Archival MARC Records and Finding Aids in the Context of End-User Subject Access to Archival Collections

Rita L. H. Czeck

Abstract

This article discusses the findings of a study to determine the extent to which archival MARC records represent chronological, geographical, personal, and corporate information contained in corresponding finding aids to archival collections. A content analysis of twenty finding aids to archival collections and their corresponding archival MARC records was conducted. The data suggest that the level of representation in archival MARC records varies depending on subject category. Geographical terms were the most likely to be represented, followed by personal names, chronological terms, and lastly corporate names. Allowing for the searching of full-text electronic finding aids would enable end users to benefit not only from the subject information present at the collection level and in the abstract, but also from the areas in finding aids that tend to get less MARC representation: scope/content notes, historical/biographical information, series summaries, and container information.

Introduction

Many archives and manuscript repositories have made finding aids available via the Internet. Websites with finding aids from hundreds of repositories nationwide may be a future alternative to searching bibliographic utilities such as the Online Computer Library Center (OCLC) for archival and manuscript collection information. Searchers may have the option to search either archival Machine Readable Cataloging (MARC) records or full-text finding aids in the same database. While the detailed information in finding aids may be useful for end users in determining relevance, it is unclear whether the finding aid format will be suitable as an initial locator

The author wishes to acknowledge her husband, David, for his encouragement and support throughout the research process.

of archival collections. Steven Hensen, one of the main developers of the MARC format for archival use and the author of *Archives, Personal Papers, and Manuscripts,* the defacto standard for archival cataloging, asserts that while finding aids may be available on-line, "it still seems likely that the pointers to such material will probably be structured catalog records."[1] The production of MARC records and their entry into bibliographic utilities, as well as the preparation of finding aids for on-line environments, represent a significant investment of time and money for archival repositories. Since MARC records contain a subset of the information provided in finding aids, what are the advantages of archival MARC records as compared to full-text finding aids for information retrieval? This article discusses a study conducted to analyze the subject information included in finding aids and the archival MARC records derived from them.

Literature Review

The main function of an archival MARC record is to abstract the most relevant information from the finding aid to provide a brief, accurate description of the collection. Among the portions of the MARC record that are typically available to end-user searching are title, author, a summary content note or abstract, and a brief historical or biographical note. Each MARC record also provides a list of index terms using a controlled vocabulary, such as Library of Congress subject headings. Index terms provide a succinct summary of the most important subject information in the finding aid. The purpose of a controlled vocabulary, which controls for synonyms and different forms of names, is to allow the end user to collocate records that are topically similar without developing an elaborate search strategy or conducting multiple searches for a given subject. The different types of subject terms listed in MARC records include geographical terms, personal names, corporate names, conferences, and occupations, as well as topical subject terms that do not refer to a specific person, place, or thing. Often, however, the list of index terms found in a MARC record is also in the corresponding finding aid, so index terms are not unique to the MARC format.

In order to evaluate what information needs MARC records are best suited to address, it is useful to examine what elements typically make up user queries. In his study of patrons at the National Archives, Paul Conway analyzed 212 initial user questions posed to front desk staff at the archives.[2] He reported the most frequent elements in these initial queries were me-

[1] Steven L. Hensen, "RAD, MAD, and APPM: The Search for Anglo-American Standards for Archival Description," *Archives and Museum Informatics* 5 (Summer 1991): 2–5.

[2] Paul Conway, *Partners in Research: Improving Access to the Nation's Archive: User Studies of the National Archives and Records Administration* (Pittsburgh: Archives & Museum Informatics, 1994).

dium, date, name, subject, place, and organization. Helen Tibbo in her study of providing access to historical literature asked historians to describe, in an open-ended format, what information would ideally be found in abstracts of historical writing.[3] The results indicate the types of information most important to history scholars are chronological, geographical, individual/group, and topical subject terms. The Getty Online Searching Project conducted by Marcia Bates and colleagues was an attempt to study how humanities scholars operate as end users of on-line databases.[4] The findings of the Getty study indicate that humanities scholars are most interested in personal names, geographical terms, chronological terms, discipline terms, and nonspecific topical subject headings when conducting on-line searches of document surrogates.

While both finding aids and MARC records incorporate personal, corporate, geographical, chronological, and nonspecific topical information, MARC records represent a subset of these data. To compare the advantages of the two formats for information retrieval, it is helpful to review studies that address differences between a full-text document and an abstract with a list of index terms using a controlled vocabulary. Studies analyzing the advantages and disadvantages of these formats in the context of on-line searching generally show that full-text searching provides a higher recall ratio, whereas abstract and index language surrogates provide a higher precision ratio (see Table 1). The recall ratio is the proportion of relevant items retrieved out of all relevant items in a database. If there are a total of fifty relevant MARC records in OCLC, and ten are retrieved, then the recall ratio would be 20 percent. The precision ratio is the proportion of relevant items retrieved out of all items retrieved. If twenty items are retrieved, for example, and ten of those items are relevant, then the precision ratio would be 50

Table 1. Retrieval Performance of Full Text, Abstracts, and Index Terms

	Full Text		Abstract and Index Terms		Abstracts		Index Terms	
	Prec*	Rec**	Prec	Rec	Prec	Rec	Prec	Rec
Tenopir	18%	74%	37%	19%	—	—	—	—
Ro	14%	84%	—	—	59%	18%	67%	21%
McKinin	37%	75%	62%	41%	—	—	—	—
Blair & Maron	79%	20%	—	—	—	—	—	—

* "Prec" = Precision
** "Rec" = Recall

[3] Helen R. Tibbo, *Abstracting, Information Retrieval and the Humanities: Providing Access to Historical Literature* (Chicago: American Library Association, 1993).

[4] Marcia J. Bates, Deborah N. Wilde, and Susan Siegfried, "An Analysis of Search Terminology Used by Humanities Scholars: The Getty Online Searching Project Report Number 1," *The Library Quarterly* 63 (January 1993): 1–39.

percent. Carol Tenopir conducted a study that compared the retrieval performance of searching full-text documents in the Harvard Business Review Online database versus searching a combination of abstracts and controlled vocabulary (or "bibliographic union").[5] She found that searching the full-text documents produced an average recall ratio of 74 percent, but only an 18 percent precision ratio. Conversely, the bibliographic union of abstracts and index terms produced a recall ratio of only 19 percent, but a precision ratio of 37 percent. Jung Soon Ro's study was a replication of the Tenopir study on a smaller scale, and the findings produced an even more dramatic difference between full text and abstract/index formats.[6] The recall ratio for full-text searching was 84 percent, while the precision ratio was only 14 percent. Searching only the abstracts produced a recall ratio of 18 percent, but a precision ratio of 59 percent. Finally, the controlled vocabulary terms produced a recall ratio of 21 percent, but a precision ratio of 67 percent. A more recent study conducted by Emma Jean McKinin and associates examined retrieval performance using the major medical databases: Medline, CCML, and MEDIS.[7] Retrieval on Medline, a database with bibliographic records that include abstracts and index terms, was compared to retrieval using the full-text databases CCML and MEDIS. Again, as in the Tenopir and Ro studies, full-text searching produced a relatively high average recall ratio (75 percent) and a relatively low average precision ratio (37 percent).[8] Searching the bibliographic records again produced a relatively low recall ratio (41 percent) and a relatively high precision ratio (62 percent).

Conversely, David C. Blair and M.E. Maron found evidence that full-text retrieval produced a high precision ratio (79 percent) and a low recall ratio (20 percent).[9] Sung Been Moon summarized the possible reasons why the Blair and Maron study produced different results from the other retrieval studies.[10] The differences may have been caused by different document types, different definitions of recall, or different methods of evaluating relevance. The most important factor is the different definition of recall used by Blair

[5] Carol Tenopir, "Full Text Database Retrieval Performance," *Online Review* 9 (April 1985): 149–64.

[6] Jung Soon Ro, "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. I. On the Effectiveness of Full-Text Retrieval," *Journal of the American Society for Information Science* 39 (March 1988): 73–78.

[7] Emma Jean McKinin, Mary Ellen Sievert, E. Diane Johnson, and Joyce A. Mitchell, "The Medline/Full-Text Research Project," *Journal of the American Society for Information Science* 42 (May 1991): 192–208.

[8] I have computed the values for full-text retrieval performance by averaging together the results of searching the CCML and MEDIS databases.

[9] David C. Blair and M.E. Maron, "An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System," *Communications of the ACM* 28 (1985): 289–99.

[10] Sung Been Moon, *Enhancing Performance of Full-Text Retrieval Systems Using Relevance Feedback* (Ph.D. diss., University of North Carolina at Chapel Hill, 1993).

and Maron. Ro and Tenopir defined the total number of relevant documents as the number of relevant documents in the union of sets retrieved by several searches on the same topic. McKinin used a similar method, but referred to it as comprehensiveness. Blair and Maron, however, sampled a subset of the document collection and examined it to assess the number of relevant documents, and used the sample to estimate the total number of relevant documents for a given query. For this reason, Blair and Maron's total number of relevant documents probably reflected a much higher percentage of the total number of documents in the database, thereby causing the recall ratio to be low. The sampling method used by Blair and Maron cannot be discounted, and in fact may be a better measure of total number of relevant documents than the method used by Tenopir, Ro, and McKinin. While the preponderance of evidence from the various studies shows that full-text retrieval will generally produce high recall and low precision ratios,[11] the findings of Blair and Maron suggest that the recall ratios found in the other studies are inflated. Blair and Maron's study did not, however, compare full-text retrieval to abstract/index term retrieval. It would have been interesting to see whether abstract/index term retrieval would have produced even smaller recall and greater precision ratios.

Given the general tendencies of full text and abstract/index term retrieval performance, there are implications for the effectiveness of archival MARC records and electronic full-text finding aids. Because precision levels tend to be low for full text, searching a database of full-text finding aids could present the user with the problem of "output overload," or the retrieval of an excessive number of irrelevant documents for a given search. Precision levels are usually higher when retrieving information from databases of abstracts and/or lists of index terms. A high precision level will result in a more manageable number of hits per search, and this is a strong argument for using the archival MARC record as an initial locator of a collection. On the other hand, if a user is concerned with finding the complete set of relevant collections, the potential for a higher recall level is an argument for searching full-text finding aids.

While the MARC format ideally represents the most relevant subject information in finding aids and provides the advantage of precision, the individual record is only as good as the quality of the cataloging. Although descriptive standards are supposed to provide consistency in descriptions from different repositories, archival cataloging is often inconsistent. Jackie Dooley notes the need for more consistent subject access to archival and manuscript collections cataloged in the MARC format.[12] She advises that

[11] The evidence presented only speaks to full-text retrieval performance in the absence of search engine techniques such as term-weighting and relevance feedback.

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

more attention should be paid to proper names, time periods, geographic places, and organizations, among other types of terms. Dooley maintains the MARC format is more than adequate to accommodate subject data, and archivists need to upgrade the provision of subject access to archival collections within the MARC structure.

Although full-text finding aids should offer greater levels of recall in information retrieval than MARC records, it is not clear to what extent finding aids represent potential subject terms that MARC records do not. Conversely, if the most important categories of subject terms, such as chronological, geographical, personal, and corporate, are often omitted or underrepresented in MARC records, the advantage of precision may not outweigh the disadvantage of low recall. The following is an analysis of the extent to which archival MARC records are likely to include or omit important subject term categories.

Methodology

In order to discover the extent to which archival MARC records are likely to represent the most important categories of subject terms, a content analysis comparing archival MARC records to their corresponding finding aids was conducted. The focus was specifically on four broad types of information: chronological, geographical, personal, and corporate. Twenty finding aids were chosen along with their corresponding MARC records in the OCLC database. All of the finding aids were chosen from the Berkeley Finding Aid Project website in order to provide electronic searching capabilities.[13] During the initial phase of content analysis, however, it became clear that searching the finding aids electronically would not provide an accurate count of subject terms. A manual count of subject terms proved to be more effective. While all of the finding aids selected for this study were chosen from the Berkeley Finding Aid Project website, they originated from three different repositories and ranged from three pages to twenty-six pages in length. Of these finding aids, two were for corporate papers, two were for family papers, and sixteen were for personal papers. Two study design factors prevented a more even mix of types of papers: all of the finding aids used for this study were chosen from the Berkeley Finding Aid website; and because the author did not have access to the Research Libraries Information Network (RLIN) database, only the finding aids at the Berkeley site that had corresponding MARC records available via OCLC were used. The OCLC limitations on MARC record length are fifty fields and 4,096 characters per record.[14] Since RLIN records do not have the same size restrictions as OCLC records, they can include more sub-

[13] <http://sunsite.berkeley.edu/FindingAids/> (accessed 7 November 1996).

[14] Electronic mail received from Tony Chirakos of the OCLC organization, March 1996.

ject information. If RLIN records had been used in this study, the results may have been very different. In addition, the finding aids did not follow a standard format, in that each repository has its own criteria for structure and inclusion of information. All of the finding aids, however, included typical finding aid elements, such as collection information, scope and content notes, historical or biographical information, and container information.

Once the finding aids were chosen, a print copy of each finding aid was visually scanned, and each instance of chronological, geographical, personal, and corporate terms was counted and recorded. Since counting subject terms is highly subjective and time-consuming, nonspecific topical subject terms, such as "trees" or "computers" were not included in the analysis. The following criteria specify what types of terms were counted in each category:

1. *Chronological terms*
   a. Individual dates, individual dates listed in a date range, or time-span indicators
      Examples: 1952, 1968–1972 (just 1968 and 1972, not the dates between 1968–1972), 1940s, Twentieth century, Middle Ages
2. *Geographical terms*
   a. Political: countries, states/provinces, counties, cities/towns
   b. Geological: deserts, rivers, mountains, oceans, seas, lakes, etc.
   c. Specific sites: buildings, dams, roads, etc.
   d. Adjectives
   e. Do not include common, unspecified terms (e.g., western states)
      Examples: Mexico, Colorado River, Hoover Dam, Mexican
3. *Personal names*
   a. A person's full name or last name
   b. Family names
      Examples: Harry Crump, Woodell, The Boyte Family
4. *Corporate names*
   a. Companies, associations, societies, institutions, foundations, etc.
   b. Subdivisions: bureaus, departments, etc.
   c. Newspapers and magazines, if a place of employment for or founded by someone listed in the finding aid
   d. Do not include common, unspecified terms (e.g., administrative committee)
   e. Do not include if in the title of a conference, meeting, forum, workshop, symposium, etc. (e.g., The Third Annual Conference of the Usenix Association)[15]

[15] Conference names have their own category, distinct from organizations. I did not investigate the representation of conference names for this study.

Examples: General Electric, Library Association, Minnesota Historical Society, U.S. Department of Labor

Some terms in the finding aids were not included in the content analysis because they were not in a context deemed useful for subject retrieval. If any of the four types of subject terms were found in the following contexts in the finding aids, they were not included in the content analysis:

1. Location of collection,
2. Encoder and encoding dates of finding aid,
3. Processing of collection,
4. Publishers, dates, and titles in bibliographical references,
5. Folder dates and chronological container information, or
6. Birth and death dates in Library of Congress authorized form of name.

Once all of the terms that fell into the four subject categories were recorded, each of these terms was compared with a print copy of the corresponding MARC record to see whether the term was represented anywhere in that format. The location of the term in the finding aid was also recorded. Each finding aid was broken into sections according to the following definitions:

1. *Collection information*
   Includes both the title of the collection and the overall date span of the collection.
2. *Abstract*
   A brief summary, usually no longer than a paragraph, recording the most important features of a collection.
3. *Scope/Content notes*
   A short section, generally one to two pages in length, describing the scope and the series and subseries of the collection, and types of materials present.
4. *Historical/Biographical notes*
   A short section usually ranging from one to two pages that provides a background of the primary person or institution related to the collection.
5. *Series information*
   Includes the series title, date span of the series, and series summaries that are normally no longer than a paragraph in length.
6. *Container information*
   A detailed listing that describes the contents of containers, typically down to the folder level. Container information can range from a few pages to hundreds.
7. *Other*
   Information that does not fall into the first six categories, such as related collections and donor information.

When considering whether a term from a finding aid was represented in its corresponding MARC record, the term had to be exactly the same in both the finding aid and the MARC record to be considered a match, except for the following cases:[16]

1. The term was obviously the same one but misspelled.
2. First and last name of a person was inverted. (For example, Carl Hammons and Hammons, Carl would constitute a match.)
3. A person's last name only, if identified by context, matched first and last name.
4. A shortened version of a corporate name (except for acronyms), if identified by context, matched the full name.
5. Dates: 1970–1980 matched 1970 to 1980; 1970's matched 1970s.

Findings

This section provides a detailed analysis of the extent to which the four types of subject terms were represented in the MARC records. The analysis presents the average percentage of representation of each type of subject category in the MARC records. This section also provides an analysis of whether the MARC records represented terms that were located in different areas of the finding aids: collection information, abstract, scope/content notes, historical/biographical notes, series information, and container information.[17]

Only one subject category, geographical, was omitted from a MARC record when there were terms from that category present in the finding aid; this occurred in only one collection out of the twenty analyzed. Aside from this occurrence, if there were terms from a given subject category in a finding aid, that category had at least some representation in the corresponding MARC record. The extent to which the subject categories were represented in the MARC records is given in Tables 2, 3a, 3b, 3c, and 3d. Table 2 shows that all of the types of subject terms—chronological, geographical, personal, and corporate—were represented on average less than 50 percent but more than 20 percent of the time. The most represented type of subject term in the MARC records was the geographical category at 41 percent. Personal names from the finding aids were represented on average 37 percent of the time. Chronological terms were represented on average 27 percent of the time, and lastly an average of 23 percent of corporate names from the finding aids were represented in the MARC records.

[16] The focus of this study was to discover whether a given concept from the finding aid was represented in the MARC record, not whether an individual searcher would be able to retrieve the term in precisely the same way from finding aid to MARC record.

[17] While only the average percentage of terms from the finding aids represented in the MARC records are provided in this article, the author can supply the raw data from which the averages were computed upon request.

Table 2. Average Percentage of Terms from Finding Aids Present in MARC Records, by Subject Category

Subject Category	Average percentage of terms present in MARC records
Chronological Terms	27
Geographical Terms	41
Personal Names	37
Corporate Names	23

Table 3a through 3d show the average percentage of terms from the finding aids present in the MARC records, but also break down the finding aids into their component parts so that further analysis is possible.

Chronological Terms

While only an average of 27 percent of chronological terms from the finding aids were represented on the whole, Table 3a demonstrates that chronological terms derived from the collection information and abstract were represented at a relatively frequent 89 percent and 75 percent, respectively. Chronological terms in collection information almost exclusively delineate the date range for the whole collection and the date range that includes the bulk of the collection, and these dates tended to have a high representation rate. The abstract of a finding aid includes the most important dates regarding the collection, and tended to have a high representation rate not only for chronological terms, but for all of the subject categories. Chronological terms from the scope/content area were represented at 46 percent, and from the series level 41 percent were represented. The scope/content area contains chronological terms in a narrative fashion similar to the abstract, but is typically much more detailed and lengthy than the abstract. Series level chronological terms are sometimes an indication of the range of the entire series. Series summaries, however, contain chronological information that indicate specific dates of events that relate to particular contents within the series. The historical/biographical section chronological terms were least likely to be represented in the MARC records at 17 percent. Often the historical/biographical section of a finding aid is simply a chronology, listing one date or date range after another in a list, with a short explanation after it.

Geographical Terms

Being the most represented of all subject categories overall in the MARC records at 41 percent, geographical terms had a higher level of representation in the abstract and scope/content sections than any other subject cate-

Average Percentage of Terms from Finding Aids Present in MARC Records by Subject Categories and Finding Aid Sections

Table 3a. Chronological Terms

Finding aid section	Average percentage of terms present in MARC records
Collection Information	89
Abstract	75
Scope/Content	46
Historical/Biographical	17
Series	41
Container Information	n/a
Other	50

Table 3b. Geographical Terms

Finding aid section	Average percentage of terms present in MARC records
Collection Information	n/a
Abstract	100
Scope/Content	67
Historical/Biographical	36
Series	43
Container Information	17
Other	n/a

Table 3c. Personal Names

Finding aid section	Average percentage of terms present in MARC records
Collection Information	100
Abstract	78
Scope/Content	61
Historical/Biographical	50
Series	91
Container Information	38
Other	56

Table 3d. Corporate Names

Finding aid section	Average percentage of terms present in MARC records
Collection Information	100
Abstract	93
Scope/Content	59
Historical/Biographical	21
Series	64
Container Information	12
Other	88

gory. All of the geographical terms (100 percent) from the finding aid abstracts were present in the MARC records. Scope/content geographical terms were represented at 67 percent in the MARC records. At the series level, 43 percent of the geographical terms were represented, and 36 percent of the historical/biographical section geographical terms were represented. Only 17 percent of the geographical terms from the container information of the finding aids were present in the MARC records. In the collection information, no geographical terms were noted because none of the collections had a title that was coded as a geographical term.

Personal Names

Personal names were second only to the geographical category in terms of representation without regard to finding aid section, at 37 percent. More specifically, though, personal names along with corporate names in the collection information had the highest representation. For personal and family papers, the title of the collection always includes some form of the personal name, and collection information personal names were represented 100 percent of the time in the MARC records. Personal names mentioned in the finding aids' series level information were represented 91 percent of the time, a far greater number than the next highest percentage of representation at the series level, being corporate names at 64 percent.

Personal names in the abstracts were represented 78 percent of the time, lower than both geographical terms (100 percent) and corporate names (93 percent). For personal names listed in the scope/content section, the representation in the MARC records was 61 percent, second only to geographical terms at 67 percent. Half of all personal names in the historical/biographical section on average were represented, a much higher level than any of the other subject categories for this section of the finding aids. Similarly, representation of personal names from the container information was significantly higher at 38 percent than any other subject category for container information.

Corporate Names

As mentioned above, all corporate names from the finding aids' collection information were represented in the MARC records. The level of representation of corporate names from the abstracts was also relatively high, 93 percent, second only to geographical terms at 100 percent. Representation of corporate names dropped off to an average of 64 percent at the series level and 59 percent from the scope/content section of the finding aids. The only subject category with less representation in both of these finding aid

areas was chronological terms with 41 percent of series level information and 46 percent of scope/content information being represented. Corporate names from the historical/biographical section of the finding aids were represented at 21 percent in the MARC records, and corporate names from the container information were represented only 12 percent of the time on average, the lowest representational level out of all the subject categories for this section of the finding aids.

Conclusion

Because of the increased accessibility of the Internet, archivists are presented with an opportunity to make in-house finding aids accessible to a wide community of searchers. Clearly, searchable and downloadable finding aids are wonderful research tools once a user has connected to a repository's website. The question remains, however, whether finding aids alone are sufficient as an initial locator of a collection, especially when searching across collections. The production of MARC records and their entry into bibliographic utilities, as well as the preparation of finding aids for on-line environments, represent a significant investment of time and money for archival repositories. Although full-text finding aids should offer greater levels of recall in information retrieval than MARC records, it is not clear to what extent finding aids represent potential subject terms that the MARC records do not. Conversely, if the most important categories of subject terms, such as chronological, geographical, personal, and corporate, are often omitted or underrepresented in MARC records, the advantage of precision may not outweigh the disadvantage of low recall.

The findings of this paper suggest that each of the subject types, chronological, geographical, personal, and corporate, are likely to be represented, at least at a minimal level, in MARC records. The level of representation varies, however, depending on subject category and section of the finding aids. Geographical terms were the most represented, followed by personal names, chronological terms, and lastly corporate names. The level of overall representation varied from 41 percent down to 23 percent. Since the purpose of a MARC record is to represent the most important information from a finding aid, it is expected that not all of the terms would be represented. The average number of terms from these important subject categories that were only present in the finding aids, however, was great. In addition, when looking at the different portions of the finding aids, the representation of terms varied considerably. Collection information should almost always be incorporated into a MARC record, since it is essentially the name and dates of a collection. This is reflected in the findings, in that personal and corporate terms from the collection information were represented at 100 percent,

and chronological terms at 89 percent. Since the abstract is intended to summarize the most important features of a collection, it would seem that most of the subject information from this section should be recorded. This, too, is borne out by the findings: the subject terms from the abstracts were represented at least 75 percent of the time, up to 100 percent for geographical terms. The rest of the sections of the finding aids were not so consistently represented. The scope/content section chronological terms were represented only 46 percent of the time, and the series chronological terms were present only 41 percent of the time. The historical/biographical section provides a background for the collection, and perhaps is not as critical for subject access, but the level of representation from this section was quite low. The container information was the least represented area, although this is not surprising since the information in this area is relatively specific and more comprehensive than the other areas of finding aids. These findings must be viewed, however, with the understanding that RLIN records may provide an even greater average percentage of relevant subject information from finding aids than do OCLC archival MARC records.

MARC records seem best suited to address searches for personal and corporate names that are central to the collection, such as a search for a person for whom the collection is named. Searching finding aids for a specific person, however, may retrieve a collection in which the person was only a minor correspondent. The person may have been considered too peripheral to be included in a MARC record, but the collection could still be retrieved by searching the full-text finding aid. A search for chronological terms in a database of MARC records may not be fruitful unless it is for the date range of the entire collection. Finding aids tend to provide a much greater number of chronological terms than MARC records, and the majority of these terms are single dates or date ranges having to do with the historical background of the subject of the collection. Searchers who have a specific date or a specific date range other than the range of the collection in mind, such as a series date range, would benefit from being able to search the full-text finding aid. Geographical terms that are prominent in the background of a person or corporation, such as where a person resided when they produced the materials in the collection, are likely to be found by searching MARC records. Searching MARC records when the collection itself is closely related to a geographical subject, such as the Central Arizona Project Association, may be useful if searching on frequently mentioned geographical features in the finding aid. Many geographical terms that specify folder contents, however, tend not to be represented in MARC records.

It is clear from these findings that a significant amount of subject information tends to be present in finding aids, but not in their corresponding MARC records. Making the full text of finding aids available through an on-

line database for subject searching would provide end users an alternative to searching MARC records in a bibliographic database. Allowing for the searching of full-text finding aids would enable end users to benefit from the subject information present not only in the collection information and in the abstract, but also from the areas in finding aids that tend to get less MARC representation: scope/content notes, historical/biographical information, series summaries, and container information. A useful alternative to searching MARC records or the entire full text of finding aids may be targeted field searching of certain sections of finding aids, e.g., collection information, abstract, and scope/contents notes.

As with MARC records, however, the database into which the full-text documents are loaded can have an impact on retrieval effectiveness. The full-text format has the potential to burden the user with excessively large retrieval sets with many nonrelevant hits, depending on the size of the database. A database that realistically reflects the hundreds of thousands of finding aids available nationwide may amount to nearly nineteen million pages of text.[18] With the increasing reliance on retrieving information from large databases, there is a need for archivists to become expert searchers so they can both act as intermediaries for their patrons and educate them to conduct searches for themselves outside of repositories. In addition, research is needed to compare the retrieval performance of full-text finding aids versus their MARC surrogates in terms of recall and precision.

[18] American Heritage Virtual Archive Project: A Proposal to the National Endowment for the Humanities (The Library, University of California, Berkeley) available at <ftp://library.berkeley.edu/pub/sgml/ead/beta/ameriher.txt> (accessed 7 November 1996).