Noah G. Huffman. Search Features and Other Characteristics of XML Retrieval Systems for EAD Finding Aids: A Content Analysis. A Master's Paper for the M.S. in L.S. degree. April, 2008. 56 pages. Advisor: Katherine M. Wisser

To date, most studies of Encoded Archival Description (EAD) have considered issues surrounding its development as a standard or its implementation at specific repositories. Few studies have examined the features and functionality of EAD retrieval systems. This study provides a content analysis of 33 XML-based EAD retrieval interfaces, with particular emphasis on search features offered, how search results are configured, how finding aids are displayed, and the types browse options offered. Findings of this research indicate that few EAD retrieval systems provide the kind of sophisticated searching and browsing features that early proponents of EAD promised. As an analysis of the current state of access to EAD finding aids, this research provides a foundation for much-needed research on the usability of EAD retrieval systems.

Headings:

Archives/Cataloging

Encoded Archival Description

Information systems

XML (Computer language)

SEARCH FEATURES AND OTHER CHARACTERISTICS OF XML RETRIEVAL SYSTEMS FOR EAD FINDING AIDS: A CONTENT ANALYSIS

by
Noah G. Huffman

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

April 2008

Approved by

_____

Katherine M. Wisser

TABLE OF CONTENTS

INTRODUCTION

For as long as repositories have collected and preserved manuscripts and other records of enduring value, the finding aid has served as the primary access tool to archival collections. Alternatively known as manuscript registers, inventories, or collection guides, archival finding aids, at the most basic level, provide users with an overview of the contents and organization of an archival collection. While finding aids vary greatly in length and level of detail depending on the collections they describe or the repositories that create them, they remain an important tool for providing access to archival materials.

Inspired by the growth of networked communication in the 1990s and its potential for reaching remote users, archivists collaborated to develop a standard markup language for finding aids known as Encoded Archival Description (EAD). Released in 1998 as an SGML standard, EAD version 1.0 and its successor, EAD 2002--an XML Document Type Definition (DTD)--have shaped the conversation of descriptive practices in the archival community perhaps more than any other single topic. Early proponents of EAD saw it as a transformative technology that could preserve the traditional structure of the archival finding aid, but free it from the constraints that traditional media had imposed. Other archivists were more guarded in their initial appraisal of EAD. Wary of the time commitment, expense, and technological expertise that implementing EAD might require, they took a more cautious "wait and see" position (Pitti and Duff, 2001). At present,

most of the archival community has embraced EAD, and even some of the smallest repositories have begun to implement the standard.

In the ten years since the release of EAD 1.0, the bulk of the professional literature has dealt with its development as an encoding standard and issues surrounding its implementation at various repositories. However, few studies have carefully examined the design and functionality of search and retrieval systems for EAD finding aids.[1] Until recently, only a few repositories had implemented any type of search and retrieval system that leveraged the underlying XML structure of EAD. Instead, most repositories provided access to EAD finding aids either through simple browsable lists on their web site, through basic full-text search engines like Google, or indirectly via links in MARC records for archival collections in traditional bibliographic databases like WorldCat. With the emergence and growing popularity of XML database technologies in the last few years, some sophisticated EAD search systems have appeared on the scene. As Anne Gilliland-Swetland envisioned in a 2001 article on the promises of EAD, these new XML technologies have the potential to create a truly integrated "archival information system" capable of exploiting the semantic richness of EAD and facilitating retrieval by diverse user groups (Gilliland-Swetland, 2001).

Although several large consortium projects and research libraries have implemented XML-based retrieval systems for EAD finding aids, there has been no empirical examination of the search features and functionality offered by these systems. In response, this study provides a content analysis of XML-based EAD retrieval systems that seeks to answer two main questions: (1) How many repositories have implemented XML-based EAD search systems, and more importantly (2) What kinds of search

functions and other unique features do these interfaces currently offer. By extension, this study also seeks to provide a foundation for future research to evaluate the effectiveness of various features common among XML-based retrieval systems for EAD finding aids. At present, it is unclear whether these more sophisticated EAD interfaces offer any improvement in the discovery and retrieval of archival information for the majority of users. By describing the current state of EAD retrieval interfaces, this study will provide data that will allow future researchers to begin answering such questions.

LITERATURE REVIEW

Since the release of EAD in 1998, most studies of the encoding standard have addressed issues such as its creation, implementation, and adoption by the archival community. For example, Daniel Pitti (1997), one of the developers of EAD, described the atmosphere surrounding the creation of the standard in the mid-1990s and hinted at the potential for EAD to transform archival description. More recently, Elizabeth Yakel and Jihyun Kim (2005) examined the adoption of EAD by the archival community. Yakel and Kim surveyed 399 repositories that sent staff to EAD workshops and discovered that only 42% of these repositories had fully adopted EAD. From their survey and from supporting research such as Roger's (1962) theory of diffusion of innovations, Yakel and Kim concluded that a number of barriers have curtailed the widespread adoption of EAD such as lack of staff expertise, insufficient funding, and a multiplicity of existing access tools already in place for archival materials.

While similar studies mirror the research by Pitti, Yakel, and Kim on the creation and adoption of EAD as a standard, few studies have carefully examined the development

of search and retrieval systems for EAD finding aids or the various features unique to these systems. In 2001, James Roth surveyed archivists at 47 institutions that implemented EAD to elicit the most popular practices for delivering EAD finding aids via the web. While Roth discovered a number of deployment methods in place, he only identified two XML-based search and delivery systems—DynaText and Panorama—that had been implemented with any success (Roth, 2001). As of 2008, both of these systems are obsolete and no longer supported.

Surveying individual archivists, Roth also found that many viewed the cost and technological expertise required to develop and maintain XML search systems as a prohibitive barrier to implementation. Lacking a "cost-effective server-side XML delivery system," Roth noted that many institutions chose instead to mount their finding aids in HTML rather than EAD (p. 226). For reasons likely similar to those that slowed the adoption of EAD as described by Yakel and Kim (2005), Roth also found that few repositories had any type of formal evaluation measures in place to monitor the effectiveness of their EAD program or to gauge end-user satisfaction with EAD finding aids or search system interfaces. The information Roth obtained from the few repositories that did evaluate their EAD programs, suggested that end-users "don't seem to care about the structure or format of their finding aids, just the content" (p. 34).

Since Roth's study, EAD has migrated from a subset of SGML to a Document Type Definition (DTD) of XML. In early 2007, EAD was upgraded to an XML schema. The emergence of several more advanced and open-sourced XML database tools and applications have also significantly changed the technological landscape since Roth's article. Nevertheless, Roth identified some of the persistent issues related to

implementation of EAD search systems and accurately described many of the attitudes archivists still hold about how user needs might or might not be addressed by certain features of these systems.

The same year as Roth's study, archivist Anne Gilliland-Swetland (2001) issued a call in the archival community for the design of a comprehensive "archival information system" that could effectively exploit the semantic richness of EAD and facilitate retrieval by diverse user groups. According to Gilliland-Swetland, "nowhere is the potential of EAD more apparent than in addressing the role of the finding aid as an information discovery and retrieval tool" (p. 210). In elaborating on the potential of EAD to improve discovery of archival materials, Gilliland-Swetland suggested ten strategies for enhancing browsing and retrieval in the ideal "archival information system" (p. 215-221). Many of Gilliland-Swetland's strategies focus on a search interface that allows fielded searching by subject, name, dates, geographic locations, and genres. This type of fielded searching is made possible by the XML structure of EAD, which requires archivists to encode or "tag" certain text within a finding aid, and in the process, give that text meaning. For example, by using EAD tags—also known as elements--such as <persname>, <geogname>, and <subject>, archivists can declare that the information contained within these tags refers to a person, a geographic location, or a subject, respectively. Such tagging practices bring with them the implicit expectation that information contained within these semantic tags will be indexed, searchable, or browsable in some sort of search system. While many archivists meticulously tag the text of finding aids in this way according to EAD guidelines, very few repositories have

implemented an EAD search interface that can effectively leverage this value-added encoding.

Some six years after suggesting her ten strategies for exploiting EAD in a search interface, most of Gilliland-Swetland's recommendations have not been realized. Although a few XML-based search systems--as discussed below--do provide fielded search capabilities, these systems are exceptions and not the norm. Considering that much of the early excitement over EAD centered around its presumed ability to support sophisticated indexing, navigation, and display (Pitti, 1999), the rarity of fully developed search systems that are truly capable of leveraging the value-added encoding provided by EAD in a consistently meaningful way is somewhat surprising. Nevertheless, Gilliland-Swetland's 2001 description of the ideal "archival information system" remains an important, although somewhat quixotic, statement about the potential of EAD search systems to improve the discovery and retrieval of archival information.

Perhaps one possible explanation for the slow development of XML-based search systems for EAD finding aids can be attributed to the current widespread practice of providing links to EAD findings aids from MARC records in traditional online public access catalogs (OPACs). Before the development of EAD in the mid-1990s, efforts by Steve Hensen (1986) and others helped establish the MARC-AMC (MAchine Readable Catalog Archival and Manuscript Control) standard for generating brief catalog records for archival collections that could co-exist with bibliographic records in library OPACs. Initially, the inclusion of archival catalog records in library databases was a huge leap forward for promoting discovery and access to archival materials, especially considering the growth of OCLC's WorldCat in the 1990s.

In the long-term, however, the widespread use of MARC-AMC records may have slowed the development of XML-based EAD retrieval systems. Because many early EAD adopters were already familiar with the MARC format, they began providing links to EAD finding aids in the MARC 856 field instead of putting their efforts into developing more sophisticated systems for searching and browsing EAD finding aids. For awhile, this strategy had some obvious benefits. As Hensen (2001) suggested, many users rely heavily on bibliographic databases like WorldCat and library OPACs, and so linking to EAD finding aids from MARC records allows users to exploit the familiar subject access and navigational metadata provided by the MARC structure. In this regard, Hensen argues, MARC and EAD should "co-exist as parts of an essential metadata structure for management and discovery of manuscript and archival materials" (p. 92).

While linking to EAD finding aids in MARC records is certainly a good idea, relying on MARC as the *only* discovery tool for finding aids is problematic. For the most part, a MARC record for an archival collection provides only a brief summary of the contents of a collection and a handful of access points. In contrast, a full EAD finding aid can provide detailed descriptions of each component of a collection and can sometimes reach a hundred pages or more. By ignoring the valuable container level information of an EAD finding aid and by restricting the total size of records, MARC is an insufficient discovery tool for most archival collections because it significantly reduces the amount of content that can be searched. In a web environment that increasingly supports full-text searching, an EAD finding aid provides much more searchable content than a MARC record. While MARC may offer the benefit of

authority control, relying solely on MARC for retrieval of EAD finding aids severely handcuffs the power inherent in the underlying XML structure of EAD. Moreover, if MARC records were the only discovery mechanism for finding aids, then finding aids could be delivered in HTML or PDF format, saving the added time and expense of encoding them in EAD.

At present, Hensen's recommendation to continue generating MARC records for archival collections is a convincing one that is likely to hold sway as long as bibliographic databases like WorldCat continue to exist. Linking to EAD finding aids from MARC records, however, is only a temporary, albeit entirely appropriate solution. If archivists continue to invest time and money into encoding finding aids in EAD, then they should expect to see the fruits of their encoding labors in the form of a more mature XML retrieval system.

During the initial excitement that surrounded the early years of EAD in the late 1990s, a few archivists at the University of Michigan and Harvard University proposed a project to develop an XML-based archival search system and union database for EAD finding aids for several participating institutions. As described by Mackenzie Smith (2000), one of the activities of the project, known as the Distributed Finding Aid Search System (DFAS), was to develop a list of Common Access Points (CAPs) that could be mapped to indexes of particular EAD elements. These indexes could then be used to support fielded searching of EAD finding aids as recommended by Gilliland-Swetland (2001). Building on existing access points in bibliographic database indexes, the investigators in the DFAS project identified nine such CAPs: Names, Dates, Titles, Places, Subjects, Repository, Contents, Summary, and Anywhere (full-text). In building

the system, designers focused on providing access to finding aids via these CAP indexes. While indexes were easily created for these categories, the DFAS investigators quickly discovered that the "lack of standardization in the application of EAD to finding aids" among repositories created several obstacles to consistent and meaningful search and retrieval. In concluding their report on the DFAS project, investigators pointed to inconsistency in encoding practices as the most significant barrier to the development of the kind of mature union catalog for archival materials that early EAD proponents had envisioned (Smith, 2000).

In response to the barriers to search and retrieval discussed by the final report on the DFAS project in 2000, archivists have more recently adopted strategies to enforce consistency of encoding more vigorously, but also to scale back the scope and expectations of EAD search systems. Successful examples of this approach can be seen in consortium projects like the Online Archive of California (OAC) and the Northwest Digital Archive (NWDA).[2] Founded in the mid-1990s, the OAC, as described by Brown and Schottlaender (2001), centralizes functions such as "monitoring to ensure consistency of EAD encoding" across finding aids in an effort to provide meaningful search and retrieval of finding aids from several participating institutions (p. 99). Rather than offer fielded searching of individual EAD elements such as <persname> and <geogname> as recommended by the DFAS report (Smith, 2000), the OAC takes a simpler approach. The default search in the OAC interface is a simple full-text search of the entire text of a finding aid. Although search functions are simplified, the OAC interface still relies on the XML structure of EAD finding aids to provide highly configured search results sets

and a number of other innovative features that are discussed in more detail in the findings below.

If the search functionality and display of retrieval sets in the OAC seems simple, the underlying technology is anything but. The OAC is powered by a suite of open-source native XML database applications and tools created by the California Digital Library and broadly known as eXtensible Text Framework (XTF). Instead of storing EAD components and structure in a relational database like most bibliographic retrieval systems, XTF allows documents to be stored intact in their native XML form. Through the use of carefully configured Java servlets, an Apache Lucene indexer, and an XML query language known as crossQuery, XTF "supports searching across collections of heterogeneous data and presents results in a highly configurable manner."[3] Because of its open source code, innovative system architecture, and advanced user-interface, the OAC is one of the more robust and promising systems currently in place for discovery and retrieval of archival information.

Despite its promises, however, XTF and other open-source XML database solutions have their disadvantages. According to Chris Prom (2007), implementing systems like the OAC that rely on XTF or a similar suite of tools is typically "well beyond the skills of even the most computer literate archivist, much less the monetary resource base to which he or she might have access" (p. 159). Additionally, native XML databases like OAC's implementation of XTF are also predicated on the consistent encoding of finding aids by all participating institutions who contribute EAD documents to the consortium. Prom argues that these two factors point to the need for a simpler,

more integrated system that can perform all of the functions needed to create, publish, and search EAD finding aids on the web.

To address the need in the archival community for a single tool to streamline the encoding, publishing, indexing, and searching of EAD finding aids, Prom and other archivists at the University of Illinois at Urbana-Champaign developed the ARCHON platform. According to Prom (2007), the ARCHON system, which was released in August 2006, "is tailored specifically to archival needs and can be implemented with little technical knowledge" (p. 157). Unlike the XTF platform, which stores EAD data in its native XML format, ARCHON utilizes a relational database model. According to Prom, the designers of ARCHON decided that, for a number of reasons, "XML made much more sense as a data interchange format rather than as a data storage format" (p. 160). Prom also suggested that the bulk of programmers have more experience working with web scripting languages that interact with relational databases than they do with XML query languages and XSLT, which are required to drive native XML databases. By allowing data to be entered into a relational database model, Prom argues that implementation of ARCHON or a similar system can streamline workflows at individual repositories by generating EAD finding aids, MARC21 records, and a searchable web interface from the same data instance. In addition, by strictly enforcing the EAD output, Prom suggests that adoption of a full-service EAD publication tool such as ARCHON will help curtail the proliferation of inconsistent local practices in encoding finding aids. In Prom's view, streamlining workflow and simplifying data entry in an integrated system like ARCHON not only lowers the barrier to entry for repositories desirous of implementing search system for their EAD finding aids, but it also ensures consistency in

encoding practices and allows archivists to concentrate on processing the enormous backlog of collections instead of constantly troubleshooting technical problems.

In light of the gradual adoption of EAD by more repositories in recent years and the development of new XML technologies like XTF, Xiaomu Zhou (2006) conducted a study of 58 repository web sites hosting EAD finding aids to determine the types of search features available for EAD. Of the 58 sites Zhou surveyed, she found that thirteen (22%) did not provide any sort of search capability (p. 105). The remainder of the sites exhibited a wide range of search interfaces and technologies. Although Zhou's study did not discriminate between sites using XML-based technologies and those that did not, she discussed several specific characteristics of search systems including the types of search fields provided, the way retrieval sets were organized and displayed, and the types of feedback mechanisms provided to users. From her observations, Zhou classified EAD search sites into some useful categories for the purposes of this study. For example, in her analysis of the display of retrieval sets on each site, Zhou identified three main presentation styles for search results: Google-style, archival professional style, and catalog-style (p. 111-112). In concluding her study, Zhou argued that "the advantages of EAD finding aids have not been fully realized" in a search and retrieval system (p. 117).

As a relatively recent analysis of some of the general search features of EAD websites, Zhou's study provides a useful framework and methodology for examining the functionality of XML-based retrieval interfaces. Moreover, Zhou suggested that future research should investigate how larger consortium project sites are implementing XML retrieval systems. Because well-funded and expertly staffed consortium projects constitute a large portion of XML-based search systems currently in place for EAD

finding aids, the present study considers the features and functions of these systems in particular.

Although this study offers a content analysis of the features of XML-based EAD interfaces, it is informed by usability studies of EAD interfaces specifically and XML-retrieval systems more broadly. In 2004, Elizabeth Yakel conducted a usability study of one EAD interface—the University of Pittsburgh's Historic Pittsburg Project. Yakel asked study participants to perform several basic search tasks and found that users rarely utilized any of the sophisticated fielded search options available on the site. In addition, Yakel noted that the use of archival jargon to describe fielded search options and other features of the system often confused users and kept them from taking full advantage of some the system's advanced functionality. To correct this problem, Yakel recommended that EAD sites "use search parameters that have more general intelligibility and meaning" (p. 74). To date, Yakel's study is one of the few to consider EAD search interfaces from a user perspective and it serves as a useful reference for evaluating features of current EAD interfaces. Yakel's study is also particularly valuable, because it examines the DLXS platform, currently one of the most widely used XML-based EAD retrieval solutions.

While Yakel evaluates EAD retrieval interfaces specifically from an archival perspective, other disciplines have generated important research on the evaluation of XML-retrieval systems more broadly. Launched in 2002, the Initiative for the Evaluation of XML Retrieval (INEX) is a large scale project to evaluate the effectiveness of XML retrieval systems (Fuhr et al., 2003). By developing a large test collection of XML encoded documents and establishing a common set of retrieval metrics, INEX has

encouraged a community of mostly Information Retrieval scholars to test all aspects of XML retrieval systems, including user expectations and satisfaction with these systems. Currently, INEX has not produced any research on EAD specifically, but research emanating from INEX helps inform the present study.

For example, one of the fundamental assumptions about EAD, and XML more broadly, is that tagging certain information in a document will enhance precision and relevance of retrieval, and allow users to identify and retrieve relevant sections of documents rather than entire documents. Usability studies of XML retrieval systems suggest that users may not prefer this added functionality. Betsi et al. (2006) found that users "expect to interact with documents" and "would feel rather uncertain if elements with no context information were retrieved" (p. 611). Similar research by Larsen et al. (2006) confirmed that, when given the choice, users clicked on links for full documents in a retrieval set 71% of the time as opposed to only 29% for links to portions of documents that were relevant to their queries (p. 664). These findings have important ramifications for XML-based EAD retrieval systems that often allow users to retrieve portions of finding aids such as series level descriptions or even container level descriptions rather than entire finding aids. Moreover, usability studies of XML-retrieval systems seem to confirm what many archivists have posited all along—that providing users with contextual information in a finding aid is just as valuable as directing them to a series or folder that might interest them.

Like retrieval of portions of XML documents, another common feature of XML-based EAD retrieval systems is the ability to conduct fielded searches, or searches restricted to the content within certain EAD elements such as <persname>, <geogname>,

<date>, or others. Potential improvements to the precision and relevance of retrieval offered by fielded searching has traditionally justified the time and effort required to encode or "tag" the content of EAD finding aids. Research conducted by Catherine Arnott Smith (2003) on retrieval of encoded clinical documents at the University of Pittsburgh Medical Center challenges these traditional assumptions. In a between-subjects study that required some physicians to retrieve XML-encoded clinical documents via a full-text search interface and required others to use an interface that permitted fielded searching, Smith found that tasks performed on the fielded search interface "required a mean number of *more* steps in the search sequence to a degree that was statistically significant" (p. 614). In addition, Smith found that tasks performed via fielded searches "had a statistically significant lower rate of precision" (p. 614).

In light of Smith's study of XML retrieval of clinical documents, and other research generated by INEX, it is still unclear if sophisticated indexing and fielded searching of EAD finding aids is really worthwhile, at least for the majority of users. Furthermore, it also begs the question, is all that effort spent "tagging" information in EAD finding aids really worth it?

Regardless of the searching options provided by an interface, a number of usability studies of EAD interfaces and of more general E-commerce websites suggest that users show a strong preference for browsing rather than searching. A usability study of EAD interfaces conducted by Chris Prom (2004) found that 74% of participants preferred browsing to searching. In addition, Prom noted that many participants expressed that "advanced search options will likely be appreciated only by a very small number of experienced users" (p. 250).

Similarly, studies of user-interactions with E-commerce sites revealed that less than 10% of users preferred searching to browsing when browsing categories were presented clearly. Moreover, even in "sub-optimal menu conditions" users still chose to browse 60% of the time (Straub, 2005). With user-preference for browsing clearly demonstrated in usability studies conducted across a number of different types of interfaces, it is clear that EAD retrieval interfaces should incorporate browse features into their overall design. As the findings of the present study suggest, browsing features must be integrated more fully into EAD retrieval interfaces.

METHODOLOGY

While the present study does not seek to evaluate the effectiveness of XML retrieval of EAD finding aids, it does seek to provide a content analysis of some search functions and other features currently offered by XML-based EAD search interfaces in order to provide a foundation for future evaluative research like the work done by Prom and Yakel. By examining EAD search interfaces across several archival repository and consortium websites, this research employed open-coding to iteratively develop categories of search functions and other features provided on each site (Babbie 2007, 384-385). Similar to Zhou's content analysis of general EAD interfaces (XML and non-XML) conducted in 2003 and published in 2006, this research considers features such as: the default search option of each interface, the types of advanced search options offered, whether or not fielded searching of particular EAD elements is allowed, the number and type of fields that can be searched, and the terminology used to label fielded search options. In addition, this study develops categories to describe differences in the display

of search results sets in each system. Categories of search results display features include: highlighting of matched words (hits), inclusion of relevant "snippets" from finding aids, and the number of metadata fields displayed and the labels used to describe those fields. Additional features considered include options for refined searching, saving search results, and sorting search results by different criteria. In the process of examining each search interface, categories were added and refined and data was reclassified until all the relevant features of a system could be adequately categorized

Using an open-coding iterative approach to categorize search functions and other features of XML-based EAD retrieval interfaces builds on previous research conducted by Zhou on more general finding aid search interfaces. Borrowing certain aspects of Zhou's methodology, findings of this research should help document the evolution of EAD search interfaces and platforms since 2003 as well as the degree to which interfaces have implemented the recommendations of usability studies by Prom, Yakel, and others.[4] Hopefully, this research will illuminate trends in the development of EAD search interfaces and provide a foundation for future, and much-needed, research to evaluate whether or not these trends have improved access to archival information.

One advantage of the content analysis method used in this study is a high degree of reliability given that the interface characteristics and search features examined represent manifest content that, for the most part, can be classified into clear categories. As a result, this study could be replicated at a later date with the same sample or with a larger sample to provide a comparative analysis of the state of EAD search interfaces over time. There are challenges, though, in the coding process. Although most features clearly fell into certain categories, some could not be categorized as easily and were

forced into existing categories. As a result, study results may imply more uniformity among EAD search interfaces than actually exists.

The sample for this study was drawn from a list of 86 EAD implementer websites currently available on the EAD Help Pages, a website and resource maintained by the Society of American Archivists (SAA).[5] This self-reported list of 86 EAD implementers represents a large portion of the population of institutions and repositories that have successfully implemented EAD programs. It is likely, however, that many EAD implementers have not reported their activities on the EAD Help Pages. As a result, these implementers are not among the 86 listed. Moreover, while the total population of EAD implementers is probably larger than 86, the number that have actually employed XML-based EAD retrieval systems is decidedly smaller, and the majority of these are likely to be early EAD implementers that *are* listed among the 86. Given the difficulty of identifying the entire population of EAD implementers, this research only examined the 86 sites listed on the EAD Help Pages. Of the 86 sites examined, 33 utilized some type of XML-based search and retrieval interface. These 33 formed the sample for this analysis. The names of repositories and consortium sites in the sample are listed in Appendix A.

Many of the 86 EAD implementer sites examined did not provide any search functionality, as Zhou reported in earlier research. Other sites did allow searching, but did so via a Google Search appliance or a similar commercial search engine imbedded in a website. Sites that used Google or similar search engines were excluded from the sample. In addition, several EAD implementers currently participate in consortium projects and provide access to their finding aids through a shared interface such as the

Online Archive of California (OAC) or the Northwest Digital Archives (NWDA). Because many of the 86 EAD implementers contribute their finding aids to the same consortium, they share the same EAD search interface. Thus, even while this study only considers 33 interfaces, many of these interfaces represent consortia projects with several contributing repositories. Although there are limitations to the type of purposive sampling conducted in this study, the sample chosen is representative of the range of features and functions currently available in XML-based EAD retrieval systems.

This study considers each search interface in the sample as a unit of analysis and examines a number of variables across search interfaces. Variables are grouped into four broad categories: Search modes, Display of Search Results, Display of Finding Aids, and Other Interface Features. Each category represents one of the central functions of an EAD retrieval interface—the search page, search results display page, display of individual finding aids, and any browse options available (see Appendix B for a fuller description of the variables in each category).

Because data collected consist mostly of discrete, nominal categories such as the existence of a particular fielded search option on a search page, or the presence of a particular metadata field in the search results display page, measurement techniques were relatively straightforward. For example, measurement of most variables fell into one of two categories: (1) A simple yes or no indicator that a feature was either present or not present in an interface, or (2) the particular term or phrase used to describe a feature of the interface. To measure other variables like browsing options and technologies used, more descriptive notes were recorded. Descriptive notes were then used to draw comparisons between each interface and to highlight any unusual or remarkable

characteristics of an interface. In sum, data analysis generally involved basic tallying of yes and no indicators for many variables, comparing terminology used in each interface, and synthesizing other more general observations about features of each interface. It should be noted that some variables were added or refined in the course of data collection, and interfaces were reexamined once all variable categories were finalized.

Each of the 33 interfaces in the sample was examined during a two week period (March 1-16, 2008). Examination began at the default search page of each interface, typically accessed from either the homepage of each repository or consortium website or by clicking on a link such as "search online finding aids." The presence of certain search features (basic search options, fielded search options, etc.) of the default search page were recorded before navigating to an advanced search page if one was available. Similarly, features of the advanced search page were recorded before navigating to any additional search pages. If the interface included fielded search options (usually in a drop-down menu), the term used to indicate the fielded search option was recorded (name, place, creator, scope and content note, etc.).

After examining all of the search features (basic, advanced, expert) of a particular interface, simple queries (usually a place name) were entered in the basic search page to examine how search results were displayed in each interface. Features noted included whether or not an interface provided collection level or series level results, the number and type of metadata fields displayed for each result (title, creator, abstract, etc.), as well as any other notable features such as the presence of "keywords in context" or highlighted search terms.

By clicking on a link in the search results display, individual finding aids were retrieved to collect information such as: the type of default finding aid views offered, any alternative views available, options for searching within finding aids, and any other notable navigation features provided from within each finding aid. During the entire process, any other noteworthy feature or function of each interface was noted such as options for browsing, the ability to refine searches, or the ability to save or email search results.

Finally, any information a site provided about technologies or system architectures was recorded along with more general comments or observations about each interface. Where possible, the total number of EAD finding aids included in each system was determined as well as the number of institutions or discreet repositories that contributed EAD finding aids to each system.[6]

FINDINGS AND DISCUSSION

Before interpreting the findings below, it is important to reiterate that the 33 EAD search interfaces examined in this study are those that employ some type of XML database platform. As a result, the sample considered here tends to provide more sophisticated features than non-XML search systems such as Google and other simple full-text only retrieval systems. In addition, the way in which search results are displayed tends to vary more across XML-based search interfaces, as they allow for more customization than popular search engines like Google. While some XML-based EAD interfaces like Harvard's OASIS (Online Archival Search Information System) are very sophisticated and provide several fielded search options in a highly configured interface,

other systems are relatively simple like the one provided by the University of Chicago's Special Collections Research Center.

Just as the level of sophistication varies greatly even among XML-based EAD search systems, so too does the total number of finding aids indexed in each system. For example, the search system implemented at the Rutgers University Special Collections provided access to only 26 EAD finding aids, whereas the Virginia Heritage Database, a consortium project of 27 repositories, indexed 7,372 EAD finding aids. The mean number of EAD finding aids available in each system was 1,796, while the median number was 830. Only thirteen of 33 (39%) search systems provided access to more than 1,000 EAD finding aids, and 4 of 33 (12%) provided access to fewer than 100 finding aids. In many instances, repository websites noted that only a fraction of their collections currently had EAD finding aids available in a search system. It should be noted, then, that the features and functions of the search interfaces described in this study pertain only to retrieval of information contained in a small number of EAD finding aids. In turn, these EAD finding aids represent only a small portion of the total number of archival collections held by repositories described in this study. While many repositories have plans to encode all of their finding aids in EAD, others have been reluctant to create EAD finding aids for all of their collections, especially smaller ones.

Search Modes

Findings in this section describe the variety of search modes and search options available across the 33 EAD interfaces examined with particular emphasis on fielded search options. With one exception, the default search mode of every interface examined

was a full-text search of the entire EAD finding aid.[7] The majority of interfaces provided a single search box on the main search page for conducting full-text searches. Some repositories provided drop-down menus on the main page for more advanced fielded search options, while others had an entirely separate "advanced search" page that included more sophisticated search modes. For the purposes of this study, any search mode other than a simple full-text search was considered an advanced search mode. Of the 33 interfaces examined, twenty (61%) provided some type of advanced search mode.

Perhaps the most common of these advanced search modes was some form of fielded searching. Fielded searching allows users to narrow searches to text that occurs within a particular EAD element or set of elements such as <persname> or <bioghist>. In theory, fielded searching allows for more precise retrieval of information in an EAD finding aid. For example, conducting a fielded search of only the contents of the <persname> element should ensure that the body of terms searched will only include names and not other text in a finding aid that may match a search query but might not be a name. Given the XML structure of EAD and the capabilities of native XML databases, it is possible to index the contents of every EAD element and to allow users to narrow searches to text that occurs specifically within any given element. As the findings in Table 1 indicate, however, current EAD search systems generally allow fielded searching of only a few select EAD elements. Table 1 notes the frequency of different fielded search option across the sample and the percentage of interfaces that included that option. Not surprisingly, the most popular fielded search option across all interfaces was "Title" or "Collection Title" (58%), while the least common were "Front Matter" (3%) and "Folder/Item Title" (3%).

Identifying what EAD elements search interfaces choose to index reveals how archivists and system designers expect users to search for information in EAD finding aids. For the most part, fielded search options in EAD search systems currently fall into two broad categories, both of which allow users to narrow searches to different kinds of information. One category of fielded searching restricts searches to individual elements (<persname> for names, <geogname> for places, <subject> for subjects, etc.). Another category of fielded searching restricts searches to larger portions or sections of finding aids, such as the Biographical/Historical Note section (<bioghist>) or the Container Listing (<dsc>). In this latter category, the indexed content is more narrative, instead of a single term or phrase that has been tagged in the first category such as a person, place, subject, etc.

Even though interfaces in the sample provided the same basic types of fielded search options, interfaces tended to assign different labels to these options. As Table 1 illustrates, a fielded search option that searched contents of the <origination> element might be labeled variously as "creator" or as "author." Likewise, a fielded search option that searched content of the <corpname> element may be labeled as "corporate name" in one system, but "organization name" in another. This inconsistency in the terms assigned to each fielded search options suggests that systems have not implemented the recommendations of Prom (2004) and Yakel (2004) to label fielded search options consistently and with terms that are intelligible to the average user. The persistence of archival jargon in fielded search labels such as "front matter" or "geographic name" certainly does not make EAD retrieval systems any more transparent for most users.

In any case, it is likely that most users never choose fielded search options, but instead conduct simple full-text searches. As a result, the inconsistencies in fielded search options provided and the labels assigned to those options might be irrelevant for the majority of users. Nevertheless, in building archival search systems, archivists and system designers should carefully consider the terminology used to label system features and functions. In turn, archivists at different repositories should collaborate to ensure that the same terms are used consistently across all search interfaces.

Table 1. Fielded Search Options, Terms Used, and Frequency (n=33)

Fielded Search Options Labels Used (in quotes)	**Frequency**	**% of Interfaces**
Title		
"Collection Title"	11	
"Title"	8	
Total Title	**19**	**58%**
Subjects		
"Subjects"	15	
"Subject Headings"	1	
Subjects Total	**16**	**48%**
Creator		
"Creator"	6	
"Provenance"	1	
"Author/Creator"	2	
"Author"	1	
Creator Total	**10**	**30%**
Names		
"Names"	4	
"Personal Name"	1	
"Name or Corporate Name"	1	
"Personal Name and Family Name"	2	
"All Names"	1	
Names Total	**9**	**27%**
Collection Overview		
"Description"	4	
"Scope and Content"	4	
"Collection/Organization Info.	1	
Collection Overview Total	**9**	**27%**

(Table continued on next page)

(Table 1 continued)

Biographical Note		
"Biography"	1	
"Biography or History Portion"	1	
"Biographical/Historical Note"	5	
"History Note"	1	
Biographical Note Total	**8**	**24%**
Call Number		
"Call Number"	4	
"Collection Number"	3	
"Reference"	3	
Call Number Total	**8**	**24%**
Places		
"Places"	6	
"Geographical Name"	2	
Places Total	**8**	**24%**
Container List		
"Container List"	2	
"Collection Inventory"	2	
"Box Inventory"	1	
"Collection Components"	1	
Container List Total	**6**	**18%**
Dates / Date Ranges		
"Date Range" (2 boxes)	3	
"Dates" (1 box)	2	
"Year" (1 box)	1	
Dates Total	**6**	**18%**
Corporate Names		
"Corporate Name"	4	
"Organization Name"	1	
Corporate Name Total	**5**	**15%**
Genre / Format		
"Form of Material"	3	
"Material Type"	1	
"Genre/Form"	1	
Genre / Format Total	**5**	**15%**
Other		
"Front Matter"	1	
"Folder / Item Title"	1	

Display of Search Results

Because many EAD retrieval interfaces are designed or customized by each repository or consortium, the display of search results varies considerably across interfaces. While the implementation of EAD has standardized the components and

content of finding aids across repositories to a degree, the archival community has developed no standard for how search results are displayed in EAD search interfaces. For example, some search systems, particularly those in Europe, returned search results at the series level or item level of EAD documents rather than at the collection level. In the University of Liverpool's EAD search system, search results were returned at the series level and at the collection level, with an option to display only collection level results. Most EAD search interfaces in the U.S. only displayed search results at the collection level.

Table 2. Level of Granularity in Search Results (n=33)

Level of Granularity	Frequency	% of Interfaces
Collection Level	29	88%
Collection Level with Series Level Hits	3	9%
Mixed (Collection Level and Series Level)	1	3%

In addition to the level of granularity, the type of information included in search results also varied across EAD search interfaces in several ways. Some interfaces provided several metadata fields in search results sets such as: the title of a collection, an abstract, the repository where the collection was held, the extent of materials in the collection, etc. Other interfaces provided only titles of collections with no supporting metadata. Some interfaces provided additional features in the search results display such as the number of times a search term occurred in the finding aid (hits), or brief excerpts (snippets) from the finding aid where a search term occurred. Table 3 shows the popularity of various metadata fields in the search results and other information included in the display of search results.

It should be noted that some interfaces provided clear labels for all of the metadata fields included in search results sets, while others offered no field labels at all. Even when field labels were present, the terms used to identify the same type of information often varied from interface to interface. For example, to label information about the physical size of a collection some interfaces used the term "extent," while others used "size." Like variations in the labels used to identify fielded search options, variations also occurred in labeling types of information provided in search results. For the sake of simplicity, variations in field labels for search results are not indicated in Table 3 as in Table 1.

Table 3. Information Included in Search Results (n=33)

Metadata Fields and Other Information Included in Search Results Display	Frequency	% of Interfaces
Collection Title	32	97%
Abstract (Full or Partial)	23	70%
Repository	17	52%
Creator	14	42%
Extent	14	42%
Number of Hits in Finding Aid	11	34%
Keyword in Context (Snippets)	6	18%
Search Terms Highlighted	5	15%

Although Table 3 indicates that there were considerable differences in both the type of information and level of detail offered in search results display pages, it does not convey how much the display of search results varied visually across EAD search interfaces. Comparing two screenshots from the Center for Jewish History and the Massachusetts Historical Society reveals just how different search results sets may appear from one interface to the next (See Appendix C for screenshots). In the Center for Jewish History interface (Figure 1), each search result occupied almost an entire browser screen

and included several metadata fields such as: the creator of the collection, the collection title, inclusive dates materials, an ID number or call number, a lengthy abstract, the language of collection materials, the size of the collection, and the file size of the EAD finding aid. In total, the Center for Jewish History interface provided eight discreet metadata fields for each search result. With so much information presented, however, only two or three search results could be viewed in a single browser screen without scrolling. In contrast, the Massachusetts Historical Society interface (Figure 2) provided only the collection title in the search results set with no additional metadata fields. With only the title displayed, roughly ten times as many search results could be displayed in a single browser screen without scrolling. Across all interfaces in the sample, the mean and median number of fields displayed for each search result was 4.[8] The most common fields displayed were the collection title (97%), abstract (70%), and repository (52%).

The ability to sort search results by different criteria was another feature offered by some EAD retrieval interfaces. Nine of the 33 (27%) interfaces examined provided some form of sorting option. Sorting by title (24%), creator (18%), or relevance (18%) were the most popular options, while sorting by contributing institution (3%) was the least popular. Table 4 indicates the frequency of different sort options in the interfaces sampled.

Table 4. Frequency of Sort Options for Search Results (n=33)

Sorting Options for Search Results	Frequency	% of Interfaces
Any Sorting Option	9	27%
By Title	8	24%
By Creator	6	18%
By Frequency / Relevance	6	18%
By Collection Date	2	6%
By Contributing Institution	1	3%

Curiously, some interfaces like the University of Michigan's Bentley Historical Library did not appear to sort search results by any default criteria. While the Bentley interface did provide options for sorting by title, creator, and frequency, the default sorting method was "not sorted."[9] It would seem that offering some sort of default sorting method, especially relevance, might mimic what users have come to expect from their experiences with popular search engines like Google. Findings of this study affirm similar findings by Zhou (2006) that a number of interfaces failed to inform users how results were sorted. In most systems, search results appeared to be sorted by relevance. If relevance was the default sorting method, then failing to disclose sorting criteria might be excusable, as most users probably consider relevance a reasonable default sorting option. However, in interfaces that did not automatically sort results according to relevance such as the Bentley system, it was disconcerting that users received no feedback about how search results were sorted or not sorted.

In many ways, the ability to customize the display of search results is one of the most useful aspects of an XML-based retrieval system for EAD finding aids. While sophisticated fielded search options may help a small pool of expert users and archivists narrow their searches to specific EAD elements or portions of finding aids, the ability to configure the display of search results affects every user of an EAD retrieval interface. As usability studies by Prom (2004) demonstrate, most users have less success when presented with multiple search options (p. 250). In light of these findings, archivists and system designers should concentrate less on offering infrequently used and complicated fielded search options and more on configuring the display of search results in a way that maximizes users' ability to scan them quickly and make relevance judgments. In

addition, archivists should develop guidelines that address the appropriate level of detail and the types of field labels that should be incorporated into the display of search results. As findings of this study show, there is currently no clear standard for determining the type and extent of information included in search results display pages.

Finding Aid Display

The display of individual EAD finding aids depends largely on XSL stylesheets that are developed by each repository or mandated by a particular consortium project. Stylesheets determine how hierarchies are visually expressed, what types of navigation features are available from within a finding aid, and several other stylistic features that often make finding aids at one institution look very different from those created by another. While many finding aid display features are determined by stylesheets irrespective of a particular search interface, there are other display features that sometimes work in close connection with the search interface such as: displaying the location of search terms in a finding aid's navigation pane, highlighting search terms in the text of a finding aid, and allowing users to view certain pieces of finding aids without viewing the entire document. Because these display features are important characteristics of the larger EAD retrieval interface they are considered here.

Many interfaces examined in this study allowed users to view finding aids in several different ways—by section, by the entire document, or some combination of both. Table 5 shows the default finding aid display type for interfaces in the sample. For the purposes of this study, the default finding aid display was the view that appeared after clicking on the collection title or other link in the search result. Ten interfaces examined

did not have a default display option but instead prompted users to choose between two or more finding aid display types.

Table 5. Default Finding Aid Display Types (n=33)

Default Finding Aid Display	Frequency	% of Interfaces
Full View	11	33%
Section View or Outline	8	24%
Full View with Collapsed Component Section	2	6%
File Level View	2	6%
No Default, Two or More Options Offered	10	30%

The most common display type was the "full view" or "entire finding aid" view. In this view, the EAD finding aid was displayed in its entirety. The next most popular view was the "section view" or "outline view." In this view, only summary information in the finding aid was displayed after clicking on a link in a search result. Summary information provided in the "section view" and "outline view" usually included the title of the collection, collection number, creator, extent, and occasionally a brief abstract. Systems with a default "outline view" for finding aids typically allowed users to navigate to other sections of the finding aid through links in the navigation panel on the left side of the screen (see Appendix D, Figure 3 for a screenshot of the Online Archive of California's default section view).

Other default finding aid views included a hybrid of the full-view and section view options. In this hybrid view, finding aids were displayed in their entirety, but the component sections (<dsc>) were collapsed. Typically the component section includes folder lists, box lists, or other similar lists of items in a collection. Viewing a particular folder or item in this hybrid view required clicking on the top-level component (a series

or box) and continuing to click until all of the subcomponents of the finding aids hierarchy were expanded. While this type of "drill down" approach reduces the initial size of the finding aid displayed on the screen—an early concern for remote users with low bandwidth--it also introduces an unnecessarily frustrating level of complexity that forces users to proceed through a series of clicks before retrieving relevant information.

Instead of offering a "full view" or an "outline view" of a finding aid, some interfaces such as the Denver Public Library's PLEADE system directed users immediately to specific components in a finding aid that match their search query (see Appendix D, Figure 4). Thus, the PLEADE system's default finding aid view was usually a single file or item. Although the file or item containing the search term was stripped out of the finding aid and displayed separately, the interface attempted to communicate that file's position in the document's hierarchy by providing links to the subseries and series where the file or item occurred. This level of granularity offered by the PLEADE system's default finding aid view was unusual, but it illustrates the variety of display methods currently offered by EAD interfaces and the difficulties some users might encounter when trying to interpret finding aids across several interfaces.

While the file-level display was the default finding aid view in the Denver Public Library's PLEADE interface, some interfaces provided other options for directing users immediately to the portions of finding aids where there search terms occurred instead of requiring users to scan the entire text of a finding aid. Interfaces achieved this functionality in several ways. One particularly innovative method was employed by the Online Archive of California, Emory University, and the University of Chicago. In these three interfaces, the location of each occurrence of a search term was indicated in the

finding aid's navigation pane. For example, if a search term occurred in several portions of a finding aid--Biographical Note, Collection Overview, Series, Subseries, etc.—the number of times the term occurred in each section was indicated with a simple red number in parenthesis next to the link for that section of the finding aid in the navigation pane. Highlighting the location of search terms in a finding aid's navigational links is an effective strategy for preserving the important hierarchical and contextual information communicated by a full finding aid, but at the same time giving users a visual representation of where their search terms appear in the finding aid's hierarchy (see Appendix C, Figure 3 for a screenshot of this feature in the Online Archive of California).

Rutgers University's EAD interface offered another method for directing users immediately to the location of their search terms in a finding aid. After clicking on a search result, the Rutgers interface loaded a full finding aid, but immediately jumped to the position in the finding aid where the search term occurred. Users could then scroll up or down for contextual information surrounding the search term. Jumping to the location of the search term in the context of the finding aid is similar to performing a "Control F" in a web browser, but the built in functionality is probably useful for more inexperienced users who many not be familiar with the Control F function.

Nine interfaces (27%) used yet another method--the "keywords in context" view--to communicate the location of search terms in a finding aid. The "keyword in context" feature was especially prevalent in interfaces built on the DLXS platform. DLXS interfaces typically provided links for three different finding aid display options below each search result—one for viewing the entire finding aid, another for viewing an

"outline view" or "section view," and a third for viewing "keywords in context."[10] Selecting the "keyword in context" option extracted all of the file-level portions of the finding aid that included a given search term and also provided links to any higher level series or subseries of which the relevant file was a part (see Appendix D, Figure 6 for an example of the "keywords in context" display option in the Cornell University interface). Other methods for indicating the location of search terms included highlighting search terms in the finding aid (51%) and the ability to search the text of a finding aid (24%).

Table 6: Methods for Indicating the Location of Search Terms in a Finding Aid (n=33)

Method	Frequency	% of Interfaces
Search Terms Highlighted in Text of Finding Aid	17	51%
Keywords In Context View Option Provided	9	27%
Ability to Search Within a Finding Aid	8	24%
Location of Search Terms Displayed in Navigation Pane	3	9%
Automatically Jump to Search Term in Full Finding Aid	1	3%

As the findings in Table 6 indicate, EAD search interfaces currently employ several strategies for communicating to users where search terms occur in a finding aid while at the same time preserving the important contextual information provided by the hierarchical structure of a finding aid. Achieving both of these two goals simultaneously--directing users to relevant portions of finding aids *and* preserving the context of the entire collection—is one of the greatest challenges for designers of EAD search systems. Usability studies are clearly needed to determine which strategy, if any, is most effective in achieving these twin goals.

Other Features of EAD Retrieval Interfaces

Some additional features of the interfaces identified in this study included the ability to refine searches or search within search results (15%), the ability to save search results in a "bookbag" or "portfolio" (12%), and the ability to email or otherwise export search results (15%). Interfaces that provided these types of features tended to function in many of the same ways as traditional online library catalogs. In addition, all four interfaces that allowed saving search results were built on the DLXS platform.

Browsing Options

The presence of fielded search options and highly customized search results displays are the hallmark of most XML-based retrieval interfaces for EAD finding aids. However, as studies of the information seeking behavior of humanists, social scientists, and even on-line shoppers suggest, more people may prefer to browse for archival information than to conduct sophisticated fielded searching of EAD finding aids.[11] If this is the case, the limited number of browse options currently available in most EAD retrieval interfaces is alarming. It should be noted that some repository websites do provide browse options separate and apart of their EAD retrieval interfaces. These high-level browse options sometimes take the form of pathfinders that include summaries of the repository's holdings in a given topical area (Civil War, Agriculture, etc.). These pathfinder-like pages may provide references or links to selected collections that archivists have deemed relevant to a particular topic.

Aside from these top-level browsing options, few repositories have implemented any sort of comprehensive topic browsing for all of their EAD finding aids. Table 7

shows the frequency of different browse options provided by interfaces in this sample. Although almost every interface allowed users to browse finding aids alphabetically by collection title, relying solely on title browsing is particularly problematic for archival materials. Titles of archival collections are generally constructed according to a collection's provenance and they usually bear the name of the individual or organization that created the collection. As a result, collection titles for archival materials are not "known-items" and they usually say very little about the topical content of a collection. For instance, if someone is interested in material on a broad topic such as the history of medicine or aviation, it is probably not useful to browse a list of collection titles like Smith Family Papers or Jones Family Papers.

Table 7. Browsing Options (n=33)

Browse Options Offered	Frequency	% of Interfaces
Alphabetical Lists (Collection Titles or Creators)	32	97%
Local Subjects / Categories	8	24%
LCSH Subjects	2	6%
Place	2	6%
Finding Aids with Digital Content	1	3%
Collections By Date	1	3%
Names	1	3%
Format (Photos / Mss.)	1	3%
Most Viewed Finding Aids	1	3%

To provide users with some way to browse topically across EAD finding aids, eight repositories (24%) used locally generated subject categories to augment the search functionality on their site. For example, the Rocky Mountain Online Archive, a consortium of 26 repositories, offered browsing by broad subjects such as "education," "journalism," "railroads," and "ranching," as well as by places like Colorado, New Mexico, or Wyoming. In contrast, ArchivesUM, the University of Maryland's EAD

interface, included more specific browsing options like “band history,” “horse racing and breeding,” and even a few selected individuals who figured prominently in their collections.

To manage these browsing categories, some consortium projects, like the Northwest Digital Archives, maintain a controlled list of acceptable categories and required participating repositories to apply at least one of these categories to every finding aid that they upload to the system.[12] Locally created categories and topics are typically encoded in a finding aid alongside more broadly used controlled vocabularies such as Library of Congress Subject Headings (LCSH) and TGM (Thesaurus for Graphic Materials) terms. When these broader locally generated browsing categories are applied consistently across all finding aids in a system, they can be harvested to generate more user-friendly browsing options in an EAD interface. Moreover, developing and maintaining a manageable list of locally generated topics and categories can serve two purposes. On one hand, providing users with a brief list of topical browsing categories can give them a sense of the type and scope of materials represented by the body of EAD finding aids in a given system. On another, browsing categories can give users a way to start exploring finding aids in a repository even if their research questions are not well-defined.

The third most popular browsing option across interfaces in the sample was the ability to browse by Library of Congress Subject Headings (LCSH). Even though many EAD implementers included a list of LCSH terms in the <controlaccess> portion of EAD finding aids, only two EAD interfaces in this sample leveraged these LCSH terms to generate any sort of browsing option. These two institutions were—not surprisingly—the

Library of Congress and the University of Chicago. Clicking on the University of Chicago's "browse by subject" feature linked to a page where users could browse all of the LCSH terms used throughout the corpus of finding aids in the retrieval system. The University of Chicago's "browse by subject" feature was particularly notable because for each LCSH term, it displayed the number of times that term was used across all finding aids in the system. With 634 EAD finding aids available in the University of Chicago's interface, the total number of browsable LSCH terms was manageable. However, if there were 10,000 finding aids in the system, browsing by LCSH terms might not be an efficient way to access the collection. Findings of this study seem to support this notion, as large consortium EAD search systems containing larger number of finding aids tended to provide a smaller set of more general and locally generated browsing options instead of allowing browsing of more granular LCSH terms. It should also be noted that, unlike MARC-AMC records in bibliographic systems, few EAD interfaces provided linkable LCSH terms in finding aids. Instead, LCSH terms usually appeared as simple lists in EAD finding aids.

The absence of topical browsing options in the majority of EAD interfaces is perhaps one of the most discouraging findings of this research, especially given that the results of usability studies indicate that most users prefer browsing over searching when they are looking for archival materials specifically (Prom, 2004) or when they are engaged in other information-seeking activities on the web (Straub, 2005). Ongoing research into faceted browsing and other methods for mapping LCSH terms to more general categories should prove useful for enhancing the browse options currently offered by EAD retrieval systems.[13] Currently, however, the lack of adequate browsing options

in EAD retrieval interfaces is one of the greatest barriers to the discovery and use of archival materials.

CONCLUSION

Despite some limitations, this research has important implications for assessing the current state of access to archival information. In particular, this study explores whether or not some of the early promises of EAD have been fulfilled. Because much of the early excitement over EAD focused on its ability to support sophisticated searching, navigation, and display, this study demonstrates how XML-based EAD search interfaces are currently delivering these features. Findings may also educate repositories that are considering XML-based EAD retrieval systems.

Given the considerable effort required to encode finding aids in EAD, archivists should expect that their encoding labors are having a marked improvement on researchers' ability to discover and use archival information. In contrast to more efficient methods for providing online access to finding aids such as encoding them in HTML or presenting them as PDF files, EAD allows archivist to "tag" certain text within a finding aid, and in the process, give that text semantic meaning. Such tagging practices bring with them the implicit expectation that information contained within these tags will be indexed, searchable, browsable, and otherwise presented to users in meaningful ways that facilitate discovery. Because less than half of the interfaces examined in this study indexed EAD elements to provide fielded search options, and even fewer used EAD elements to develop browsing categories, this research begs the question: Is all that tagging really worth it?

The analysis of search results display pages and finding aid display options indicates that there is some cause for optimism. A number of EAD retrieval interfaces are making use of value-added encoding to customize how search results are displayed. Likewise, some interfaces have developed innovative methods for identifying the location of search terms in a finding aid, but at the same time preserving the important contextual and hierarchical organization communicated by a traditional finding aid. As XML technologies become more pervasive on the web, it is likely that other industries will develop more sophisticated retrieval systems that might serve as useful models for EAD interfaces. The recent implementation of the Endeca platform for some research library catalogs is one example of this trend.[14]

At some point in the near future, the growth of Web 2.0 technologies may render some of the findings of this study irrelevant. As more and more archival repositories and libraries begin pushing their content into existing web-based platforms like Flickr, Footnote, or Google Books instead of building their own content management systems, it is quite possible that systems may emerge organically from outside of the archival community that allow users to share, distribute, and repurpose information contained in EAD finding aids in innovative ways.[15]

Because the technical skills of archivists vary considerably, and because many archivists face growing backlogs of collections that need processing, few in the profession have time to fully investigate the technological issues addressed in this study. At the same time, users of archives are becoming increasingly web-savvy and they are demanding more sophisticated online discovery tools. By describing the current state of EAD interfaces, this study gives archivists the knowledge necessary to begin evaluating

how the profession is currently meeting these emerging user demands. By providing some concrete quantitative data about the types of features and functionality currently offered in EAD retrieval interfaces, this study can help archivists form some assumptions about whether or not the promises of EAD have been realized and whether or not current encoding practices need to be reconsidered. At the same time, future research is needed on a number of fronts. Usability studies are needed to explore user interactions with increasingly sophisticated EAD retrieval interfaces. In addition, more classical information retrieval studies are needed to evaluate the relevance and precision currently offered by XML-based EAD interfaces.

Despite the widespread adoption of EAD and the development of the XML-based EAD retrieval interfaces considered in this study, the fact remains that there are few systems currently in place that approach the mature and broadly accessible "archival information system" that Gilliland-Swetland envisioned in 2001. On the verge of the ten year anniversary of EAD, it is still unclear whether this type of system is even possible. Nevertheless, archivists should persevere. Just as the archival community initiated the development of EAD as an encoding standard in the 1990s, the critical need for similar standards for the functionality of a broadly accessible EAD retrieval system should demand the attention of archival professionals. If archivists fail to reach a consensus on these important issues, users of archives will suffer, cast adrift in a sea of competing and confusing EAD retrieval interfaces that often do more to impair than to improve the discovery of the materials that archivists so diligently collect, preserve, and describe.

NOTES

[1] For a discussion of the development of EAD as a standard see: Pitti, D. (1997). Encoded Archival Description: The Development of an Encoding Standard for Archival Finding Aids. *The American Archivist,* 60, 268-283; for a discussion of implementation issues see: Yakel, E. & Kim, J. (2005). Adoption and Diffusion of Encoded Archival Description. *Journal of American Society for Information Science and Technology (JASIST)*, 56, 1427-1437.

[2] Finding Aids database of the Online Archive of California, <http://www.oac.cdlib.org/>, accessed on 21 September 2007.; Finding Aids database of the Northwest Digital Archive, < http://nwda-db.wsulibs.wsu.edu/nwda%2Dsearch/>, accessed on 21 September 2007.

[3] Information on the development, implementation, and documentation of XTF can be found at: <http://www.cdlib.org/inside/projects/xtf/>, accessed on 21 September 2007.

[4] Although Zhou's study was published in 2006, she notes that all data from her research "came from the investigation conducted at the end of 2003" (p. 105.) In the three years between Zhou's data collection and publication, there were likely numerous changes to the EAD interfaces she examined.

[5] EAD Implementers are listed on the EAD Help Pages at <http://www.archivists.org/saagroups/ead/implementors.html>

[6] Some search interfaces listed the number of finding aids available, but in some cases the number of finding aids was determined by either a) cutting and pasting alphabetical browse lists into a word document and counting the lines, or b) conducting a search without entering a term into the search box and counting the number of search results. The study was unable to determine the total number of finding aids for three of the 33 interfaces in the sample.

[7] The default access method for EAD finding aids at the University of Maryland (ArchivesUM) was a series of three drop down menus with various browsing options (e. g. alphabetical, unit/department, subject); See <http://www.lib.umd.edu/archivesum/about.jsp>

[8] Some interfaces clearly labeled metadata fields included in search results, while other did not. In addition, some interfaces strung several metadata fields together in one line of text.

[9] For unsorted search results see: Bentley Historical Library, University of Michigan, <http://quod.lib.umich.edu/cgi/f/findaid/findaid-idx?&page=simple&c=bhlead>

[10]Seven EAD search interfaces examined in this study used the DLXS (Digital Library eXtension Service) platform developed by the University of Michigan. These include: Bentley Historical Library at the University of Michigan, Cornell University, the Kentuckiana Digital Library, the University of Minnesota, the University of Pennsylvania (Annenberg Rare Book and Manuscript Library), the University of Pittsburgh (Historic Pittsburgh Finding Aid Project), and the University of Wisconsin. For information on DLXS see: <http://www.dlxs.org/>

[11] For research on the information seeking behavior of social scientists see: Meho, L. I., & Tibbo, H. R. (2003). Modeling the Information-Seeking Behavior of Social Scientists: Ellis's Study Revisited. *Journal of the American Society for Information Science and Technology* 54, 570-587. For research on on-line shoppers' preference for browsing over searching see: Katz, M. A. & Byrne, M. D. (2003). Effects of Scent and Breadth on Use of Site-specific Search on E-Commerce Web Sites. *ACM Transactions on Computer-Human Interaction*, 10(3), 198-220.

[12] For the list of 111 acceptable browsing categories developed by the Northwest Digital Archives Consortium see: <http://orbiscascade.org/index/cms-filesystem-action?file=nwda/browsingtermsalphajan302006.pdf>

[13] Development of the Simple Knowledge Organization System (SKOS) holds promise for improving browsing of hierarchical classification schemes and controlled vocabularies like LCSH. For information about SKOS see: <http://www.xml.com/pub/a/2005/06/22/skos.html>.

[14] The Endeca platform was initially developed to serve corporate clients, but has been adopted by a number of libraries. Endeca provides innovative guided navigation and browsing of bibliographic holdings. Currently, North Carolina State University and the Triangle Research Library Network's Search TRLN interface have implemented Endeca. While current implementations of Endeca leverage existing MARC data, the same technologies show promise for unlocking the potential of XML documents like EAD.

[15] Recently the Library of Congress had begun contributing photographs from their collections to Flickr and the National Archives and Records Administration (NARA) currently has an agreement with Footnote to digitize and host selected records. While these types of web-based collection management sites currently provide mostly digital surrogates of original documents, it might be possible for similar web-based finding aid repositories to emerge where users could browse, share, annotate, and repurpose information contained in EAD finding aids. For information about Library of Congress photographs on Flickr see: <http://www.flickr.com/photos/Library_of_Congress>; for NARA records in Footnote see: <http://go.footnote.com/nara/>

BIBLIOGRAPHY

Babbie, E. (2007). *The Practice of Social Research.* Eleventh Edition. Belmont, CA: Thomson Wadsworth.

Betsi, S, Lalmas, M., & Tombros, A. (2006). User Expectations from XML Element Retrieval. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 611-612.

Brown, C. B. & Schottlaender, B. E. (2001). The Online Archive of California: A Consortial Approach to Encoded Archival Description. In *Encoded Archival Description on the Internet*. Binghamton, N.Y.: Haworth.

Cornish, A. (2004). Using a Native XML Database for Encoded Archival Description Search and Retrieval. *Information Technology and Libraries*.

Feeney, K. (1999). Retrieval of Archival Finding Aids using World-Wide-Web Search Engines. *The American Archivist*, 62, 206-228.

Fuhr, N., Malik, S., & Lalmas, M. (2004). Overview of the Initiative for the Evaluation of XML Retrieval (INEX) 2003. *Proceedings of the Second INEX Workshop*.

Gilliland-Swetland, A. J. (2001). Popularizing the Finding Aid: Exploiting EAD to Enhance Online Discovery and Retrieval in Archival Information Systems by Diverse User Groups. In *Encoded Archival Description on the Internet.* Binghamton, N.Y.: Haworth, 199-225.

Hensen, S. L. (2001). Archival Cataloging and the Internet: The Implications and Impact of EAD. *Journal of Internet Cataloging*, 4, 75-95.

--------. (1986). The Use of Standards in the Application of the AMC Format. *American Archivist*, 49, 31-40.

Katz, M. A., & Byrne, M. D. (2003). Effects of Scent and Breadth on Use of Site-specific Search on E-Commerce Web Sites. *ACM Transactions on Computer-Human Interaction,* 10(3), 198-220.

Kim, J. (2004). EAD Encoding and Display: A Content Analysis. *Journal of Archival Organization,* 2, 41-55.

Larson, B., Tombros, A, & Malik, S. (2006). Is XML Retrieval Meaningful to Users?: Searcher Preferences for Full Documents vs. Elements. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 663-664.

Meho, L. I., & Tibbo, H. R. (2003). Modeling the Information-Seeking Behavior of Social Scientists: Ellis's Study Revisited. *Journal of the American Society for Information Science and Technology* 54, 570-587.

Pitti, D. V. (2006). Technology and the Transformation of Archival Description. *Journal of Archival Organization,* 3, 9-22.

--------. (1999). Encoded Archival Description: An Introduction and Overview. *D-Lib Magazine* 5(11), < http://www.dlib.org/dlib/november99/11pitti.html>.

--------. (1997). Encoded Archival Description: The Development of an Encoding Standard for Archival Finding Aids. *American Archivist,* 60, 268-283.

Prom, C. (2007). A Unified Platform for Archival Description and Access. *Proceedings of the 2007 Conference on Digital Libraries*, 157-166.

--------. (2004). User Interactions with Electronic Finding Aids in a Controlled Setting. *American Archivist,* 67, 234-268.

Rogers, E. M. (1962). *Diffusion of Innovations*. New York: Free Press of Glencoe.

Roth, J. M. (2001). Serving Up EAD: An Exploratory Study on the Deployment and Utilization of Encoded Archival Description Finding Aids. *American Archivist,* 64, 214-237.

Smith, C. A. (2003). The Effect of XML Markup on Retrieval of Clinical Documents. *American Medical Informatics Association Annual Symposium Proceedings,* 614-618.

Smith, M. (2000). DFAS: The Distributed Finding Aid Search System. *D-Lib Magazine*, 6(1).

Straub, K. (2005). The Answer You're Searching for is 'Browse'. *UI Design Newsletter*.

Yakel, E. (2004). Encoded Archival Description: Are Finding Aids Boundary Spanner or Barriers for Users. *Journal of Archival Organization*, 2, 63-77.

Yakel, E., & Kim, J. (2005). Adoption and Diffusion of Encoded Archival Description. *Journal of American Society for Information Science and Technology (JASIST)*, 56, 1427-1437.

Zhou, X. (2006). Examining Search Functions of EAD Finding Aids Web Sites. *Journal of Archival Organization*, 4, 99-118.

APPENDIX A: Sample of 33 EAD Interfaces Examined

Taken from a list of 86 EAD implementers at:
<http://www.archivists.org/saagroups/ead/implementors.html>

INSTITUTION (repository)	**LINK TO EAD SEARCH INTERFACE** (all links accessed March 1-16, 2008)
Arizona Archives Online	http://aao.lib.asu.edu/index.html
Bentley Historical Library	http://quod.lib.umich.edu/cgi/f/findaid/findaid-idx?&page=simple&c=bhlead
Brown University	http://dl.lib.brown.edu/bamco/
Cambridge (UK)—Janus	http://janus.lib.cam.ac.uk/
Center for Jewish History	http://www.cjh.org/collections/findingaids.php
Cornell University Archival Guides	http://dlxs.library.cornell.edu/cgi/f/findaid/findaid-idx?page=index&c=rmc
Denver Public Library (Western History and Genealogy)	http://history.denverlibrary.org/
Duke University	http://library.duke.edu/digitalcollections/rbmscl/inv/
Emory (MARBL)	http://marbl.library.emory.edu/FindingAids/rqst.php
Five Colleges Archives and Manuscripts Collections	http://asteria.fivecolleges.edu/index.html
Harvard University (OASIS)	http://oasis.lib.harvard.edu/oasis/deliver/advancedsearch?_collection=oasis
Kentuckiana Digital Library	http://kdl.kyvl.org/cgi/f/findaid/findaid-idx?;page=simpleext;xc=1
Library of Congress	http://lcweb2.loc.gov/faid/
Massachusetts Historical Society	http://www.masshist.org/findingaids/
New York University Archives and Special Collections	http://dlib.nyu.edu/findingaids/?fq=collectionId%3Afales

Northwest Digital Archive	http://nwda.wsulibs.wsu.edu/index.shtml
Online Archive of California	http://www.oac.cdlib.org/
Rocky Mountain Online Archive	http://rmoa.unm.edu/index.php
Rutgers University Special Collections and University Archives	http://www2.scc.rutgers.edu/ead/
Southern Utah University (SUUPER Search)	http://archive.li.suu.edu/archive/
Syracuse University Special Collections Research Center	http://library.syr.edu/information/spcollections/findingaids/index.html
Texas Archival Resources Online (TARO)	http://www.lib.utexas.edu/taro/index.html
United Methodist Church Archives	http://archives.gcah.org:8080/exist/archives/gcahcat.xml
University of Chicago Library	http://ead.lib.uchicago.edu/
University of Liverpool Special Collections and Archives	http://sca.lib.liv.ac.uk/ead/index.html
University of Maryland Archives and Special Collections	http://www.lib.umd.edu/archivesum/index.jsp
University of Minnesota	http://discover.lib.umn.edu/findaid/
University of Pennsylvania (Annenberg Rare Book & Manuscript Library)	http://ead.library.upenn.edu/cgi/f/findaid/findaid-idx?page=index&c=findaid
University of Pittsburgh (Historic Pittsburgh Finding Aids Project)	http://digital.library.pitt.edu/ead/
University of Vermont Special Collections	http://cdi.uvm.edu/findingaids/index.xql
University of Virginia (Virginia Heritage Database)	http://ead.lib.virginia.edu/vivaead/eadform.pl
University of Wisconsin Digital Collections	http://digicoll.library.wisc.edu/cgi/f/findaid/findaid-idx?page=home;c=wiarchives;cc=wiarchives
Yale University	http://webtext.library.yale.edu/finddocs/fadsear.htm

APPENDIX B: Variables by Category

***Search Options*:**

1. Search modes offered: Default search, advanced search, or expert search
2. Fielded searching options:
    a. Types of fielded searching offered: title, subject, creator, name, place, date, scope and content note, biographical note, component level information, etc.
    b. Terms used to label fielded searching options: subject vs. topic, name vs. person, etc.

Display of Search Results

1. Granularity of search results: collection level, series level, item level, or other
2. Number and type of metadata fields included in retrieval sets (e.g. title, creator, date, abstract, repository, etc.)
3. Presence of other features in search results sets: highlighted search terms, keyword in context or "snippets," sorting options, and relevance ranking options

Display of Finding Aids

1. Default finding aid display: Full finding aid, section view, item-level view
2. Finding aid views offered (same as above)
3. Ability to search within the text a finding aid
4. Other navigation options offered
5. Highlighting of search terms in the finding aid

Other Interface Features

1. Total number of finding aids included
2. Number of institutions or repositories contributing finding aids to the search system.
3. Ability to conduct refined searches, or search within search results.
4. Ability to save search results, or email search results or finding aids
5. Browsing options offered: Browsing by locally created subjects, browsing by Library of Congress Subject Headings, lists of collection creators, time periods, etc.)

APPENDIX C: Display of Search Results, Example Screenshots

Figure 1. Display of Search Results at the Center for Jewish History

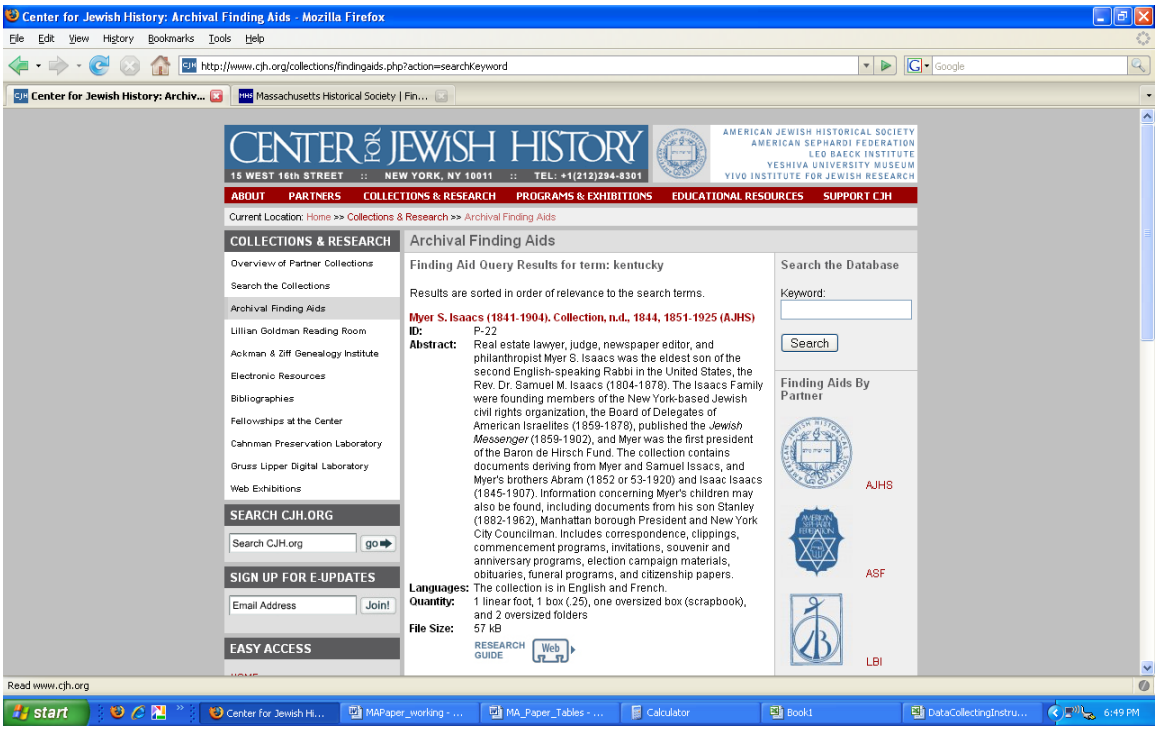

Figure 2. Display of Search Results at Massachusetts Historical Society

Appendix D: Finding Aid Display Features, Example Screenshots

Figure 4. Default Section View of Finding Aid, Online Archive of California (OAC)

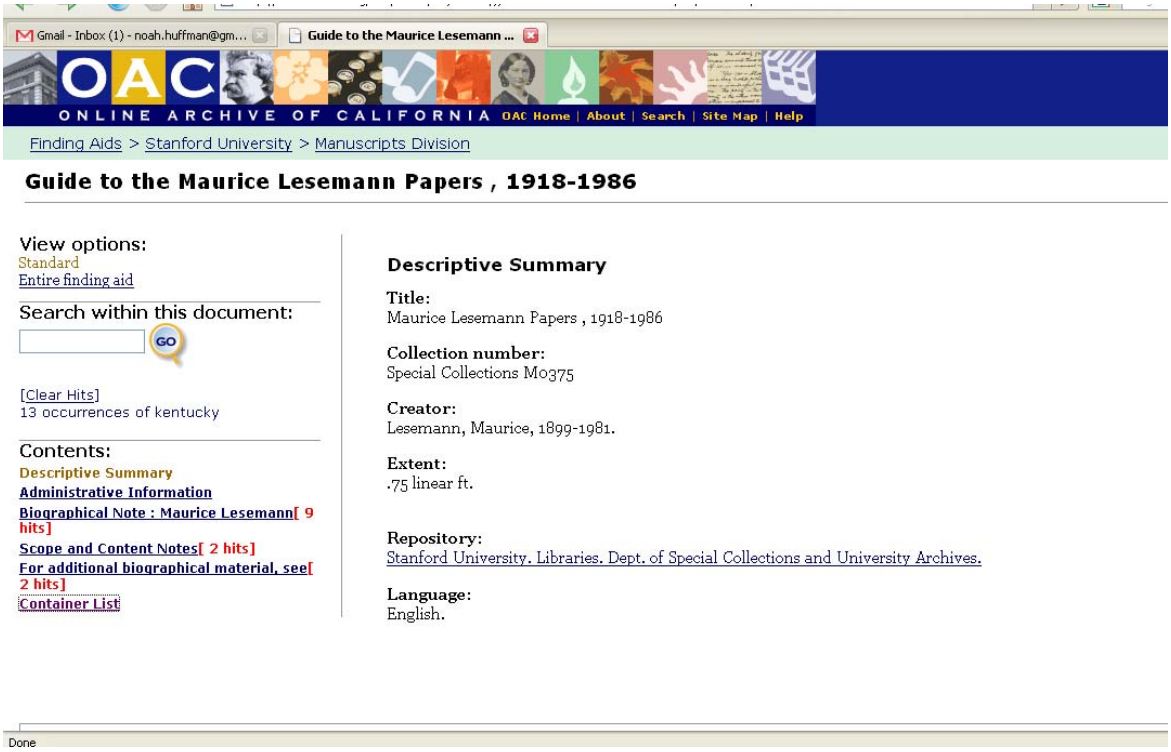

Figure 5: Default Item-Level Display of Finding Aid, Denver Public Library (PLEADE)