

Using the Open Archives Initiative Protocols with EAD

Christopher J. Prom
University of Illinois Archives
19 Library, 1408 W. Gregory Dr.
Urbana, IL, 61801
+1 217 333 0798
prom@uiuc.edu

Thomas G. Habing
University of Illinois Engineering Library
1301 W. Springfield Ave.
Urbana, IL, 61801
+1 217 244 4425
thabing@uiuc.edu

ABSTRACT

The Open Archives Initiative Protocols present a promising opportunity to make metadata about archives, manuscript collections, and cultural heritage resources easier to locate and search. However, several technical barriers must be overcome before useful OAI records can be produced from the disparate metadata formats used to describe these resources. This paper examines Encoded Archival Description (EAD) as a test case of the issues to be addressed in transforming cultural heritage metadata to OAI. While EAD and OAI may appear to be incompatible, a mapping would be both useful and technically feasible. The authors suggest that it will be necessary to create numerous OAI records from one EAD file. In addition, the findings indicate that further standardization of EAD markup practices would enhance interoperability.

Categories and Subject Descriptors

H.3.7 [Information Storage and Retrieval]: Digital Libraries – Collection, Dissemination, Standards, Systems issues, User issues.

General Terms

Design, Experimentation, Standardization.

Keywords

Open Archives Initiative, Encoded Archival Description, Interoperability.

1. INTRODUCTION

Over the past 10 years, libraries, museums, and archives have undertaken numerous digital library projects to provide access to the unique photographs, manuscripts, and archives held in their custody. Although these projects vary in their exact goals, audiences, and technical specifications, all seek to make previously difficult-to-access cultural heritage materials available through the Internet. The most prominent projects (such as American Memory, the Online Archive of California, and Making

of America) are well known to librarians, but innumerable others have also been undertaken.

One issue which must be confronted when examining these projects is the question of how researchers will actually locate materials of scholarly or personal interest. Planners of digital libraries have used a wide variety of system designs and metadata formats to organize, describe, and provide electronic access to their holdings. These resources are not optimally accessible to researchers who have no prior knowledge of their existence. In many cases, the user must find digital library resources by following a footnote in a book or by using an institutional website, an online public access catalog, or a locally-developed database. In addition, information is typically transformed into HTML before being provided through a browser, sacrificing metadata that exists in the native format.

As a result, many items in digital libraries cannot be easily found using traditional web search engines [15]. They are part of the so-called "deep web" [3]. In addition, researchers are not using electronic resources as effectively as they might. A recent analysis of research methods used by humanities scholars found confusion and regret over the lack of uniformity in the operation of electronic resources and inefficient use of archival finding aids which could potentially facilitate the research process [5]. For this reason, librarians and archivists should reconceptualize how information about special collections and cultural heritage resources is presented to users.

The Open Archives Initiative (OAI) protocols present one possible way to make hidden cultural heritage resources easier to find and use. OAI arose as a promising attempt to develop a relatively "low barrier" interoperability framework for sharing metadata about Pre-Prints (on line repositories of articles not yet accepted for publication) from a variety of domains and content providers. But OAI's advocates soon realized that the framework could be applied to other data types. The research library community in particular has expressed an interest in using OAI as a common format to expose metadata for use by harvesters and search engines. As the developers of OAI are well aware, the feasibility of this hope needs to be demonstrated, not assumed [13]. Accordingly, the Andrew W. Mellon Foundation has undertaken a Metadata Harvesting Initiative, funding seven projects that test the application of harvesting and search technologies using OAI protocols [18].

As one of the seven Mellon grantees, the University of Illinois at Urbana-Champaign is testing the feasibility of harvesting, integrating, and searching metadata about cultural heritage resources and special collections which are described using standards such as Encoded Archival Description, the Text Encoding Initiative (TEI), Machine Readable Cataloging

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL'02, July 13-17, 2002, Portland, Oregon, USA.

Copyright 2002 ACM 1-58113-513-0/02/0007...\$5.00.

(MARC), and locally developed web-accessible databases [17]. While many technical (not to mention political and financial) barriers need to be overcome before the OAI protocols can effectively index cultural heritage resources, one of the most pressing technical issues confronted in the Illinois OAI project is also one of the most difficult: how to develop usable OAI records from disparate metadata sources.

This paper concentrates solely on issues to be addressed when converting Encoded Archival Description (EAD) records into OAI records. EAD is a good test case because it has proven to be one of the most successful of the SGML/XML document type definitions used by the archival, museum, and library community. It is among the most widely used metadata formats employed by cultural heritage and archives projects, and it is used to encode metadata about a wide variety of resources, from large manuscript collections to individual photos, letters, and artifacts. But perhaps more importantly, EAD uses a very flexible metadata structure. It allows for a wide range of tagging practices and encoding schema and presents challenges which test the applicability of OAI to disparate resources. This paper explains the structural problems associated with converting EAD to OAI, analyzes EAD best practice guidelines (as well as a sample of finding aids donated to the Illinois OAI project), and provides a recommended procedure for creating OAI metadata records from EAD files.

Archivists reading this paper may wonder why we should bother to convert EAD to OAI records. Are not the two metadata formats intended for very different purposes? Don't they use the word "archives" in very different ways? There is something to these concerns and a few caveats are in order. First, we do not propose that the full capabilities of EAD finding aids could be subsumed by OAI. Searching the EAD finding aids directly may yield richer results, at least if search mechanisms for EAD ever become more robust than those currently available. Second, EAD metadata can be made more discoverable and interoperable if it is able to be included in OAI systems because it can then be aggregated with other digital library metadata. Finally, even if the attempt to expose EAD metadata available to OAI systems is not successful, the attempt can teach us much about the challenges to be addressed in constructing a harvest and search mechanism for archival materials.

Attempting to apply the OAI metadata harvesting protocols to EAD therefore has broad implications. In particular, the analysis provided here suggests that in some cases it may be useful to create many OAI records from one source record, compelling a reexamination of some fundamental assumptions about the OAI protocols. In addition, the analysis of EAD encoding guidelines suggests that the archival community could help enhance the interoperability of finding aids by regularizing tagging practices or developing a single best practice guide to replace the numerous standards currently in existence.

2. EAD BACKGROUND

Encoded Archival Description is the SGML/XML metadata format used by archives, manuscript libraries, and museums to encode finding aids. Although finding aids take many forms, they generally provide content describing the nature and provenance of

a manuscript collection or archives. They usually include an inventory of the actual materials in the collection, at a file folder or (less frequently) item level. In simple terms, an EAD record is to archives what a MARC record is to books [17]. However, EAD records are much more lengthy, complex, and loosely structured than MARC records because archives are much more lengthy, complex, and loosely structured than books. While a typical MARC record describes a single book or serial, an EAD file provides detailed descriptions of the often hundreds or thousands of files or items which comprise a manuscript collection or an archival record series. At the same time, it must be realized that EAD records provide collective description, because archival description is collective in nature. EAD files range in size from several kilobytes to several megabytes.

EAD acts as a wrapper for collective archival description, but it does so with differing levels of specificity depending on the nature of the collection and the finding aid. Archival practice in constructing finding aids varies widely from institution to institution, and EAD was designed to accommodate differences while encouraging as much uniformity as possible by standardizing commonly used data elements.

An EAD record has two main components: metadata about the finding aid (i.e. the electronic document describing the collection), and metadata about the collection described in the finding aid. Metadata about the finding aid is contained within required elements nested within the high level element `<eadheader>`. The information encoded in the `<eadheader>` provides summary information about the finding aid, including its title, author, and date of creation.

Metadata about the collection described in the finding aid is encoded in the high level element `<archdesc>`. The `<archdesc>` usually comprises the bulk of the finding aid and may include numerous sub-elements, most of which are repeatable. In general, however, `<archdesc>` utilizes two types of elements—those that describe the collection as a whole, and those that describe subordinate components of the collection such as a series of files, an individual file, or an item. Elements that describe parts of the collection are included, appropriately enough, within EAD's "description of subordinate components" (`<dsc>`) tag. Within the `<dsc>`, information about the components of the collection is encoded in hierarchically nested `<cN>` tags (where N represents a number from '01' to '12'). It is also possible to use generic `<c>` tags recursively in place of the numbered component tags, but doing so is much less common.

The distinction between `<dsc>` and non-`<dsc>` of elements is not as clear as it might appear at first glance because every element which can be used to describe the collection as a whole is also available at any point within the 12 hierarchical components of the `<dsc>`. An overview of the `<archdesc>` adapted from the *EAD Application Guidelines* is provided in Table 1.

While EAD encourages archivists to use collective and multilevel description—practices which indeed follow from the archival theories of *fonds* and provenance [10]—the manner in which these practices are applied varies tremendously from finding aid to finding aid. EAD encourages a wide range of possible tagging practices.

Table 1: Overview of <archdesc>

```
<archdesc>
  <did>
  <add>
  <admininfo>
  <arrangement>
  <bioghist>
  <controlaccess>
  <dao> and <daogrp>
  <note>
  <odd>
  <organization>
  <scopecontent>
  <dsc>
    <c01>
      <did>
      (repeat as above)
    <c02>
      <did>
      (and so forth, to level <c12>)
```

Table 2 provides an example of how two different institutions might encode the same information. While this example seems trivial, the implications are not. The machine handling of EAD documents poses special problems because it is difficult to predict how an individual depository will apply archival practices of collective description. In particular, it is difficult to know how an institution will use the hierarchical elements.

Although this paper cannot hope to provide an exhaustive list of machine handling problems that one might reasonably expect to be encountered in converting EAD to OAI records, three bear particular emphasis at this point, since they will impinge upon the success of any scheme which aims to produce OAI-compliant records from an EAD file.

First, how can one provide full context for OAI records drawn from EAD's <dsc> (description of subordinate components) section? It is relatively easy to map the non-<dsc> descriptive information into OAI—i.e. to create one collective OAI record. Such a record could either describe the finding aid itself (as a digital object) or the materials described in the finding aid (the physical archives). But such a simplistic record will often not uncover the complete richness of the materials described, nor will it meet the needs of many researchers. For example, consider the hypothetical case of a researcher looking for letters to Theodore Dreiser. An EAD finding aid is used to describe a manuscript collection called the "Woodrow Wilson Papers," which consists of 5,000 items, including a few Dreiser letters. Unless Dreiser is listed in the top level of the Wilson EAD file, any OAI record produced from the EAD file cannot provide enough information to direct a researcher to the material in which she is interested. To provide such specificity, it is necessary to map metadata from

EAD's description of subordinate components. Here several problems immediately arise. First, should one map each level, or only the lowest level in each hierarchy? How will metadata from one level be inherited to the next? Will enough metadata be provided in the final record to make it usable? How might an OAI-based system point the user back to the full context of the hit? Can the system preserve the organic integrity of the finding aid? None of these questions would necessarily arise when a human being uses a finding aid since the level of description can often be inferred from contextual clues. But if an isolated record is created in OAI, the context may be difficult to preserve.

For instance, consider the example shown in Table 3, which is taken from an actual EAD finding aid encoded by the Minnesota Historical Society [11]. In order to provide a useful OAI record for the folder containing "Minutes of Board Meetings," it is necessary that an OAI system provide the user information from prior levels in the EAD hierarchy. It is not useful for researchers to know that a folder contains meeting minutes (level <c04>), unless they also know the minutes are those of the Hispanic Ministry Advisory Board (level <c03>) created by the Archdiocesan Office of Hispanic Ministry (level <c02>), stored in a series of material related to Hispanic Organizations in Minnesota (level <c01>), which is part of the Irene Gomez-Bethke Papers (<archdesc><did>), found at the Minnesota

Table 2: Encoding Differences

```
<c02>
  <did><container type="box">23</container>
    <unittitle>Correspondence to J. R. R. Tolkien,
      <unitdate>1945</unitdate></unittitle>
    <physdesc>includes 21 letters</physdesc>
  </did>
</c02>
      -vs.-
<c02 level="subseries">
  <did><container type="box">23</container>
    <unittitle>Tolkien, J. R. R. (John Ronald
      Reuel),<unitdate>1945</unitdate></unittitle>
  </did>
    <c03 level="file"><did>
      <physdesc>21 letters.</physdesc></did></c03>
</c02>
```

Historical Society. Providing access to these layers via an OAI-based system is essential if the record is to be usable to a user. This does not necessarily mean that an individual OAI record needs to contain metadata from all levels of the EAD hierarchy, but that a system manipulating EAD-OAI records must be able to point the user to the "hit" in its original context within the full finding aid.

Table 3: Hierarchical Inheritance

```
<archdesc type="inventory" level="collection">
<did id="a1"><unittitle>Irene Gomez-Bethke papers</unittitle><unitdate>1970-1993.</unitdate></did>
<c01>
  <did><unittitle>Hispanic Organizations in Minnesota:</unittitle></did>
  [other c02 levels not shown]
  <c02>
    <did><physloc>151.H.1.1B</physloc><container>1</container><unittitle>Archdiocesan Office of Hispanic
      Ministry:</unittitle></did>
    [other c03 levels not shown]
    <c03>
      <did><unittitle>Hispanic Ministry Advisory Board:</unittitle></did><scopecontent><p>Advised the
        archbishop.</p></scopecontent>
      [other c04 levels not shown]
      <c04>
        <did><unittitle>Minutes of Board Meetings, 1986-1989.</unittitle></did>
      </c04>
    </c03>
  </c02>
</c01>
</archdesc>
```

The second problem which one might expect to encounter in the machine handling of EAD records is related to the first: How does the computer (and thus its human user) know what level of materials are being described? Ideally, OAI records from EAD should specify the level of granularity because researchers will be interested in knowing whether the search results describe an individual letter, a file folder, a series of files, or an entire collection. Unfortunately, the EAD DTD only rarely requires a standardized description for the level of materials being described. Both the <archdesc> and the <cN> tags include a LEVEL attribute, which allows the values of record group, subgroup, series, subseries, collection, file, fonds, item, and otherlevel. (If otherlevel is used, the value is set using an OTHERLEVEL attribute; any alphanumeric value is allowed.) However, the LEVEL attribute is required only in the <archdesc> tag, and an examination of selected finding aids gathered for the Illinois OAI project indicates that it is rarely used beyond <c01>.

Finally, flexibility in the DTD means that it may be difficult to produce a general purpose tool for exposing OAI records to a harvester. Any institution which wishes to provide OAI metadata must transform their metadata into OAI-compliant records for exposure to OAI harvesters. The OAI FAQ page acknowledges that this involves quite a bit of work on the part of providers since "[r]esponding to protocol requests also involves accessing or extracting your metadata. Giving a time estimate for this is difficult since the nature of the task is entirely idiosyncratic" [14]. Transformation of an EAD record to OAI is most easily accomplished using an XSLT stylesheet, but it is unlikely each institution with EAD files will be able to write or adapt its own stylesheet and associated scripts for OAI. Many archives have already invested a substantial amount of time and effort in EAD itself as a preferred metadata format and experience many problems using XSLT stylesheets to transform their EAD to HTML, as the archives of the EAD listserv testify [2]. While it is unlikely they will write another set of stylesheets, they may adopt

an off-the-shelf tool that can produce OAI from their existing EAD files. Producing such a tool is a major goal of the University of Illinois OAI project.

Our attempt to map EAD to OAI is predicated on the assumption that one EAD file should be mapped to multiple OAI records. Other alternatives exist, but they hold distinct drawbacks. For example, one could simply map the collective "top level" description from an EAD file to a single OAI record. This would provide collection-level access. But such a simple record would not help users uncover the richness buried in many finding aids. For example, users could not directly find digital objects referred to in the far reaches of many files. On the other hand, it would be theoretically possible to create one very large OAI record for each EAD file including all of the metadata in the <dsc>. However, such an OAI record would lack the specificity and precision needed to be effectively manipulated in an OAI system.

3. ENCODING STANDARDS AND PRACTICES

The problems discussed above do not necessarily mean it will be impossible to map an EAD file to OAI using a general purpose tool. Luckily the archival profession has taken some steps to alleviate potential problems which might arise in the machine handling of EAD records. The most significant has been the development of EAD workshops sponsored by the Society of American Archivists and the Rare Book School, as well as the development of best practice documents and encoding guidelines. Institutions have realized that there needs to be a standardized practice for EAD. As noted in the American Heritage Virtual Archive Project encoding guidelines:

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, [using EAD] does not in and of itself ensure that machine-readable finding aids will

be easily communicated between repositories, nor facilitate the building of union databases. 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. . . . Predictability and stability are essential for the existence of communities. [1]

It is possible that training and encoding standards have informally served to standardize mark-up practices. Unfortunately, different communities have developed different standards. For this reason, it is necessary to examine the existing encoding guidelines and a sample of finding aids from different institutions. Doing so will help us determine whether enough uniformity exists under current practices to allow for development of a general tool to provide OAI records from EAD metadata. Such a tool could then be tested against actual EAD files.

At the time this article was written, eight best practice encoding standards were available electronically for examination on a website provided by the Research Libraries Group's EAD Advisory Group [4]. In addition, the EAD Cookbook provides commonly used encoding templates and is used by many institutions. Finally, the *EAD Application Guidelines* offers basic guidance on recommended tagging practices [16]. Examining these ten guidelines can help gauge the consistency of tagging practices.

The guidelines were examined for their treatment of several issues related to the consistency of encoding, including use of the <eadid> tag to facilitate identification, the manner in which the collection title, abstract, controlled subjects, and depository information are recorded at top level of finding aid, recommendations for use of the LEVEL attribute to indicate the nature of materials being described, the use of controlled vocabularies and standard formats of names and dates, the recommended tagging practice for representing the collection's hierarchy and for recording file and item titles. Admittedly, these issues are not the only ones which might impede the machine handling of EAD finding aids in an OAI environment. But examining them provides a reasonable basis for predicting expected problems which might be encountered. Accordingly, the appropriate sections of each guideline were examined and loaded into a database to allow for comparison and analysis.

The results of this analysis indicate that the encoding guidelines share a general consensus over the proper way to encode a finding aid. For example, all of the guidelines recommend that a title, originator, and abstract should be encoded in the appropriate <archdesc><did> elements, so that metadata for the collection as a whole is encoded in a predictable location. In addition, six of the guidelines recommend that the LEVEL attribute for the <archdesc> should be set to "collection". This indicates that many institutions use a single EAD file solely to encode metadata for entire collections of personal papers or archival records. While it is possible that some institutions are using distinct EAD files to encode descriptions of individual series, subseries, files, or even items, this does not appear to be standard practice. No examples were found in a cursory search of the records donated to the University of Illinois OAI project. At the <archdesc> level of the finding aid, the encoding guidelines usually recommend using a controlled vocabulary and name authority file, which may

produce more interoperable OAI records. The description of subordinate components in the collection (e.g. series, file, and item titles) is also handled fairly consistently by the encoding guidelines, which recommend that information should be encoded hierarchically within <unittitle> tags in the various <cN> elements. All of the guidelines recommend that information should not be encoded redundantly. Information should only be encoded at the highest possible level of the tree, so that child elements inherit metadata from their parents. An informal examination of EAD files contributed to the University of Illinois OAI project indicates that these recommendations have usually been followed, so it is probable that useful OAI records can be harvested from subordinate parts of an EAD file.

Yet in spite of the encoding guidelines' agreement on some basic matters, important differences do exist. While none of these differences necessarily mean that EAD finding aids cannot be successfully transformed into OAI, they do need to be dealt with when designing a transformation process.

For example, recommendations for the use of the LEVEL attribute within the subordinate components of the finding aid are not consistent. Only one encoding guideline requires that the LEVEL attribute be set for every <cN> element, including those describing files or items. Most other guidelines recommend that it be set only at the higher levels, such as "series" or "subseries." It is common for the level to be specified at the <c01> level; the templates accompanying the EAD Cookbook, to cite only one example, set the level at <c01> to "series" by default. Yet below the <c01>, it may be very difficult for a computer to tell exactly what is being described if the LEVEL attribute is not set. For example, if a component level shows <unittitle>Robert Redford</unittitle>, does this level describe several boxes of materials, a file folder, or an individual item? A little human intuition can help. For example, if Redford is the last node on a tree, it is likely that the materials described are a file or an item. But if the Redford element includes several children, it is likely to be a series or subseries of materials. Although the actual level will always be ambiguous unless it is specified in the LEVEL attribute of the <cN> tag, predicting a level at the point of transformation from EAD to OAI may be both possible and, on balance, beneficial.

Another inconsistency in EAD practice which can be mitigated at the time of transformation to OAI emerges from the treatment of the <eadid> tag in the <eadheader>. As noted in the *EAD Application Guidelines*, this element is intended to provide an identifier which will be unique across the entire range of finding aids encoded worldwide. Several options are mentioned by the *Application Guidelines*, including the use of URLs, file names, locally developed identifier numbers, or SGML catalog entries. Accordingly, the *EAD Application Guidelines* provide several possible methods for encoding an identifier for the EAD file. Many EAD files in SGML format use catalog entries, which can be used to point to a source file location. But SGML catalog entries are not available in XML, and institutions creating XML finding aids appear to have taken a variety of approaches in the <eadid>. Harvard uses an eight digit code providing the physical location of the collection and a number "for the finding aid." The Library of Congress uses persistent identifiers (URNs) to register handles for each finding aid. The University of Michigan uses the Bentley Historical Library call number. The EAD Cookbook

recommends using a "value that is suitable as a computer file name."

Since so many methods may be used to represent the ID, the information in the <eadid> element will not be sufficient to derive the location of the EAD file itself. Under current practice, OAI metadata providers who wish to provide EAD metadata will need to provide a URL in a Dublin Core identifier element which is added at the point of transformation. If providers want to export EAD directly through OAI, <eadid> usage may need to be standardized. For example, the value could be limited to URLs or to URNs like those used by the Library of Congress [9].

4. A RECOMMENDED EAD TO OAI CROSSWALK

An examination of EAD encoding guidelines shows that while it should be possible to map OAI to EAD, the process will not be straightforward. The OAI Protocols require that OAI repositories export OAI records in unqualified Dublin Core. Additional metadata formats (such as qualified Dublin Core or EAD itself) may also be exported, but they must be specified and described in an XSD schema. An XSD schema does not presently exist for EAD, so it appears likely that dcterms (qualified Dublin Core) would be the most useful metadata format in which to export records if they are to be compared to other OAI records generated from non-EAD sources. Accordingly, the following discussion proposes a recommended procedure to map EAD into Dublin Core and dcterms. This option is preferred over unqualified Dublin Core since it would provide better functionality and the ability to show the relationships between OAI records drawn from a single EAD file.

4.1 Mapping the Top Level

The *EAD Application Guidelines* include two recommended mappings to Dublin Core, one for the finding aid, and another for the resources described in the finding aid. This ambiguity mirrors one in Dublin Core. Although the Dublin Core standard originated in a need to describe electronic resources, it may now also be applied to collections of documents, books and other tangible items [6]. In EAD, metadata about the finding aid is ostensibly encoded in the <eadheader>, but certain types of useful metadata are not included in the recommended mapping to Dublin Core or can be included only in a cumbersome fashion [12]. By the same token, researchers are likely to be more interested in the material described in the finding aid than in the finding aid itself, but not all of the pertinent information is encoded in the <archdesc> where one would expect to find it. As a result, neither mapping given in the *Encoding Guidelines* would generate records adequate enough for manipulation in an OAI system. It is necessary to take a flexible approach in mapping the top level of an EAD finding aid into an OAI record. The recommended mapping given below draws metadata from both the <eadheader> and <archdesc>, but emphasizes the <archdesc>.

In the schema proposed here, OAI records describing an entire EAD finding aid may be referred to by other OAI records drawn from the same source EAD document. For that reason, it is very important that the top-level record be as complete, accurate, and concise as possible. The base record needs to include a URL

pointing to the source EAD file as well as references to related parts of the collection which are described in other OAI records. Although the *Encoding Guidelines* suggest that the <eadid> tag be mapped into a Dublin Core identifier, encoding practices with the <eadid> tag are simply too disparate to use it as the basis for file identification and retrieval. In the schema being tested by the University of Illinois OAI project, the file's URL serves as both the OAI record identifier and as a DC identifier.

To allow for effective file manipulation, sorting, and retrieval, it is also necessary to modify the *Encoding Guidelines'* recommended mapping to the DC type field. The *Application Guidelines* recommend mapping the LEVEL attribute of the <archdesc> element, but the LEVEL attribute contains many possible values not approved in the Dublin Core Metadata Initiative Type Vocabulary [7]. In addition, it should be noted that although the DC Type working group has discussed forming a list of approved subtypes, no consensus on the need for such a list has emerged. Some working group members have suggested that domain- and application-specific types should be defined [8]. Given the uncertainty at this time, it appears sensible to use multiple DC type tags in OAI records describing archival materials. An approved type such as "Text" could be included in the first type tag, and an additional type tag could be set to "Archives or Manuscripts" in order to refine the general type. Finally, a third type tag could be used to provide more specificity by encoding the LEVEL attribute of the <archdesc>.

Appendix 1 provides a suggested schema (in XPath syntax where space allows) for mapping the base OAI record from EAD into Dublin Core and dcterms. Table 4 provides a sample of the XSLT code used to produce part of the top level record.

4.2 Mapping the Description of Subordinate Components

Once a top level OAI record is created for each EAD file, it is also necessary to map the description of subordinate components to create separate but linked OAI records. The suggested approach attempts to replicate the hierarchical structure of EAD, limit redundancy and allow for adequate resource discovery. It replicates the hierarchical model of EAD by using Xpointers within dcterms "IsPartOf" and "HasPart" fields to point to the part of the source EAD document where the original reference is found [19]. While current web browsers do not support XPointer, it is hoped that references can be manipulated via a traditional server-side XSLT transformation, so that users of the search interface can see the "hit" in the context of the entire finding aid. Each OAI record includes references to both the immediate parent and to any children of the <cN> element being mapped. Although the primary benefit of this approach lies in facilitating file retrieval, it has a subsidiary benefit in allowing more immediate access to the original context for the citation. From an archival point of view, this serves an important function in providing the full provenance and context for an item or file retrieved. Appendixes 2 and 3 show the suggested mapping (given in XPath syntax where possible) from the context of the current <cN> node being processed, and sample OAI output records generated from the <dsc> of two sample finding aids.

Table 4: XSLT Code for top level HasPart

```
<!--hasPart-->
<xsl:for-each select="/ead/archdesc/dsc/c01/did/unittitle">
  <xsl:element name="dcterms:hasPart">
    <xsl:element name="rdf:Description">
      <xsl:attribute name="rdf:about">
        <xsl:value-of select="$identifier"/><xsl:text>#xpointer(//c01[</xsl:text><xsl:number
          count="c01"/><xsl:text>])</xsl:text>
      </xsl:attribute>
      <xsl:element name="dc:identifier">
        <xsl:value-of select="$identifier"/><xsl:text>#xpointer(//c01[</xsl:text><xsl:number
          count="c01"/><xsl:text>])</xsl:text>
      </xsl:element>
      <xsl:element name="dc:title">
        <xsl:value-of select="normalize-space(text()*[not(self::unitdate)])"/>
      </xsl:element>
      <xsl:element name="dc:date">
        <xsl:value-of select="./unitdate|../unitdate"/>
      </xsl:element>
    </xsl:element>
  </xsl:element>
</xsl:for-each>
```

5. IMPLICATIONS

EAD is a very flexible and complex metadata format. This means that it cannot be mapped into OAI in a straightforward fashion. Some information will inevitably be lost in transformation, and EAD records will remain most fully searchable in their native format.

However, an analysis of encoding guidelines for the creation of EAD finding aids indicates that EAD metadata may be consistently enough structured to allow for the generation of OAI-compliant records that can be harvested, aggregated with non-EAD records, and searched using tools provided by OAI harvesters. The suggested mapping relies on archival theory and practice, but also makes compromises on the premise that it will generate useful OAI records that can be searched alongside records developed from other sources, such as MARC records and locally-developed databases.

Stay tuned. The success or failure of this approach will be measured during development of the harvester and search portal for the University of Illinois's OAI project.

6. REFERENCES

- [1] American Heritage Virtual Archive Project, "The Encoded Archival Description Retrospective Conversion Guidelines," February 24, 1999. Available at <http://sunsite.berkeley.edu/amher/upguide.html>.
- [2] Archives of ead@listserv.loc.gov. Available at <http://listserv.loc.gov/listarch/ead.html>.
- [3] Bergman, M. K., The Deep Web: Surfacing Hidden Value. JEP: the Journal of Electronic Publishing 7,1 (Aug 2001). Available at <http://www.press.umich.edu/jep/07-01/bergman.html>
- [4] Best Practices & Guidelines for RLG EAD Advisory Group Members. Available at http://www.rlg.org/primary/eadac_practices.html.
- [5] Brockman, William S., Laura Neumann, Carole L. Palmer, and Tonyia J. Tidline, Scholarly Work in the Humanities and the Evolving Information Environment. Washington DC: Digital Library Federation, December 2001. Available at <http://www.clir.org/pubs/reports/pub104/contents.html>.
- [6] DCMI Frequently Asked Questions (FAQ). Available at <http://dublincore.org/resources/faq/#whatisaresource>.
- [7] DCMI Type Vocabulary. Available at <http://dublincore.org/documents/dcmi-type-vocabulary/>.
- [8] DCMI Type Working Group. <http://dublincore.org/groups/type/>.
- [9] The Digital Object Identifier. <http://www.doi.org/>.
- [10] Hensen, Steven, 'NISTF II' and EAD: The Evolution of Archival Description. American Archivist 60, 3 (1997): 284-96.
- [11] Irene Gomez-Gethke: An Inventory of Her Papers. Available at <http://www.mnhs.org/library/findaids/00039.html>.
- [12] Kiesling, Kris. Metadata, Metadata Everywhere--- but Where is the Hook?. OCLC Systems & Services 17, 2 (2001): 84-88.
- [13] 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://www.acm.org/jcdl/past-event-conf.shtml>

[14] Open Archives Initiative FAQ.
<http://www.openarchives.org/documents/FAQ.html>.

[15] Prom, Christopher J. The EAD Cookbook and the Future of Descriptive Standards: A Survey and Usability Study. Mss. currently under review by American Archivist.

[16] Society of American Archivists, Encoded Archival Description Application Guidelines, ver. 1.0. Chicago, IL: Society of American Archivists (1999). Available at <http://www.loc.gov/ead/ag/aghomet.html>.

[17] University of Illinois at Urbana-Champaign Open Archives Initiative Metadata Harvesting Project.
<http://oai.granger.uiuc.edu/>

[18] Waters, Donald J. The Metadata Harvesting Initiative of the Andrew W. Mellon Foundation. ARL Bimonthly Report 217 (August 2001). <http://www.arl.org/newsltr/217/waters.html>.

[19] W3C XML Pointer, XML Base and XML Linking.
<http://www.w3.org/XML/Linking>.

APPENDIX 1: TOP LEVEL MAPPING

DC.dcterms	Value from EAD file	Notes
Identifier	URL of EAD file being transformed	
Title	/ead/archdesc/did/unittitle /ead/eadheader/filedesc/titlestmt/titleproper	
Date	/ead/archdesc/did/unittitle/unitdate /ead/archdesc/did/unitdate	
Creator	/ead/archdesc/did/origination	Include data encoded in any child tags as well as in the node
Type	"Text" "Archives or Manuscripts" /ead/archdesc/@level	use three type tags
description.abstract	/ead/archdesc/did/abstract	used to encode a brief description
Description	/ead/archdesc/scopecontent	provides more lengthy description
format.extent	/ead/archdesc/did/physdesc/	
Subject	/ead/archdesc/controlaccess//corpname	Repeatable; select each of the nodes listed in this grouped set: corpname, famname, function, genreform, name, occupation, persname, subject, title
coverage.spatial	/ead/archdesc/controlaccess//geogname	
Publisher	/ead/archdesc/did/repository eadheader/filedesc/publicationstmt/publisher	
Language	/ead/archdesc/@langmaterial	
relation.HasPart.identifier	number of the current <c01> node being processed	See XSLT code in Table 4.
relation.HasPart.title	unittitle of current <c01> node being processed	
relation.HasPart.date	unitdate of current <c01> node being processed	

APPENDIX 2: <DSC> MAPPING FROM CONTEXT OF <cN> Node

DC.dcterms	Value from EAD file	Notes
Identifier	Prefix: URL Suffix: position of current <cN> node expressed as an Xpointer	Example: http://web.library.uiuc.edu/ahx/ead/ua/2620058/2620058.xml#xpointer(//c01[1]/c02[1]/c03[3])
Title	did/unittitle .	Choose entire <cN> node if unittitle tag does not exist
Date	did/unittitle/unitdate did/unitdate	
Creator	did/origination	
Type	"Text" "Archives or Manuscripts" @level of current <cN> node	Use three type tags
description.abstract	did/abstract scopecontent in the current <cN> node	
Format	Physdesc	
Subject	controlaccess//corpname	Repeatable; see Appendix 1
coverage.spatial	controlaccess//geogname	
relation.HasPart.identifier	Prefix: URL of EAD file Suffix: XPointer corresponding to the child of <cN> node being processed	Use <xsl:for-each> to process each <cN+1> child of the current node
relation.HasPart.title	cN+1/did/unittitle cN+1	
relation.HasPart.date	cN+1/did/unittitle/unitdate cN+1/did/unitdate	
relation.IsPartOf.identifier	Prefix: URL of EAD file Suffix: XPointer corresponding to the parent of <cN> node being processed	Use ".." syntax to select node <cN-1> of node being processed
relation.IsPartOf.title	../did/unittitle ..	
relation.IsPartOf.date	../did/unittitle/unitdate ../did/unitdate	

APPENDIX 3: SAMPLE OAI RECORDS FROM EAD's DESCRIPTION OF SUBORDINATE COMPONENTS

Thick Record

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://bolder.grainger.uiuc.edu/cornell/RMA00001.xml#xpointer(//c01[1]/c02[1]/c03[1]/c04[1])">
    <dc:identifier>http://bolder.grainger.uiuc.edu/cornell/RMA00001.xml#xpointer(//c01[1]/c02[1]/c03[1]/c04[1])
    </dc:identifier>
    <dc:title></dc:title>
    <dc:date>June 17, 1828 - September 22, 1830</dc:date>
    <dc:type>text</dc:type>
    <dc:type>archives or manuscripts</dc:type>
    <dc:type>file</dc:type>
    <dc:description>August 23, 1830. E.B. Cornell to Ezra Cornell from Manlius: "It's very sickly about here now, there is about 2
      hundred patients under the physician's care."</dc:description>
    <dc:subject>Family correspondence</dc:subject>
    <dc:subject>family health</dc:subject>
    <dc:subject>travel</dc:subject>
    <dc:subject>Quaker Meeting</dc:subject>
    <dc:subject>personal finances</dc:subject>
    <dc:subject>Cornell, Elijah</dc:subject>
    <dc:subject>Cornell, E.B.</dc:subject>
    <dc:subject>Eddy, Otis</dc:subject>
    <dc:subject>Cornell, Mary Ann.</dc:subject>
    <dcterms:spatial xmlns:dcterms="http://purl.org/dc/terms/">DeRuyter, N.Y.</dcterms:spatial>
    <dcterms:spatial xmlns:dcterms="http://purl.org/dc/terms/">Manlius, N.Y.</dcterms:spatial>
    <dcterms:spatial xmlns:dcterms="http://purl.org/dc/terms/">Ithaca, N.Y.</dcterms:spatial>
    <dcterms:isPartOf xmlns:dcterms="http://purl.org/dc/terms/">
      <rdf:Description rdf:about="http://bolder.grainger.uiuc.edu/cornell/RMA00001.xml#xpointer(//c01[1]/c02[1]/c03[1])">
        <dc:identifier>http://bolder.grainger.uiuc.edu/cornell/RMA00001.xml#xpointer(//c01[1]/c02[1]/c03[1])
        </dc:identifier>
        <dc:title>Correspondence :: Ezra Cornell Correspondence ::</dc:title>
        <dc:date>1828-1845</dc:date>
      </rdf:Description>
    </dcterms:isPartOf>
  </rdf:Description>
</rdf:RDF>
```

Thin Record

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://bolder.grainger.uiuc.edu/cornell/RMA01466.xml#xpointer(//c01[4]/c02[2]/c03[38])">
    <dc:identifier>http://bolder.grainger.uiuc.edu/cornell/RMA01466.xml#xpointer(//c01[4]/c02[2]/c03[38])
    </dc:identifier>
    <dc:title>Voodoo Death, [1942]</dc:title>
    <dc:type>text</dc:type>
    <dc:type>archives or manuscripts</dc:type>
    <dc:type>file</dc:type>
    <dcterms:isPartOf xmlns:dcterms="http://purl.org/dc/terms/">
      <rdf:Description
        rdf:about="http://bolder.grainger.uiuc.edu/cornell/RMA01466.xml#xpointer(//c01[4]/c02[2])">
        <dc:identifier>http://bolder.grainger.uiuc.edu/cornell/RMA01466.xml#xpointer(//c01[4]/c02[2])
        </dc:identifier>
        <dc:title>Notes and Files</dc:title>
      </rdf:Description>
    </dcterms:isPartOf>
  </rdf:Description>
</rdf:RDF>
```