and 4 are reached



Search the Web for documents in any language

Tip: To find all pages pointing to a URL, try: link:mypage.html. More tips

What's Up

Read all about AltaVista's fresh new look and services
Internet bandwidth: cable modems vs. telephone technology
Visit the Finance zone for hot stock quotes & charts
How to make your own streaming video on the Web!

Zones

Entertainment
Finance
Health
Travel

FIGURE 1. AltaVista Default Search Interface

[15] Information on search engines is drawn from the help pages of HotBot and Altavista and from Search Engine Watch, <http://www.searchenginewatch.com>.

[16] <http://www.altavista.digital.com/av/content/about_our_technology.htm>; <http://www.hotbot.com/help>.

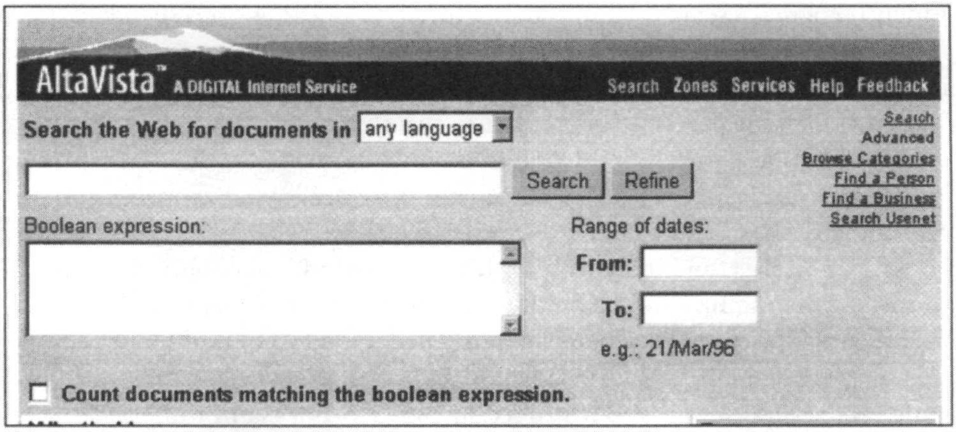

AltaVista™ A DIGITAL Internet Service
Search Zones Services Help Feedback
Search the Web for documents in any language
Search
Refine
Boolean expression:
Range of dates:
From:
To:
e.g.: 21/Mar/96
□ Count documents matching the boolean expression.
Search
Advanced
Browse Categories
Find a Person
Find a Business
Search Usenet

FIGURE 2. AltaVista Advanced Search Interface

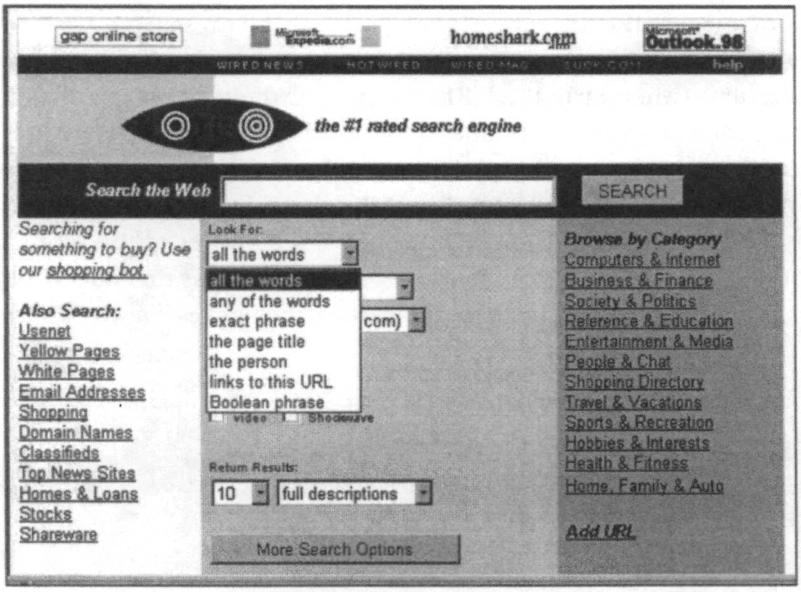
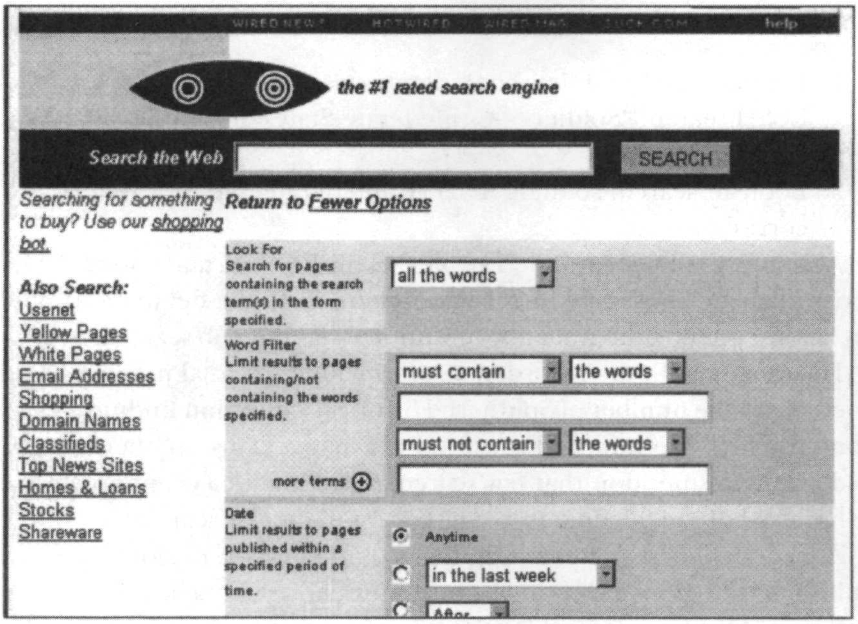


FIGURE 3. HotBot Default Search Interface

by users who choose the "Advanced" option in AltaVista or the "More Search Options" link in HotBot.

The first set of searches in this study was based upon twenty-five topical subject headings taken from the ninety finding aids discussed above. The subject headings ranged in length from a single word to more complex listings including up to five terms. Each of the headings was first entered into DRA, the University Library's on-line catalog, to determine the number of electronic



Searching for something to buy? Use our shopping bot.

Return to Fewer Options

Look For
Search for pages containing the search term(s) in the form specified.
all the words

Word Filter
Limit results to pages containing/not containing the words specified.
must contain | the words
must not contain | the words

Date
Limit results to pages published within a specified period of time.
● Anytime
○ in the last week
○ After

FIGURE 4. HotBot Advanced Search Interface

finding aids at the University in which it appears.[17] Then, three searches were performed on each subject heading in each search engine. The first search was done according to the default search option presented to the user upon first encountering the search engine. In AltaVista, this is a keyword search for any of the terms entered in the search box. In HotBot, the default search is described as one that looks for "all the words," or a keyword search for documents that contain all of the words entered in the search box, but does not require them to be in a phrase or in proximity to one another. All terms were uncapitalized in the default searches; both AltaVista and HotBot will search for both the capitalized and uncapitalized form of a term that is not capitalized by the searcher. In the second topical subject heading search, terms were entered as they appear in the finding aid, including capitalization and punctuation. In AltaVista, the search terms were enclosed within quotation marks, while in HotBot, the searcher chose the "exact phrase" search option. The third search took advantage of the "advanced" features of the search engines, treating multiple word terms such as "fugitive slaves" as exact phrases that must be found in searched documents, and entering Boolean operators between distinct terms in a subject heading. In these searches, only proper nouns were capitalized. For example, the subject heading "Soldiers—Confederate States of America—

[17] The search was for <subject heading> AND Southern Historical Collection. This was done for each subject heading. The number of electronic finding aids was determined by the researcher's count among the results returned.

Correspondence" would have been entered into each search in the following formats:

1. Default search: soldiers confederate states of america correspondence
2. LCSH search: "Soldiers—Confederate States of America—Correspondence"
3. Boolean search: soldiers AND "Confederate States of America" AND correspondence

Although the LCSH search is atypical of the keyword and natural language queries likely to be used by Internet searchers, both the default and Boolean searches are intended to imitate conventional on-line web searches.

For each search, the recorded results include the total number of results retrieved and the number of Southern Historical Collection finding aids found among the first one hundred retrieved web pages. This number was chosen based on the assumption that few searchers are likely to examine more than one hundred of the documents retrieved in a particular search.

A second set of searches was performed upon twenty personal names used as subject headings in finding aids. The twenty names were taken from the same sampling of finding aids used for the topical subject heading searches. A pilot study was done to determine the proper way in which to structure name queries to each search engine.[18] Two searches were performed upon each name in each search engine. The first was identical to the default searches described above for topical subject terms. The second search instructed the search engines to search for personal names appearing either in inverted form ("Little, Rufus") as they would appear as subject headings, or in traditional order ("Rufus Little") as they would appear in other portions of a finding aid. In AltaVista, this search was structured as a Boolean query on the "Advanced Search" screen. ("Little, Rufus" OR "Rufus Little") In HotBot, the same effect was achieved by simply entering the name and instructing the program to search for "the person." (The pilot study demonstrated the ability of this search feature to correctly manipulate personal names consisting of multiple names, such as "Eliza Anne Marsh Robertson.") For each search, the researcher recorded the total number of documents found and noted whether or not the finding aid from which the personal name was taken appeared among the first one hundred search results. A note was also made of other pointers to the Southern Historical Collection that appeared among the first one hundred results, such as a finding aid other than the one from which the name was taken or a subject guide to the manuscript collection.

[18] For example, the default search in Altavista disregards single letter terms, such as initials, while the HotBot default search includes initials among the searched terms, but disregards punctuation.

Findings

Topical Subject Heading Searches

The searches on topical subject headings proved to be largely unsuccessful in directing the searcher toward the Southern Historical Collection holdings. Complete results of the searches may be found in Tables 6 and 7 in the appendix. As shown in Table 1 below, only ten out of seventy-five AltaVista searches included SHC finding aids among their first one hundred results; in HotBot, only thirty-six out of seventy-five searches retrieved any of the desired documents.

The two search engines produced very different result sets. In addition to retrieving SHC inventories in a larger portion of total searches than did AltaVista, the HotBot search engine generally produced larger sets of successful results. There appears to be little congruence between the retrieval sets found by the two search engines. There were eight subject headings for which both HotBot and AltaVista retrieved SHC inventories with at least one of the three searches built upon each heading. In only three of these cases were all of the SHC inventories retrieved by AltaVista also retrieved by HotBot. In each case, the HotBot searches retrieved a greater number of SHC inventories than did AltaVista. HotBot retrieved SHC inventories with searches for nine additional subject headings, for which AltaVista retrieved no relevant entries.

More importantly, few of the searches retrieved more than a small portion of the total number of on-line SHC finding aids containing the relevant subject headings. Only thirteen searches retrieved a total number of inventories greater than 30 percent of the total available inventories for the relevant heading.[19] It must be noted that a direct comparison between the total number of on-line SHC inventories containing a particular subject heading and the number of SHC finding aids retrieved in a particular search is very rough, and greatly biased toward retrieving all relevant finding aids, because of the structure of the searches. For example, a default or Boolean search based on the subject heading "Reconstruction—Tennessee" would be likely to retrieve inventories containing

Table 1. Number of Searches Retrieving SHC Inventories

Search	No. of searches retrieving SHC inventories
HotBot Default	9
HotBot Advanced	17
HotBot Advanced II	10
AltaVista Default	1
AltaVista Advanced	6
AltaVista Advanced II	3

[19] All but one were HotBot searches.

the headings "Reconstruction—Alabama" and "Diaries—Tennessee" as well as the heading upon which the search is based. Also, phrases used as subject headings may appear elsewhere within the text of a finding aid. "Fugitive slaves" and "Plantation life" are prominent examples. The ability to search for finding aids on the basis of terms beyond those chosen for subject indexing is, of course, considered to be one of the potential strengths of on-line, searchable archival holdings information. Although today's Web search engines should retrieve all relevant SHC finding aids based on these subject heading searches, in actuality, only 46 of the 150 total searches performed on topical subject headings provided the searcher with a pointer to Southern Historical Collection holdings in the first one hundred hits, while most of the searches that did retrieve SHC inventories retrieved only a few of the relevant documents in this number of hits.

There appear to be two barriers to effective retrieval of on-line finding aids through topical subject heading searches. The first is the prevalence of large retrieval sets, particularly for default searches. In AltaVista, all but two of the default searches produced retrieval sets of over 300,000 items; fourteen of the searches retrieved more than one million items. The SHC documents could be anywhere in those one million hits, but are unlikely to be discovered by the searcher. The HotBot default searches were more restrictive than those in AltaVista, searching for documents containing *all* of the keywords entered, rather than those including *any* of the search terms. Nevertheless, only four of the searches retrieved fewer than one thousand items. Retrieval sets were smallest for the subject heading searches. Only six AltaVista searches and six HotBot queries retrieved more than one hundred web pages. Boolean searches in AltaVista produced retrieval sets ranging in size from 25 to 11,366 listings; in HotBot, Boolean searches resulted in retrieval sets ranging from 1 to 13,860 listings, with twelve of the twenty-five searches retrieving more than one thousand items. In short, while we may surmise that SHC finding aids were among the thousands of hits produced by each search, very few searches based on topical subject headings produced retrieval sets that are likely to be examined in full by a typical searcher. It is also important to note that many of the SHC finding aids appearing among the top one hundred results for a search were ranked in the lower portion of that list, often appearing among search results fifty-one through one hundred. Table 2 shows the number of finding aids for each

Table 2. Ranking of Retrieved Finding Aids

	AltaVista Default	AltaVista LCSH	AltaVista Boolean	HotBot Default	HotBot LCSH	HotBot Boolean
Total Retrieved	1	18	3	126	96	174
Top 10	0	9	0	18	52	21
Top 25	0	18	2	40	87	52
Top 50	0	18	3	78	95	114

search type that appeared in the top half, quarter, and tenth of the examined search results.

With the exception of the Library of Congress Subject Heading searches, for which the result sets often contained fewer than twenty items, it is clear that only a small portion of the desired search items were found near the top of the list of results.

The nature of non-Southern Historical Collection materials retrieved by each search varied widely. Some of the retrieved documents were clearly relevant to the search performed, including finding aids to manuscript collections held by repositories other than the SHC, documents other than inventories describing SHC holdings, and other resources closely related to the subject heading used to develop the search. Other retrieved items could be related only tenuously, if at all, to the terms searched. For example, while the majority of the items retrieved by searches on the term "American Colonization Society" were easily recognized as relevant to the history of the Society, a large portion of the citations retrieved by searches on the term "Plantation life" referred to web sites related to travel and recreation in the contemporary southern United States. The problem of large retrieval sets may be attributed in part to retrieval of irrelevant items by search engines, but it is created in large measure by the quantity and variety of information available on-line.

The second factor limiting the effectiveness of these searches is the likelihood that many of the Southern Historical Collection's on-line finding aids have not been indexed by search engines. Although most retrieval sets were too large to be examined in full, a significant number included fewer than one hundred listings, allowing the researcher to ascertain that they did not include all of the relevant SHC inventories. As explained previously, search engine indexes are created by programs that continuously and automatically "crawl" the Internet, indexing the web pages that they encounter. It seems likely that a large portion of the archival finding aids available on-line have not yet been encountered and indexed by the most widely used Internet search resources.

In an effort to increase the number of relevant documents retrieved, default, Library of Congress Subject Heading, and Boolean searches were repeated on a subset of five subject headings, with the addition of the phrase "Southern Historical Collection" to each search. The phrase is found in the heading of each finding aid. Such a search is, of course, much more precise than is likely to be performed by an actual Internet searcher. A researcher already aware that the documents he or she seeks are in the Southern Historical Collection is likely to go directly to the collection's website, or to contact the repository by other means. In the absence of a local search utility, however, a researcher might resort to the use of Internet search engines to search across the finding aids in a particular location. These specialized searches produced the results presented in Table 3. The first set of results listed for each search engine, HotBot I and AltaVista I, refer to the original searches, the second,

Table 3. Documents Retrieved by Searches Including Phrase "Southern Historical Collection"

North Carolina—History—Civil War

	Default		LCSH		Boolean	
	Total	SHC in top 100	Total	SHC in top 100	Total	SHC in top 100
HotBot I	27,030	0	12	7	13,860	0
HotBot II	299	37	7	7	290	33
AltaVista I	6,456,420	0	6	2	11,942	0
AltaVista II	462	2	0	0	186	4

N = 50[20]

American Colonization Society

	Default		LCSH		Boolean	
	Total	SHC in top 100	Total	SHC in top 100	Total	SHC in top 100
HotBot I	10,661	0	595	0	595	0
HotBot II	28	2	16	2	16	2
AltaVista I	6,110,570	0	513	0	513	0
AltaVista II	386	0	1	0	1	0

N = 7

Soldiers—Confederate States of America—Correspondence

	Default		LCSH		Boolean	
	Total	SHC in top 100	Total	SHC in top 100	Total	SHC in top 100
HotBot I	835	15	13	13	0	0
HotBot II	91	36	13	13	8	0
AltaVista I	858,780	0	0	0	229	0
AltaVista II	313	9	0	0	13	1

N = 72

Women—Confederate States of America

	Default		LCSH		Boolean	
	Total	SHC in top 100	Total	SHC in top 100	Total	SHC in top 100
HotBot I	5074	1	9	1	765	2
HotBot II	139	36	2	1	83	38
AltaVista I	561,560	0	15	0	895	0
AltaVista II	340	0	0	0	41	3

N = 18

[20] N = number of SHC finding aids containing the subject heading entered.

Table 3. *(continued)* Documents Retrieved by Searches Including Phrase "Southern Historical Collection"

Slavery—North Carolina

	Default		LCSH		Boolean	
	Total	SHC in top 100	Total	SHC in top 100	Total	SHC in top 100
HotBot I	10,088	0	29	21	7605	0
HotBot II	183	51	25	21	181	51
AltaVista I	361,000	0	1	0	1623	0
AltaVista II	332	83	0	0	134	4

N = 73

HotBot II and AltaVista II, to searches including the phrase "Southern Historical Collection."

As was expected, the specialized searches greatly reduced the total number of items retrieved, in two cases reducing the retrieval set from more than six million to fewer than five hundred items. In doing so, these searches appear to have increased the number of SHC finding aids included among the first one hundred citations. In the AltaVista default search for "North Carolina—History—Civil War," for example, it is likely that the two finding aids retrieved in the specialized search were present in the result set for the first search on the terms, but only appeared among the top one hundred results when the result set was reduced from 6.4 million to 462 items. Nevertheless, the results of the specialized searches support the above conclusions concerning the difficulty of locating archival finding aids by use of Internet search engines. Nine of the thirty specialized searches did not include any SHC finding aids among the first one hundred items listed; only three found finding aids in a quantity equaling that available among SHC electronic finding aids containing the relevant subject heading. These were the default and Boolean searches in HotBot for "Women—Confederate States of America," and the default search in AltaVista for "Slavery—North Carolina." All of these searches retrieved a greater number of SHC finding aids than those which actually contain the relevant subject heading. This may be easily attributed to the fact that the individual terms "women," "Confederate States of America," "slavery," and "North Carolina" appear frequently throughout many of the SHC's electronic finding aids. Of these three searches, it is notable that the HotBot Library of Congress Subject Heading search for "Women—Confederate States of America" retrieved only one finding aid while the AltaVista LCSH search for "Slavery—North Carolina" found none. In short, even when searches are customized to favor items from a particular repository, only a small portion of relevant on-line documents are retrieved through Internet search engine searches.

Personal Name Searches

The searches based on personal names used as subject headings in SHC finding aids was conducted according to standards slightly different than those based on topical subject headings. Rather than being used to calculate the number of relevant finding aids retrieved, search results were examined to determine whether or not they contained the particular finding aid from which a personal name heading was taken. The results, found in Table 8 in the appendix, were suggestive of the same trends found for the topical subject heading searches. The default searches in both search engines produced unmanageably large retrieval sets. Only one AltaVista default search retrieved fewer than 100,000 items. HotBot default retrieval sets were considerably smaller; nevertheless, sixteen of the twenty total searches produced more than one hundred web page citations.

The advanced searches, which instructed each search engine to look for both the inverted and traditional form of a name, provided relatively small retrieval sets, allowing the researcher to determine with certainty whether each search had retrieved the desired document. Two of the HotBot "name" searches produced more than four thousand results. These searches, for "William Mercer Green" and "Peter Boyd Martin" included the relatively common names "William Green" and "Peter Martin" among their search options. However, the search engine appears to have given extra weight to documents containing the complete name as entered by the searcher. In both cases, the desired finding aids appeared in the first position in the list of search results. Nevertheless, as indicated in Table 4, the majority of the searches did not locate the desired finding aids. The most successful search type, the HotBot advanced "name" search included only 50 percent of the desired finding aids among its results; other search types were considerably less successful.

Another significant trait of the results found for personal name searches was the prominence of additional pointers to the Southern Historical Collection among the web pages listed. These additional resources included:

- The on-line alphabetical index to SHC finding aids, titled "Index of /mss/inv/". This document is a list of the titles of on-line finding aids, linked to the inventories themselves. Search results might include the index if the personal name used in the search is included in the title of a collection.[21]
- *Records of Ante-Bellum Southern Plantations from the Revolution Through the Civil War* (1996). This is an on-line guide to a set of microfilm reproductions of Southern Historical Collection manuscript collections produced by University Publications of America. It includes adaptations of SHC finding aids, whose content is very close to that of the original documents. Although this is neither an SHC publication nor, strictly speak-

[21] <http://www.lib.unc.edu/mss/inv/>.

Table 4. Percentage of Personal Name Searches Retrieving SHC Finding Aids

	AltaVista Default	AltaVista Advanced	HotBot Default	HotBot Advanced
Names found in first 100 search results (out of 20 searches)	0%	10%	20%	50%

Table 5. Additional References to SHC Retrieved in Personal Name Searches

	AltaVista Default	AltaVista Advanced	HotBot Default	HotBot Advanced
SHC finding aid index	0	0	4	6
Microfilm index	1	15	3	11
Other SHC finding aid	0	2	2	2
Business History Resources	2	2	2	2
African-American Resources	1	3	0	0

ing, a direct pointer to the repository itself, it is cited here because of the frequency with which it appears in search results. It does not, of course, include links to the Southern Historical Collection or its holdings information.[22]

- An SHC finding aid other than the one from which a personal name was taken.
- An on-line index to resources for the study of business history in the Southern Historical Collection, produced by the SHC and linked to relevant finding aids.[23]
- A guide to resources in North Carolina repositories suitable for the study of African-American history, compiled by an SHC staff member.[24]

The frequency with which these resources appeared among search results is indicated in Table 5. Their prevalence suggests alternate means by which on-line archival holdings information may be made accessible to Internet searchers.

Conclusions

The findings of this study suggest that the Internet is too large and heterogeneous a search ground in which to locate archival holdings information by commonly used Internet searching practices and tools. The problems of searching for archival materials in large bibliographic databases, including

[22] Originally found at <http://www.asprs.org/upa/guides/plantj1.htm>. Currently available at <http://www.lexis-nexis.com/cispubs/guides/southern_hist/south.htm>.

[23] *Selected Business History Resources in the Southern Historical Collection,* <http://www.lib.unc.edu/mss/bushist.htm>.

[24] Timothy D. Pyatt, ed., *Guide to African-American Documentary Resources in North Carolina* (University Press of Virginia, 1996); available at <http://www.upress.virginia.edu/epub/pyatt/nchome.html>.

inconsistent use of terminology and enormous retrieval sets, are magnified and compounded on the Internet. The findings also suggest that electronic finding aids may not be well suited to serve as pointers to archival collections. Instead, they may be most appropriate as resources for researchers who have already located a repository's on-line resources through other means.

While the ability to place archival holdings information on the Internet, and to make that information electronically searchable is certainly an invaluable development for archives and their users, this study provides evidence of the continuing necessity of expertise and professional assistance in the traditional means of locating archival materials, whether through a logical, provenance-based search or through the use of subject guides. These sorts of searches may be conducted in both print and electronic environments. Repository and subject guides may be made available on-line, and categorically organized resources such as Internet directories like Yahoo[25] may be valuable tools in a provenance-based search by allowing a user to connect electronically to specific repositories.

This study provides indirect support for the creation of cooperative databases of archival finding aids, allowing researchers to conduct full-text searches across a large collection of documents provided by multiple repositories, but limited to archival holdings information. On-line subject guides and archival clearinghouses are also likely to be invaluable tools for researchers searching for holdings information on-line. Additionally, the study suggests that researchers will benefit from studies to determine the most effective means of searching and the development of searching technologies specifically adapted to the structure of on-line archival resources. Also, the study supports the continued creation of MARC records linked to electronic finding aids, suggesting that, despite the problems of retrieval of relevant records and manageable search sets from bibliographic utilities like OCLC and RLIN, MARC records remain a more valuable and reliable means of locating archival resources than do lengthy, full-text documents, particularly given the increasing size of the World Wide Web. In fact, these findings suggest that continued research into more precise and sophisticated indexing and searching of MARC AMC records should remain a focus of archivists' efforts.

The results of the personal name searches, in which retrieval sets frequently contained pointers to Southern Historical Collection materials other than the finding aids from which the search terms were taken, suggest the value to archivists of creating electronic versions of resources such as repository guides and subject bibliographies. While these resources do not contain the complex and detailed descriptive information that makes collection inventories uniquely valuable, they may be better suited to the realities of Internet searching than are complete full-text inventories. A subject bibliography or repository guide may highlight the most prominent topics found within a col-

[25] <http://www.yahoo.com>.

lection and can use repetition of significant terms to increase its opportunity for a high rank among search results and may be used to direct a searcher to a repository's on-line resources, where a more complete and refined search may be conducted.

Electronic finding aids are certainly resources of great value for disseminating archival holdings information. They allow users to access remotely detailed information about collections and to search the text of this information. Additionally, they may be linked to digitized versions of collections, to records already available in electronic formats, or to other resources that may be of value to researchers. Nevertheless, the creation of on-line finding aids, most of which duplicate material already available on paper, requires the devotion of a significant amount of time, money, equipment, and expertise on the part of institutions that may find all of these resources to be in short supply. Archivists interested in the creation of electronic holdings information must clearly explore and define the needs that these resources are expected to fulfill and the ways in which they might replace, supplement, or duplicate other descriptive tools.

Table 6. AltaVista Topical Subject Heading Searches

		AltaVista Default		AltaVista Advanced		AltaVista Advanced II	
TERM	DRA	Total	SHC	Total	SHC	Total	SHC
American Colonization Society	7	6,110,570	0	447	0	n/a	n/a
American Party	3	4,950,355	0	1290	0	n/a	n/a
Confederate States of America	300+	431,980	0	3168	0	n/a	n/a
Confederate States of America. Navy	18	1,803,090	0	27	0	550	0
Cotton growing—North Carolina—History—19th century	4	7,865,440	0	0	0	25	0
Diaries—South Carolina	21	379,560	1	1	0	1565	0
Family—North Carolina—Social life and customs—19th century	238	9,625,880	0	0	0	218	0
Freedmen	46	3232	0	n/a	n/a	n/a	n/a
Fugitive slaves	10	62,160	0	1006	0	n/a	n/a
Horsebreeders—North Carolina—History—19th century	1	4,905,978	0	0	0	0	0
Hospitals, military—Confederate States of America	5	3,001,900	0	1	1	150	0
North Carolina—Economic conditions—19th century	33	8,642,290	0	2	2	283	0
North Carolina—History—civil war	50	6,456,420	0	9	2	11,366	0
Plantation life	59	492,910	0	993	0	n/a	n/a
Presbyterian Church—Southern States—*History—19th century*	1	6,652,350	0	0	0	51	1
Reconstruction—Tennessee	3	453,570	0	11	0	7208	0
Slave bills of sale	41	1,329,920	0	46	1	n/a	n/a
Slave records—North Carolina	19	4,134,070	0	1	1	41	1
Slave trade—United States	1	690,770	0	13	0	4317	0
Slavery—North Carolina	73	361,000	0	20	5	5469	0
Slaves—Medical care	6	4,815,860	0	1	0	1277	0
Soldiers—Confederate States *of America—Correspondence*	72	858,780	0	2	0	182	0
Tobacco—North Carolina—History—19th century	6	5,615,770	0	0	0	743	0
Women—Confederate States of America	18	561,560	0	18	0	761	0
Women—Education—Mississippi—History-19th century	2	3,082,230	0	1	0	2136	0

Table 7. HotBot Topical Subject Heading Searches

		HotBot Default		HotBot Advanced		HotBot Advanced II	
TERM	DRA	Total	SHC	Total	SHC	Total	SHC
American Colonization Society	7	10,661	0	595	0	595	0
American Party	3	562,674	0	1421	0	1421	0
Confederate States of America	100+	12,231	0	3123	1	3123	1
Confederate States of America. Navy	18	2822	0	17	1	601	0
Cotton growing—North Carolina—History—19th century	4	706	7	0	0	38	9
Diaries—South Carolina	21	3204	0	3	3	1,743	0
Family—North Carolina—Social life and customs—19th century	238	1006	66*	27	24*	158	85*
Freedmen	46	3749	0	n/a	n/a	n/a	n/a
Fugitive slaves	10	4386	0	1260	0	1260	0
Horsebreeders—North Carolina—History—19th century	1	1	0	0	0	1	0
Hospitals, military—Confederate States of America	5	603	0	1	1	110	3
North Carolina—Economic conditions—19th century	33	2788	25*	3	3	248	37*
North Carolina—History—civil war	50	27,030	0	12	7	13,860	0
Plantation life	59	38,943	0	1056	0	1056	0
Presbyterian Church—Southern States—History—19th century	1	1,069	1	1	1	39	2
Reconstruction—Tennessee	3	7738	0	8	1	7738	0
Slave bills of sale	41	2714	4	53	10*	53	10*
Slave records—North Carolina	19	4014	1	6	6	66	15
Slave trade—United States	1	17,577	0	14	0	5,086	0
Slavery—North Carolina	73	10,088	0	29	21*	7,605	0
Slaves—Medical care	6	8044	0	2	1*	1,501	1
Soldiers—Confederate States of America—Correspondence	72	835	15	13	13	0	0
Tobacco—North Carolina—History—19th century	6	1435	6*	1	1	874	6
Women—Confederate States of America	18	5074	1	9	1	765	2
Women—Education—Mississippi—History-19th century	2	3080	0	2	1	2,433	1

* Each of these result sets included duplicate citations of particular finding aids. The total given in this table excludes these duplicate listings.

Table 8. Personal Name Searches

	AltaVista Simple		AltaVista Advanced		HotBot Simple		HotBot Advanced	
Name	Total	SHC	Total	SHC	Total	SHC	Total	SHC
Ballard, Rice C.	167,880	N	1	N (2)	6943	N	12	N (1,2,3)
Blackwell, Margaret E.	183,460	N	1	N	9577	N	90	N(1,3)
Bones, James	481,510	N	45	N	52,294	N	97	Y
Burwell, Elizabeth Gayle	175,819	N	0	N	239	N (1)	0	N
Burwell, George W.	3,829,462	N	3	N (2)	2226	N (1)	34	Y (1,2)
Capehart, Susan Bryan Martin	1,569,967	N	1	N (2)	92	Y (2)	3	Y (2)
Green, William Mercer	450,200	N	8	Y (2,3)	12,573	N	4155	Y (2)
Hunt, Sophia Hughes	931,990	N	1	N (2)	1761	N	37	Y (2)
Ker, John	321,640	N	84	N (2,3,5)	6119	N (1)	108	Y (2)
Little, Benjamin Franklin	1,041,160	N	11	N (2,5)	27,253	N	281	N (1)
Little, *Rufus*	193,430	N	29	N (2)	12,281	N	26	N
Martin, Peter Boyd	525,400	N	1	N (2)	26,944	N	5168	Y
Massenburg, Lucy C.	216,200	N	0	N	51	N (2,3)	10	N (2)
Massenburg, Nicholas Bryar	271,024	N (4)	4	N (4)	1	N (4)	3	N (1,4)
McDowell, Thomas David Smith	4,550,080	N	7	N (2)	11,622	N	5	N (2)
Meade, Rebecca Beverley	315,780	N	1	N (2)	139	Y (2,3)	18	Y (2)
Platt, Eleanor Meade	353,850	N	1	N (2)	140	Y	12	Y (2)
Prudhomme, Phanor	5222	N (2,4,5)	4	N (2,4,5)	6	N (1,4)	5	N (1,4)
Robertson, Eliza Anne Marsh	1,040,450	N	1	N (2)	505	N	0	N
Swann, Ann Sophia Green	436,200	N	2	Y (2)	160	Y	3	Y (2)

Numbers (1)–(5) refer to additional SHC documents retrieved in personal names searches (see Table 5): (1) SHC finding aid index (2) *Records of Ante-Bellum Southern Plantations from the Revolution Through the Civil War* (3) Other SHC finding aid (4) *Selected Business History Resources in the Southern Historical Collection* (5) "Guide to African-American Doucmentary Resources in North Carolina"