Christopher J. Prom
August 29, 2002

****

****

Does EAD Play Well with Other Metadata Standards?: Searching and Retrieving EAD using the OAI Protocols

Abstract: The Open Archives Initiative Protocol for Metadata Harvesting has been suggested as a simple method by which cultural heritage metadata might be exchanged and searched. However, much of the archival description encoded in Encoded Archival Description (EAD) files may not be optimally accessible. By analyzing EAD encoding patterns found in a sample of finding aids, this article explores some of the challenges which arise in attempting to make EAD files interoperable and searchable. The analysis shows that many EAD files lack key metadata elements or use non-standard encoding patterns. The findings suggest that institutions using EAD could benefit from the application of tighter best practice recommendations. In addition, application of the OAI protocols on top of EAD may help harmonize and eliminate encoding differences, providing for better search and retrieval mechanisms.

Does EAD Play well with Other Metadata Standards?:
Searching and Retrieving EAD using the OAI Protocols

"You know you've achieved perfection in design, not when you have nothing more to add, but when you have nothing more to take away."

- Antoine deSaint Exupéry

This paper provides some preliminary conclusions regarding the characteristics of EAD-encoded finding aids. The analysis was conducted to help improve the retrievability of EAD encoded data in a non-EAD environment. The conclusions are preliminary in two senses: First, since the results are based on an analysis of the encoding patterns used in a sample of finding aids; second, in the sense that the research project on which the paper is based has not yet been fully completed.

Background: OAI and Archival Interoperability

This research was completed for the University of Illinois Open Archives Initiative Metadata Harvesting Project.[1] The UIUC OAI project is testing a method by which information about manuscripts, archives, photographs, and other cultural heritage items may be exchanged and searched, regardless of the metadata formats in which such information was originally created and stored. Specifically, the project has been testing the feasability of using the Open Archives Initiative Protocols for Metadata Harvesting (OAI-PMH) for searching descriptions of cultural heritage materials, including those described in EAD files. The project is very much a proof-of-concept project, and previous work has led us to conclude that some difficult, but not insurmountable, barriers need to be overcome before useful OAI records can be produced from EAD.

This research focuses on only one of the issues that will need to be confronted in evaluating the utility of EAD finding aids in an OAI context: Interoperability. Specifically, it examines the question of whether EAD finding aids are consistently enough structured so that they might be searched alongside descriptions of cultural heritage materials that are drawn from other metadata formats, such as TEI, METS, MARC, or relational databases.

As noted elsewhere, it may be useful to be able to search EAD metadata alongside metadata encoded in other formats by using the Open Archives Initiative protocol for Metadata Harvesting (OAI-PMH).[2] The OAI-PMH has been suggested as a relatively simple means by which metadata from various formats might be exchanged in a common format.[3] OAI's purpose has been fully summarized elsewhere,[4] but in general it is important to keep several things in mind regarding OAI. First, OAI is an international initiative centered on increasing the interoperability of digital libraries, broadly construed. OAI originated in the scientific community’s desire to exchange information regarding what has been termed "archived" pre-printed scientific papers. However, the protocol is content neutral and may be applied to any type of metadata, including metadata describing archives, manuscripts, photographs, or other cultural heritage materials. Second, OAI employs a model of "metadata harvesting," (not distributed searching), and it is therefore much less technically complex than previous digital library interoperability standards, such as Z39.50. [5] In general, metadata is exchanged in a simple Dublin Core format, although interest groups are also allowed to exchange metadata using any format that can be described using an XML Schema. (EAD could therefore be exchanged under OAI in its native format, if such a schema were developed to supplement the DTD which is maintained by the EAD Working Group.) Third, OAI views the digital world into terms of data providers and service providers. Again, the distinction is described in detail elsewhere,[6] but in general a data provider is any institution that agrees to provide its metadata in OAI format, using a simple Dublin Core/XML syntax. Service providers are those institutions or individuals who "harvest" this Dublin Core metadata from the data providers and build value added services (such as search portals) that make use of the harvested metadata. Fourth (and finally), OAI does not specify how, when, or under what conditions harvested metadata might be used. It simply specifies a common protocol under which the metadata might be exchanged.

The OAI standard is clearly growing is use and influence, as evidenced by the great deal of interest it occasioned at the recent ACM/IEEE Joint Conference on Digital Libraries; seven papers specifically mentioned OAI in their title and described OAI-based projects; other papers also dealt with the topic in a less formal manner. In particular the National Science Digital Library core architecture is built around OAI, meaning that the protocol will be well supported. The Andrew Mellon Foundation has also funded seven OAI projects. In the realm of cultural heritage materials, The University of Illinois, University of Michigan, Emory, and the University of Virginia are exploring the use of OAI with Mellon grants. The Digital Libraries Federation is also conducting an evaluation of OAI.[7] It appears to have a bright future.

OAI's advocates hope that widespread adoption of the protocol will allow for better digital library interoperability and in particular for better search and retrieval results than that provided by web search engines such as Google. Web search engines generally do not exploit metadata and often miss items found in the deep or hidden web. The emergence of OAI raises interesting questions for archivists using EAD or other descriptive mechanisms. If the content from an EAD file is to be searched alongside content from other EAD files or other metadata records—whether in an OAI or non-OAI environment—it is important that the metadata is be consistently structured. This point is not new,[8] but it takes on new significance given the emergence of OAI.

The EAD team on the Illinois OAI project was charged with developing an effective mapping from EAD to OAI/Dublin Core. As noted above, our general approach has been discussed elsewhere, but in general we decided that to provide the most functionality, we would need to produce many OAI records from one EAD file. Specifically, we produce a "top level" or master record based on the <eadheader> and <archdesc>, which contains metadata describing the entirety of materials in an EAD finding aid. In addition, we believe it will be useful to produce mini-records for each component within the <dsc>. This will allow us to search the entire finding aid while preserving appropriate context for any hits found via the search mechanism.

However, after we began mapping records to OAI, we noticed that many records were not very useful. In fact, many were empty or lacking key metadata elements. This included both the OAI records produced from the top level of the finding aid and those from the <dsc>. This result indicated to us that either the metadata elements were missing from the finding aid or were coded differently than best practices recommend. To gain a better understanding of these problems, we decided to test EAD encoding patterns, in the hope that a full analysis would help us develop a more effective transformation to OAI, particularly for materials found in the description of subordinate components. (To the best knowledge of the author, such as analysis has not been previously conducted.) At the same time, it must be noted that a systematic analysis of EAD encoding patterns will have broader applicability beyond the OAI environment. It may help us understand how EAD is actually being used, potentially shaping the future revisions to the DTD or the development of best practice guides and content standards.

Research Objectives and Methodology

This study's main objective was to evaluate the effects of potential interoperability problems related to EAD records in an OAI environment. Specific objectives and working hypotheses are summarized in Figure 1.

Figure 1: Study Objectives and Hypotheses

Objectives	Working Hypotheses
Main Objective: Evaluate the effects of potential interoperability problems of EAD records in an OAI environment	**Main Hypothesis:** OAI is predicated on a one-to-one representational model, and is intended to deal with relatively small records. Large file sizes, inconsistent encoding practices, and the hierarchical (and redundant) nature of EAD metadata may hamper the interoperability of EAD finding aids in an OAI environment.
O1: Evaluate the existence of metadata (specific elements) in <eadheader> needed for generation of a "top level" OAI record.	**H1:** Lack of key metadata elements (or miscoding of those elements) in some finding aids will impede interoperability
O2: Evaluate the consistency of metadata encoding at the top level of the <archdesc> by examining (1) existence of selected elements in the <did> (2) used of controlled access elements, including use of vocabularies, and (3) use of normalized dates and names	**H2:** Lack of key metadata elements, normalized dates or use of standard vocabularies will negatively affect interoperability and search capabilities.
O3: Evaluate extent to which encoding of <dsc> follows the generally recommended model. 1) only one or two <dsc> elements 2) use of proper elements to encode main metadata such as unittitle and subjects, scopecontent 3) use of level attributes in the <archdesc> and <c0x> elements. (4) existence of <persname> and <title> tags in the <dsc>.	**H3:** Due to major encoding inconsistencies, metadata drawn from the <dsc> will not to be useful in OAI environment.

It should be noted that the analysis in this paper provides only a minimal measure of interoperability requirements, since additional finding aid elements would need to be transformed for display purposes. For example, coding differences among <container> elements could affect display properties significantly. Similarly, the analysis does not touch on broader issues, such as whether the finding aids adhere to external descriptive standards such as APPM or ISAD(G). In other words, the main objective of the analysis was simply to study machine handling problems that might arise, in order to help us develop a more effective mapping from EAD to OAI.

The methodology used to do this is largely based on suggestions previously made by Anne Gilliland-Swetland for testing the scope, structure and consistency of encoding EAD finding aids contributed to the Online Archives of California.[9] But while Gilliland-Swetland's proposed methodology seems to have relied upon a time-consuming manual tabulation process (and no results have been published), our analysis was automated, allowing for quicker results.

To test the hypotheses, a statistical sample was taken from a population finding aids. The sample was evaluated for data integrity, depth, and consistency of encoding according to existing best practice recommendations such as the OAC Best Practice Guide and seven other best practice guides previously analyzed. A complete list of conditions and values tested is shown in the appendix, but a few points about the nature of the analysis need emphasis, since they correspond to the study objectives and hypotheses. First, extensive tests were run to check for the existence of certain elements typically encoded in the <eadheader> that are needed for basic file handling. For example, it is useful to know how consistently or inconsistently the <eadid> element in encoded, since the element is used to identify the finding aid among the entire universe of finding aids. Similarly, we tested for encoding patterns of elements in the <filedesc> such as the <titleproper> and <publicationstmt>. Second, the finding aids were tested for the existence of "top level" metadata in the <archdesc>. In particular, is information about the title and dates of the material provided in the descriptive identification? How are controlled access terms applied? Is a physical description given? Can these data elements be consistently extracted? Third, the elements in the description of subordinate components <dsc> were analyzed for consistency of encoding. How many levels are used in the typical finding aid? Were <unittitle> tags and level attributes extractable? How often are <scopecontent> notes and controlled access terms applied at this level of the finding aid? Finally, It should be noted that in all three of the categories, tests were run to measure the existence of certain "abusive" tagging processes. Did institutions mix genenic <c> tags with the numbered components? Was the <frontmatter> (which is intended purely for formatting) used in place of information in the <eadheader> ? Were an excessive number of <dsc> tags used or were they buried in component levels?

The sample of finding aids was drawn from a population of 8,327 finding aids contributed by 57 institutions to the UIUC OAI project. The calcualted sample size was 367, which ensured a confidence interval of +/- 5%, with a 95% confidence level. The 8,327 finding aids were grouped by institution, and every 22nd file from the population was selected, giving 378 files in the main sample, or an oversample of 11 files. In addition, 22 files from institutions whose files were not selected in the original sample were separately analyzed, as a subsample. (These institutions had contributed less than 22 finding aids to the UIUC project, and it seemed possible that encoding patterns among institutions that had encoded a relatively small number of finding aids might vary from those that had encoded a large number, warranting a separate analysis.)

To simplify the analysis of the files, finding aids in the sample were grouped together, and a XSLT stylesheet was applied to all files to normalize certain insignificant characteristics, such as tag cases or paths to the DTD in the DOCTYPE declaration. After normalization, another XSLT stylesheet was applied to each of the files in the sample. This stylesheet used XSLT's capabilities as a query language to test the files for the conditions noted in the appendix, and the results were exported to a delimited text file. (Some of the XSLT code used is shown in Figure 2.) During post-transformation processing, the text files were joined into a single delimited text file, which was imported to a database to allow for summarization and analysis.

Figure 2: XSLT code used to test existence of <controlaccess> in third component of <dsc>

```
<xsl:choose>
  <xsl:when test="/ead/archdesc/dsc/c01/c02/c03|/ead/archdesc/dsc/c/c/c">
     <xsl:choose>
      <xsl:when  test="/ead/archdesc/dsc/c01/c02/c03//controlaccess[not(ancestor::c04)]|
                    /ead/archdesc/dsc/c/c/c/controlaccess|
                    /ead/archdesc/dsc/c/c/c/*/controlaccess[not(parent::c)]">
            <xsl:text>y</xsl:text>
      </xsl:when>
      <xsl:otherwise>
        <xsl:text>n</xsl:text>
      </xsl:otherwise>
    </xsl:choose>
  </xsl:when>
  <xsl:otherwise>
    <xsl:text>na</xsl:text>
  </xsl:otherwise>
</xsl:choose>
<xsl:text>","</xsl:text>
```

EAD Metadata Characteristics

Population and Sample Characteristics

The population included files ranging in size from 1,547 to 13,728,394 bytes, and all files were EAD (not EADGRP) files. Five thousand five hundred twenty-eight (66%) of the 8,327 files in the population were contributed by institutions participating in the Online Archives of California (OAC). Since OAC has in the past used retrospective conversion and best practice guidelines, it is possible that the OAC files may display a higher level of encoding uniformity than other files. On the other hand, several of the OAC institutions contributed a very small number of files, or, in a few cases, only one file. This may lead to less encoding consistency than might be expected among institutions with a more established encoding workflow. Second, 2,100 of the remaining files (25%) were contributed by institutions in the US Midwest (The University of Minnesota, University of Michigan, University of Chicago, University of Illinois, University of Iowa, Minnesota Historical Society, and Michigan State University). This is not surprising since our Open Archives Initiative project focuses on CIC institutions. The remaining records (9%) were contributed by the Texas Archival Resources On-line project, the Harvard University Libraries, and Cornell University Library. In general, the EAD files in the population were produced by some of the leading institutions and projects. Nevertheless, the population represents a wide range of institutions and tagging practices, at least in the United States. No EAD records from outside the US were included in the population.

The smallest file in the sample of 378 records examined was 1,996 bytes, and the largest was 4,66,667 bytes, although one file in the supplemental sample of twenty two finding aids exceeded 8MB in size. The average file size was 71,379 bytes, indicating, as expected, that most EAD files are moderately large. Of the 378 files, 363 (96%) were apparently used to describe a "collection" of material, since the mandatory <archdesc> level attribute was set to collection.[10]

Encoding Characteristics

The raw results of the analysis are shown in the appendix. In general they confirm the main hypothesis that some EAD files exhibit inconsistent encoding practices and lack key metadata elements, although many files are structured in accordance with current best practice guidelines. This can be seen by a more detailed explanation of the findings regarding the three specific hypotheses listed earlier.

Objective/Hypothesis One: Elements in <eadheader>

The lack of key elements or attributes in the <eadheader> will affect the machine handling of many EAD files, since these are key fields that would be needed for manipulation in an OAI system. This is most immediately apparent in the use of the <eadid> element, which is a required subelement of the <eadheader>. The element includes source, systemid, and type attributes which are used to provide information about the naming schema employed, but many insitutions are not using these elements. While 335 of 378 (or 87.3%) of the finding aids included the type element, only 71 (18.8%) specified a source; similarly only 72 (19%) specified a systemID. Forty-three (11.4%) did not specify any of the attributes, meaning the information encoded therein would be impossible to use as a formal identifier. This included finding aids from five institutions.

Of those finding aids that did include a type attribute, 263 (69.6%) used an SGML catalog entry. The high percentage is a reflection of the fact that this is required for inclusion in the OAC; relatively few non-OAC institutions used it. Most of the others used a simple file name (16.1%) or URL (2.9%). For OAI purposes, it would be easiest to simply identify all EAD files placed on public web servers by using the URL/URI, and ignoring the <eadid> tag, since its use is so variable.

The encoding patterns for other elements in the <eadheader> indicate that they may be more useful for OAI, but key metadata was missing for a small percentage of finding aid. For example, 47 institutions (12.4%) did not list a finding aid publisher; 10 (2.6%) chose not to list (or did not properly encode) the name of the repository where the materials are located. Even more significantly, 95 (25.1%) of the finding aids did not include a publication date for the finding aid. Although the publication date could possibly be inferred from the date in the <profiledesc>, doing so would necessitate a loss of specificity, and it should also be noted that 28 institutions (7.4%) did not include a <profiledesc> date; eleven (2.9%) did not provide either a publication or a profile date.

Use of the most essential element, <titleproper>, was much more consistent; only 1 (.3%) of the institutions did not provide data in the <titleproper> field. The element itself is required by the DTD. Upon close examination it was noted that this finding aid contained an apparent encoding error. The <titleproper> element did exist as required, but was empty even though the institution had provided a subtitle. (This example, as well as others found during the analysis may provide some rationale for converting the DTD to an XML schema, which would allow for data typing.) Finally, it should be noted that 302 institutions (79.9%) included a titlepage in the frontmatter. Although the titlepage is meant to allow institutions to repeat selected information from the <eadheader> it is possible that some new information may be introduced. Given the increasing use of XML, the existence of frontmatter in the DTD is not strictly necessary since XSLT allows programmers to reorder text from anywhere in an finding aid. Redundancy in the <frontmatter> complicates computer processing of the records.

Objective/Hypothesis Two: Consistency of Metadata at Top Level of Archdesc

In an OAI environment, consistency of markup in the top levels of the <archdesc> is important because data from the top level is used to produce an OAI record describing the entirety of materials described in a finding aid. It has often been noted that archival description is collective by nature, and the consistency with which that collective description is applied will increase interoperability. Analysis of EAD metadata in the top level of the <archdesc> shows that such consistency is lacking for a substantial minority of finding aids.

In most cases, essential metadata does exist and is able to be extracted. For example, all 371 (98.1%) of the finding aids examined included a <unittitle> element in the descriptive identification <did>. All but 34 (9.0%) included a <unitdate>[11], all but 38 (10.1%) a <creator> element, and all but 20 (5.3%) a <physdesc> element. Although these elements are not formally required by the DTD, their widespread usage indicates that they can be extracted from most finding aids for the automatic creation of OAI records. However, some seemingly essential elements are missing from many finding aids. Two hundred thirty-one finding aids (61.2%) did not include an abstract, and 121 (32%) did not include a <scopecontent> element. Although the function of these elements overlaps somewhat, either may be used to provide a brief free text description of the collection; 61 (16.1%) of the finding aids did not include either field, impeding their interoperability. Similarly, 165 (43.7%) of the finding aids in the sample did not use any controlled access subject or name terms at the top level of the finding aid.

Although these results suggest the metadata encoded EAD sometimes lacks key metadata elements, it should be noted that the minority of finding aids that are encoded to high standards would be able to contribute high quality, searchable metadata to an OAI harvester. However, this number appears to be a very small minority. Only seventy-two (19%) of the finding aids specified usage of a controlled vocabulary for subjects by listing the vocabulary in the source attribute for the <controlaccess> element; all but one of these used Library of Congress subject headings (LCSH); the other used locally defined subjects. These findings would seem to indicate that even if content standards are being applied in an EAD environment, they are often coded incorrectly and thus will be difficult, if not impossible, to be effectively processed by computers.

In most finding aids, the top level of the <archdesc> contains adequate metadata to produce a minimally useful OAI record, but many finding aids could probably benefit from the application of more rigorous encoding practices. Under present circumstances, it may be advisable for OAI service providers to draw OAI records describing a collection as a whole from MARC-AMC records, if such records have been produced. Although such a strategy runs the risk of producing duplicate or near duplicate OAI records in a service provider's database, such duplicates could possibly be filtered out by the service provider.

Objective/Hypothesis Three: Consistency of Encoding the <dsc>

The description of subordinate components <dsc> is the heart of most finding aids, and corresponds most closely to the traditional archival idea of the register, inventory, or box listing. Three hundred and forty-two (91.5%) of the finding aids in the main sample included a <dsc>, as did twenty-one (95.5%) of the finding aids in the subsample. It describes the components of the collection, series, or item described in the entire EAD document, and the ability to search for and retrieve items encoded in the <dsc> may be very important to users, especially since an increasing number of items described in the finding aid may be available via digital surrogates. Fifty three of the finding aids in the main sample (14%) included pointers to digital archival objects.

Accurate, consistently structured description in the <dsc> is important, because poorly structured markup is difficult to extract. Liz Shaw has cited the infamous example of the <unitdate> tag, which may legally appear at multiple and unpredictable nodes in the <dsc>.[12] However, it is uncertain how the availability of flexible markup options is actually reflected in actual markup. No analysis of files has been previously done, and without such an analysis, we have no way of knowing what interoperability problems need to be overcome.

Clearly, there is a lot of data that could potentially be made use of. Statistics drawn from the main sample indicate that the average finding aid contains 19.9 <c01> elements, 92.4 <c02> elements, and 110.4 <c03> elements; altogether, the 378 finding aids contain 84,017 levels of description through only the 3rd level (deeper levels were not measured.) Most sampled institutions (283 of 342 finding aids; 82.7%) appear to be encoding all the subordinate components in one <dsc>, which accords with current best practice recommendations of the OAC and other bodies, but fifty-seven institutions (16.7%) use two <dsc>s, and two use three or more.[13]

How is this data structured? First, it should be noted that very few institutions are specifying the level of materials being described by any given component level. EAD's component elements include a level attribute which can be used to specify the level at which materials are being described (such a series, item, file, record group, collection). Only 181 of the 336 (53.9%) institutions that used <c01> or the equivalent generic <c> used the level attribute, and usage declined at component levels two and three, where the level attribute was used by 54 of 222 (24.3%) and 28 of 125 (22.8%) of finding aids respectively. Furthermore, the application of these attributes appears to have been handled in a relatively pro forma fashion. It was noted earlier that the vast majority of institutions set the <archdesc> level to collection; it might similarly be noted that 68% of institutions using a level attribute at <c01> or its generic equivalent set the attribute to "series" and that the majority of the rest (22.7%) were set to item, reflecting the specific guidelines of one project. Similarly, 78.6% of those using it at <c03> set the level to file.

If it is difficult to tell what level of materials are being described in the component levels, it is also difficult in many cases to extract useful metadata. Using a very liberal XPATH syntax, I attempted to find any <unittitle> tags within the first component level encountered within the finding aid.[14] For the <c01> or equivalent <c>, the search succeeded only 60.1% of the time. Similar results were found for the second and third component levels. In many cases the search failed because information is encoded in a tabular display format, using the <drow> and <dentry> tags—151 (44.2%) of the 342 finding aids with a <dsc> included such tabular formatting, which is deprecated in most best practice guides. In other cases, institutions appear to have encoded information in other tags or in the component tag without further differentiation. Whatever the case, the lack of consistency between finding aids makes it difficult to develop a general purpose mapping out of the <dsc>. By implication, it would also be very difficult, if not impossible, to construct a search tool that accurately includes all metadata encoded by different institutions in the common EAD format.

The usage of more specific elements in the component levels appears to be relatively rare. For example, only 31% of finding aids included <scopecontent> element in any of the <c01> elements; 31% used a <title> element; 8.2% used the <persname> element; only 2.1% included any <controlaccess> elements; and none included any <creator> elements. (These figures reflect usage *at any point within the <dsc>*.) Although including these elements may be useful for display purposes or possibly for searching in a local context, the fact they are used relatively infrequently makes them of marginal utility for search and retrieval, especially since the analysis of the sample found no evidence that any authority controls have been applied. Free text searching of the entire component (excluding any component children) would be just as effective.

When examining data across a body of finding aids, it is clear that most data is very loosely structured within the subordinate components. It is unlikely that the markup applied here will do much to increase search and retrieval results, however useful the markup may be in specifying and structuring outputs within the context of an individual institution or project. These results might be taken to imply that metadata drawn from the <dsc> will not to be useful in OAI environment. However, better mappings can be developed so that, at a minimum, metadata for a unititle can be extracted from each level. Although metadata in an OAI environment may therefore seem less rich than EAD metadata in its native format, the "dumbing down" process of creating an OAI record may actually harmonize the data by smoothing over encoding differences.

Implications

The results reported in this paper suggest that much of the information encoded in EAD metadata is not optimally structured to allow for reprocessing and searching in an OAI environment. By extension, it is apparent that encoding inconsistencies would also affect search and retrieval results in a native EAD environment. In addition, normalized dates or use of standard vocabularies are applied infrequently.

Application of stricter encoding schema (or even a more restrictive DTD) could make EAD records much more useful for searching, both in their native context, or converted to OAI. Similarly, the findings here seem to indicate that many of the element options in the DTD add unnecessary complexity to encoding, resulting in inconsistent encoding patterns; in fact many of the options are rarely used. (For example, only 5 of 378 institutions include a <bioghist> element in the <dsc>.) Similarly, the inclusion of procedural markup elements in the DTD (such as <table>, <drow>, <dentry>, and <div>) severely limits the interoperabilty of EAD files which is essentially a descriptive markup protocol. Although some of these options are eliminated in the 2002 version of the EAD Document Type Definition, these findings cast some doubt on the long-term utility of EAD as a data structure standard, at the very time when data content standards are being harmonized.[15]

Advocates of Encoded Archival Description have often argued that one reason EAD should be adopted is because it will provide archives and libraries with the ability to exchange and search information in a common format regarding their holdings. The reasoning is similar to the arguments for adopting MARC-AMC. Specifically, the use of EAD is encouraged as a potential means to aggregate archival description, including collection-level description and inventories or registers, into union databases. It is argued that EAD can possibly provide enhanced cross-collection and possibly cross-repository searching. Many institutions have begun to aggregate their finding aids based at least partly on this hope; prominent projects include the Online Archive of California, Texas Archival Resources Online, the Five Colleges Finding Aid project, and, overseas, the massive British Access2Archives project, which includes an astounding 50,000 EAD Records stored in an XML database.[16]

Yet at the same time, it is uncertain whether the long-term benefits in interoperability will actually be worth the time, money, and aggravation that have been expended doing this encoding. Questions have been raised from a variety of perspectives. For example, Helen Tibbo has noted that "Archivists do not have a great track record in the area of user studies. . . . To date, there is very little evidence regarding the use and efficacy of any electronic access tools in the archival domain."[17] Elizabeth Shaw points out that from the point of view of a computer programmer, EAD's permissive data model may undermine the very interoperability it is intended to foster, an opinion that would seem to be validated by the research conducted for this paper.[18] From a more theoretical perspective, many computer scientists and information theorists are questioning the utility of complex metadata as librarians and archivists currently conceive it. (As Elizabeth Liddy noted during a panel session devoted to metadata at the recent Joint Conference on Digital Libraries, techniques of natural language processing may provide users as much access to information as the rigorous encoding of information in a structured format.[19]) Kris Keisling has also implied that EAD metadata is not optimally structured to be searched and found using current methods employed by web search bots to index websites.[20]

It is not an overstatement to say that we do not actually know how useful EAD metadata actually is for complex search and retrieval. However, the results of this research cannot be taken to argue, as earlier hypothesized, that EAD files are too lengthy and complex to be used in an OAI environment. In this respect, the Open Archives Initiative Protocols may actually help improve access. As Michael Nelson has pointed out, you do not solve interoperability problems by ratcheting up the level of complexity.[21] Here, perhaps, lies OAI's great advantage as a possible search and retrieval mechanism for EAD metadata. If used as a "front end" to EAD metadata, OAI records could potentially mitigate the encoding differences found between finding aids and institutions. As demonstrated in the beta test of the UIUC OAI search interface, it is possible to search EAD at a very deep level in an OAI context using harvested metadata.[22] While this paper shows that much metadata is lost in conversion to OAI using the general purpose mapping similar to that used in the initial stages of the UIUC OAI project,[23] it should be possible to develop a more effective mapping that will smooth over encoding differences. The data assembled for this study should make that a more feasible task, although the specific manner in which that may be accomplished is a subject for further work—and the topic of another paper.

Appendix: Tested Values and Raw Results

This table shows the database fields created from 378 sample OAI records and the 22 additional records tested from institutions not included in the original sample.

Field	Purpose	Main Sample Results	Subsample Results
FileName	Identification number		
EADIDTypeExists	Does eadid have a type attribute?	Yes: 335 (87.3%) No: 48 (12.7%)	Yes: 22 (100%)
EADIDSourceExists	Does eadid have a source attribute?	Yes: 71 (18.8%) No: 307 (81.2%)	No: 22 (100%)
EADIDSystemIDExists	Does eadid hae a systemid attribute?	Yes: 72(19%) No: 306 (81%)	No: 22 (100%)
EADIDTypeValue	Value of eadid/@type	SGMLCatalog: 263 (69.6%) File: 61 (16.1%) url/uri: 11 (2.9%) NA (no type attribute): 48 (12.7%)	SGMLCatalog: 22 (100%)
EADIDSourceValue	Value of eadid/@source	Not calculated	NA
EADIDSystemIDValue	Value of /ead/@systemid	Not calculated	NA
TopPublisher	Does the finding aid list a publisher?	Yes: 331 (87.6%) No: 47 (12.4%)	Yes: 21 (95.5%) No: 1 (4.5%)
TopRepository	Does the finding aid list a repository?	Yes: 368 (97.4%) No: 10 (2.6%)	Yes: 22 (100%)
PubStmtDate	Is there a publication date in eadheader?	Yes: 283 (74.9%) No: 95 (25.1%)	Yes: 22 (100%)
PubStmtDateNormal	Is publication date normalized?	No: 283 of 283 possible (100%)	No: 22 of 22 possible (100%)
ProfileDescDate	Is there a creation date for EAD file?	Yes: 350 (92.6%) No: 28 (7.4%)	Yes: 21 (95.5%) No: 1 (4.5%)
ProfileDescDateNormal	Is creation date it normalized?	No: 350 of 350 (100%)	21 of 21 (100%)
TopTitleProper	Does finding aid have a titleproper in filedesc?	Yes: 377 (99.7%) No: 1(.3%)	Yes: 22 (100%)
TopSubTitle	Does titleproper have a subtitle?	Yes: 71(18.8%) No: 307 (81.2%)	No: 22 (100%)
TopTitlePage	Does finding aid include frontmatter/titlepage?	Yes: 302 (79.9%) No: 76 (20.1%)	No: 22 (100%)
TopUnittitle	Does <archdesc> have a unittitle?	Yes: 371 (98.1%) No: 7(1.9%)	22 (100%)
TopDate	Does archdesc have a unitdate?	Yes: 344 (91%) No: 34 (9.0%)	Yes: 12 (55%) No: 10 (45%)
TopDateNormal	Does is unitdate normalized?	No: 344 of 344 (100%)	No: 12 of 12 (100%)
TopCreator	Does /ead/archdesc/did/origination exist?	Yes: 340 (89.9%) No: 38 (10.1%)	Yes: 20 (90.9%) No: 2 (9.1%)
TopCreatorConrtrol	/ead/archdesc/did/origination/persname[@source]|/ead/archdesc/did/origination/corpname[@source] exist?	Yes: 60 of 340 (17.6%) No: 280 of 340 (82.4%)	No: 20 of 20 (100%)

Field	Purpose	Main Sample Results	Subsample Results
ArchdescLevel	Values of /ead/archdesc/@level	Collection: 263 (96.0%) Record Group: 8 (2.1%) Series: 5 (1.3%) Item 1(.3%) Otherlevel: 1 (.3%)	
TopAbs	Does archdesc have an abstract?	Yes: 147 (39.8%) No: 231 (61.2%)	Yes: 1(4.5%) No: 21 (96.5%)
TopScope	Does archdesc have a scopcontent?	Yes 257 (68%) No: 121 (32%) NOTE: 61 finding aids (16.1%) did not include either an abstract or scopecontent note.	Yes: 17 (77.3%) No: 5 (26.7%)
TopAbsLength	Length of abstract in characters.	Not calcuated	Not calculated
TopScopeLengh	Length of Scopecontent note in characters.	Not Calculated	Not caluclated
TopPhysdesc	Does a physical description exist?	Yes: 358 (94.7%) No: 20 (5.3%)	Yes 15 (68.2) No: 7 (31.8%)
TopCtrlAccNumber	Number of control access terms used	None: 165 (43.7%) Avg: 6.6 Max: 199	None: 19 (86.4%) Avg: .6 Max: 9
TopNameSource	Are personal names authority controlled?	Yes: 62 of 213 (29.1%; 16.4% of whole) No: 151 of 213 (70.9; 83.6% of whole)	No: 0 of 3 (0%)
TopNameSourceValue	Which authority is used?	LCNAF: 58 of 213 (27.2%) AACR2: 2 of 213 (.9%) Othersource: 2 of 213 (.9%)	na
TopSubjectSource	Are subject authority controlled?	72 of 213 (33.8%; 19%)	2 of 3 (66.6%)
TopSubjectSourceValue	Which authority control system is used?	LCSH: 71 of 72 (98.6%; 18.8% of whole) Othersource: 1 of 72 (1.4%; .3% of whole)	2 of 3 (66.6%; 9.1% of whole)
TopLang	Does archdesc have a language attribute?	Yes: 57 (15.1%) No: 321 (84.9%)	Yes: 22 (100%)
TopLangValue	Which language is listed?	Eng: 57 of 57 (100%)	Yes: 22 of 22 (100%)
DscExist	Does finding aid have <dsc>?	Yes: 342 (91.5%) No: 36 (9.5%)	Yes: 21 (95.5%) No: 1 (4.5%)
DscNumber	How many subordinate components are there?	zero: 36 (9.5%) one: 283 (74.8%) two: 57 (15.1%) three: 1(.3%) four: 1 (.3%)	zero: 1 (4.5%) one: 14 (63.7%) two: 6 (27.3%) four: 1 (4.5%)
DscType1	Value of type attribute for first dsc element	indepth: 221 of 342 (64.6%) combined: 66 of 342 (19.3%) analyticover: 54 of 342 (15.9%) othertype: 1 of 342 (.3%)	in-depth: 10 of 21 (47.6%) analyticover: 8 of 21 (38.1%) combined: 3of 21 (14.3%)

Field	Purpose	Main Sample Results	Subsample Results
DscType2	Value of type attribute of second dsc element	indepth: 46 of 59 (78%) othertype: 12 of 59 (20.3.%) Analyticover: 1 of 59 (1.7%)	indepth: 6 of 7 (85.7%) analyticover:1 of 7 (14.3%)
C01Number	Number of c01 or equivalent c elements	Total: 7,510 Avg: 19.9 Min: 0 Max 594	Total: 1,320 Avg: 60 Min: 0 Max 576
C02Number	Number of c02 or equivalent c elements	Total: 34,911 Avg: 92.4 Min: 0 Max: 4,394	Total: 4,980 Avg: 226.4 Min: 0 Max: 1,787
C03Number	Number of c03 or equivalent c elements	Total: 41,596 Avg: 110.4 Min: 0 Max:17,017	Total: 1771 Avg: 80.5 Min: 0 Max: 597
TabFormatting	Does finding aid use <drow> formatting?	Yes: 151 of 342 (44.2%) No: 191 of 342 (55.8%)	Yes: 14 of 21 (66.7%) No: 7 of 21 (33.3%)
The following were tested from first example found in finding aid			
C01Level	Does c01 or equivalent c have level attribute?	Yes: 181 of 336 (53.9%) No: 155 of 333 (46.1%)	Yes: 16 of 20 (80%) No: 4 of 20 (20%)
C02Level	Does c02 or equivalent c have level attribute?	Yes: 54 of 222 (24.3%) No: 168 of 222 (75.7%)	Yes: 7 of 19 (36.8%) No: 12 of 19 (63.2%)
C03Level	Does c03 or equivalent c have level attribute?	Yes: 28 of 125 (22.4%) No: 97 of 125 (77.6%)	Yes: 4 of 8(50%) No: 4 of 8 (50%)
C01LevelValue	Value of c01 level	series: 123 of 181(68%) item: 41 of 181(22.7%) otherlevel: 10 of 181 (5.5%) subgrp: 3 of 181(1.7%) file: 3 of 181 (1.7%) recordgrp: 1 of 181 (.6%)	Not Calculated
C02LevelValue	Value of c02 level	file: 23of 54 (42.6%) subseries: 21 (38.9%) item: 5 (9.3%) series: 2 (3.7%) otherlevel: 2 (3.7%) recordgrp: 1 (1.9%)	Not calculated
C03LevelValue	Value of c03 level	file: 22 of 28 (78.6%) item: 4 (14.3%) series: 1 (3.6%) subseries: 1 (3.6%)	Not calculated
C01Unititle	Is <unittitle> used at first component level?	Yes: 202 of 336 (60.1%) No: 134 (39.9%)	Yes: 16 of 20 (80%) No: 4 of 20 (20%)
C02Unititle	Is <unittitle> used at second component level?	Yes: 126 of 222 (58.8%) No: 96 (42.2%)	Yes 5 of 19 (26.3%) No: 14 (73.7%)
C03Unititle	Is <unittitle> used at third component level?	Yes: 72 of 125 (57.6%) No: 53 (42.4%)	Yes: 4 of 8 (50%) No: 4(50%)
The following were tested for any point in the finding aid:			
C01Scope	Does c01/scopecontent exist?	Yes: 104 of 336 (31%) No: 232 (69%)	Yes: 9 of 20 (45%) No: 11 of 20 (55%)

Field	Purpose	Main Sample Results	Subsample Results
C02Scope	Does c02/scopecontent exist?	Yes: 47 of 222 (21.2%) No: 175 (78.8%)	Yes: 3 of 19 (15.8%) No: 16 of 19 (84.2%)
C03Scope	Does c03/scopecontent exist?	Yes: 26 of 125 (20.8%) No: 99 of 125 (19.2%)	No: 8 of 8 (100%)
C01Controlaccess	Does c01/controlaccess exist?	Yes: 7 of 336 (2.1%) No: 329 (98.1%)	No: 20 of 20 (100%)
C02Controlaccess	Does c02/controlaccess exist?	No: 222 of 222 (100%)	Yes: 3 of 19 (15.8%) No: 16 of 19 (84.2%)
C03Controlaccess	Does c03/controlaccess exist?	Yes: 1 of 125 (.8%) No: 124 of 125 (99.2%)	No: 8 of 8 (100%)
ComponentDsc	Do component levels contain a nested <dsc>?	Yes: 3 (.8%) No: 375 (99.2%)	No: 22 (100%)
Component tables	Do component levels use <table> element for formatting?	Yes: 1 (.3%) No: 377 (99.7%)	No: 22 (100%)
C01AndC	Are both component styles used together?	No: 378 (100%)	No: 22 (100%)
Dao	Is dao element used?	Yes: 4(1.1%) No: 374 (98.9%)	No: 22 (100%)
DaoGrp	Is daogrp element used?	Yes: 49(12.9%) (42 at Michigan State, 7 at Calher) No: 329 (87.1%)	No: 22 (100%)
DaoDesc	Is daodesc element used?	Yes: 43 (11.4%) No: 335 (88.6%)	Yes: 1 (4.5%) No: 21 (95.5%)
DaoLoc	Is daoloc element used?	Yes: 49(12.9%) (42 at Michigan State, 7 at Calher) No: 329 (87.1%)	Yes: 3 (13.6%) No: 19 (86.4%)
DaoCount	Number of dao elements used	Total: 548 Min: 0 Max: 540	NA
DaoLocCount	Number of daoloc elements used	Total: 1647 Min:0 Max: 687	Total: 15,611 Min: 0 Max: 12,714
FindAidTag	Is the deprecated <findaid> element used?	Yes: 7 (1.9%) No: 371 (98.1%)	No: 22 (100%)
TitleExists	Is <title> tag used in the <dsc>?	Yes: 117 (31%) No: 261 (69%)	Not calculated
PersnameExists	Is <persname> tag used in the <dsc>?	Yes: 31(8.2%) No: 347 (91.8%)	Not calculated
BioghistExists	Is <bioghist> tag used in the <dsc>?	Yes: 5 (1.3%) No: 373 (98.7%)	Not calculated
CreatorExists	Is <creator> tag used in the <dsc>?	Yes: 0 No: 378 (100%)	Not calculated

Endnotes

[1] http://oai.grainger.uiuc.edu/. The author thanks the Andrew Mellon Foundation for supporting this project and his research.

[2] Christopher J. Prom and Thomas G. Habing, "Using the Open Archives Initiative Protocols with EAD," *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, 176-77.

[3] http://www.openarchives.org/ Accessed July 22, 2002.

[4] Lagoze, Carl, and Herbert Van de Sompel, The Open Archives Initiative: Building a Low-Barrier Interoperability Framework, *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries* (2001): 54-62. Available at http://doi.acm.org/10.1145/379437.379449;

[5] Michael L. Nelson, "Better Interoperability through the Open Archives Initiative," *The New Review of Information Networking* 7 (2001): 133.

[6] Nelson, "Better Interoperability," 134.

[7] http://www.diglib.org/architectures/testbed.htm

[8] For example, the American Heritage Virtual Archive Project Retrospective Conversion Guidelines, which were last updated in February 1999, note that "Given the flexibility of EAD, the choices with respect to type, sequence, and quantity of information, as well as varying levels of detail of encoding, it does not in and of itself ensure that machine-readable finding aids will be easily communicated between repositories, nor facilitate the building of union database. Finding aids in union databases will need to share a degree of uniformity, both to make them easily intelligible for users as they navigate from a finding aid from one institution to that of another, and to make them manageable in a computer environment. This uniformity applies to both the intellectual content, and the machine-readable representation or encoding of that content. Predictability and stability are essential for the existence of communities." http://sunsite.berkeley.edu/amher/upguide.html. Accessed July 23, 2002.

[9] Anne J. Gilliland-Swetland, "Evaluation Design for Large-Scale, Collaborative Online Archives: Interim Report of the Online Archive of California Evaluation Project," *Archives and Museum Informatics* 12:3-4 (1998): 192-96.

[10] Of the remaining 15 files, 8 (2.1 %) described record groups, 5 (1.3%) described series, and one each (0.3%) described an item or "otherlevel".

[11] Somewhat anomalously, 10 of the 22 finding aids in the subsample (45%) did not include a unitdate. Otherwise, results in the subsample paralleled those in the main sample.

[12] Elizabeth J. Shaw, "Rethinking EAD: Balancing Flexibility and Interoperability," *New Review of Information Networking*,7 (2001): 122.

13. Three finding aids (.8%) of the 378 sampled included a <dsc> as the child of a component level. Although technically legal according to the DTD, such a practice would severely affect interoperability.

14. Essentially, the syntax searched for any <unittitle> elements describing the current level but not any child components or in <drow> elements, which require separate handling.

15. Statement of Principles for the Custard Project, http://www.archivists.org/news/custardproject.asp. Accessed August 9, 2002.

16. Meg Sweet, message to EAD Listserv, July 18, 2002. http://listserv.loc.gov/cgi-bin/wa?A2=ind0207&L=ead&F=&S=&P=2699

17. Helen R, Tibbo, "Primarliy History: Historians and the Search for Primary Source Materials," *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries* (2002): 1-10.

18. Elizabeth J. Shaw, "Rethinking EAD: Balancing Flexibility and Interoperability," *New Review of Information Networking*,7 (2001): 117-131.

19.See http://www.ohsu.edu/jcdl/main.cgi?opt=sked-pan for a description of the panel. Accessed July 22, 2002.

20. Kristi Keisling, "Metadata, metadata, everywhere - but where is the hook?," *OCLC Systems and Services* Vol 17:2 (2001): 84-88.

21. Michael L. Nelson, "Better Interoperability through the Open Archives Initiative," *The New Review of Information Networking* 7 (2001): 136.

22. http://oai.grainger.uiuc.edu/oai/search/

23. Christopher J. Prom and Thomas G. Habing, "Using the Open Archives Initiative Protocols with EAD," *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, 178-79.