Encoded Archival Description: Data Quality and Analysis

Luis Francisco-Revilla[1], **Ciaran B. Trace**[2], **Haoyang Li**[3], **Sarah A. Buchanan**[4]
[1]Texas Advanced Computing Center, [2,3,4]School of Information
The University of Texas at Austin
Austin, TX, 78701, USA
[1]revilla@tacc.utexas.edu, [2]cbtrace@ischool.utexas.edu, {[3]haoyang.li [3]sarahab}@utexas.edu

ABSTRACT

In order to authenticate the meaning of collections and to preserve their evidentiary value, archivists create documents (*finding aids*) that describe the provenance and original order of the records (MacNeil, 1995). Metadata standards such as Encoded Archival Description (EAD) enable finding aids to be encoded, searched, and displayed online. However, recent research has begun to draw attention to problems with the quality of EAD finding aid data and metadata, and the encoding practices by which finding aids are created. Since the next frontier in archival description involves reusing finding aid data for advanced information visualization techniques that support additional ways of engaging with collections, there is a pressing need for further study of data quality and how it might impact information visualization. This work analyzes a set of 8729 finding aids aggregated by the Texas Archival Repository Online (TARO) using VADA, a visual analytic tool for finding aids. The results show previously unidentified problems that have significant impact on the ability to visualize this data. The paper explains how these problems relate to both EAD's design and the actual encoding practices of EAD, and provides recommendations for improving the quality of finding aid data.

Keywords

Encoded Archival Description, finding aids, archives, metadata quality, evaluation of document standards.

INTRODUCTION

Finding aids are the defining document for the field of archival science. The overall archival workflow hinges around them, as they represent the outcome of the work of archivists, and the beginning of the work for archival users (Trace & Dillon, 2012). Organizations and individuals create and accumulate records as byproducts of everyday activity. These collections are uniquely valuable, as they provide evidence about the past activities of their creators. Finding aids capture information and metadata about the provenance and original order of the records; contexts which are crucial to archival users in understanding the significance of the material. By understanding the nature of finding aids, and the process by which they are created, it is possible to devise metadata encoding standards that allow for the standardization, display, discovery, aggregation, analysis, and sharing of descriptive records through time.

The Society of American Archivists (SAA) defines a finding aid as "a single document that places the materials in context by consolidating information about the collection, such as acquisition and processing; provenance, including administrative history or biographical note; scope of the collection, including size, subjects, media; organization and arrangement; and an inventory of the series and the folders."[1] According to SAA, a finding aid functions as both:

1. a tool for information discovery in a collection of records
2. a description of records that gives the repository physical and intellectual control over the materials, and helps users access and understand the materials

As the Library of Congress (LOC) explains, finding aids provide a gateway that "helps users discover and navigate through the thousands of boxes and folders that house each collection"[2]. The finding aid summarizes the scope of the collection, conveys information about the individuals and organizations involved, and gives an inventory of its contents. Therefore, a finding aid provides a comprehensive overview of a collection, and progressively detailed descriptions of its components parts.

A key facet of archival description is that it incorporates the notion that the arrangement and description of collections is by necessity hierarchical. This hierarchy instantiates the external and internal contexts that archivists believe must

77th ASIS&T Annual Meeting, October 31- November 4, 2014, Seattle, WA, USA.

[1] http://www2.archivists.org/glossary/terms/f/finding-aid

[2] http://www.loc.gov/rr/ead/

be associated with the collection in order for that collection to maintain its evidentiary value (Trace & Francisco-Revilla, 2014). The hierarchy unfolds as various divisions or levels that are described in the finding aid. This approach provides a consistent structure across collections.

- *Fonds* or *Record Group*: all records of an entity/individual with the same provenance
- *Subgroup*: subset of records with a distinct external structure of provenance (external context)
- *Series*: group of similar records created, received or used in the same function or activity, and filed accordingly (internal context, which captures the documentary context of the material)
- *Subseries*: a set of documents within a series, distinguished from the whole by a filing arrangement such as type, form or content
- *File*: set of documents related to same matter or event

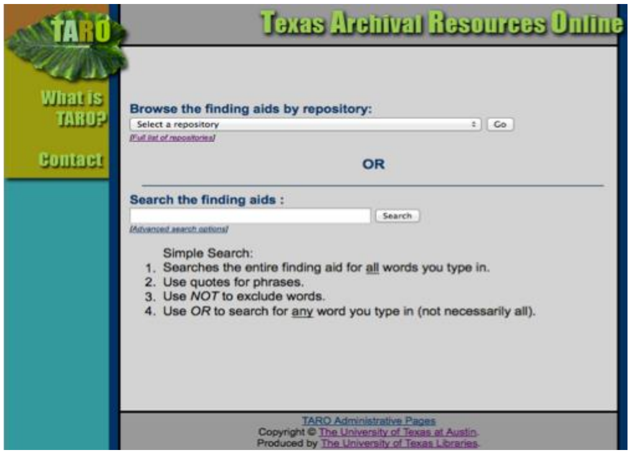
A standard encoding model is one mechanism that can be used to publish, navigate, and search finding aids online. Currently, most American archival institutions encode finding aids in XML, following the Document Type Definition (DTD) or schema for Encoded Archival Description (EAD).

EAD is an international descriptive markup standard for the archival community. Its ongoing development is handled by SAA Standards Committee's Technical Subcommittee for Encoded Archival Description (TS-EAD), with the documentation hosted by LOC. In developing EAD, the archival community sought to:

- Help the archival community demarcate the critical components of archival description
- Represent this data in a manner that protects the content and intellectual organization of the materials
- Provide access to archival description such that people can "discover or locate archival materials at any time and from any place"
- Ensure the sustainability and the longevity of this data over the long term (Pitti, 1999).

EAD was created as a flexible data model in order to facilitate the encoding of legacy finding aids, accommodate the significant variations in procedure across archives, and to allow for the emergence of new archival descriptive practices (Shaw, 2001). This permissiveness of the EAD data model has been an important factor in its widespread adoption within the archival community (Shaw, 2001). However, this has also created significant difficulties achieving the larger goals of EAD, vis-à-vis the processing of data for resource discovery, aggregation, and cross-repository sharing (Shaw, 2001). Shaw describes it as a tension between "attempting to describe the data retrospectively (that is, as it already exists)," and "trying to find a common, standard way of describing the data so that we can create processes that enable us to share it widely" (Shaw 2001, p. 119).

This paper looks at how the tension between data access and data aggregation play out within Texas Archival Resources Online (TARO), a consortium site established in 1998 that now facilitates access to over eight thousand collections from 36 archival repositories throughout the state of Texas.[3] The authors' long-term research goal is to use a subset of the TARO data to visualize cross-institutional patterns and trends in the descriptive and encoding practices relating to archival arrangement, an area of archival practice whose efficiency and effectiveness has come under increasing scrutiny in recent years (Trace & Francisco-Revilla, 2014). However, this paper is a case study of the quality of the TARO finding aid data that relates to archival arrangement. Data quality is a foundational issue that must be addressed before robust data visualization and analysis can occur in this area. The following section provides background information, explaining the design of EAD and its most important characteristics. The next section discusses the challenge with EAD to facilitate both the creation of finding aids and their subsequent aggregation and analysis. Next, the paper discusses VADA, a tool that helped to uncover and highlights issues with regard to the quality of TARO finding aid data. The paper then discusses the quality of the aggregations of finding aids, identifying particular issues and connecting them to specific aspects of EAD. The paper finishes by summarizing the contributions of the work, and discussing the implications of those challenges for information visualization.


TARO
Texas Archival Resources Online
What is TARO?
Contact
Select a repository
Go
[Full list of repositories]
OR
Search
[Advanced search options]
TARO Administrative Pages

Browse the finding aids by repository:

Search the finding aids :

Simple Search:

1. Searches the entire finding aid for all words you type in.
2. Use quotes for phrases.
3. Use *NOT* to exclude words.
4. Use *OR* to search for any word you type in (not necessarily all).

Copyright © The University of Texas at Austin.
Produced by The University of Texas Libraries.

Figure 1. Texas Archival Resources Online

ENCODED ARCHIVAL DESCRIPTION

EAD is said to be "the latest in a series of efforts by archivists to help remote researchers identify collections relating to their topics" (Ruth, 2001, p. 29). Prior to the advent of database and markup technologies, researchers relied on a variety of methods to identify relevant archival collections including consulting printed and microfilm

[3] http://www.lib.utexas.edu/taro/

guides, footnote and reference chasing, consulting knowledgeable individuals including colleagues and archivists, and making in person trips to archival institutions (Ruth, 2001).

The advent of database and markup technologies fundamentally changed the way in which researchers accessed descriptions of archival collections. Beginning in the 1960s, the library community developed a standard for the communication of bibliographic records in machine-readable format, Machine-Readable Cataloging (MARC). Networked computer databases developed by library consortiums (Research Libraries Information Network (RLIN) and Online Computer Library Center, Inc. (OCLC)) further facilitated this data exchange through centralized union catalogs.

In the 1980s, archivists built on the developments of their library colleagues by adopting the USMARC Archival and Manuscripts Control (MARC AMC) format to serve as the standard for the organization, storage and exchange of summary or collection level information about archival collections in machine-readable form. As a data structure standard, MARC allowed archivists to demarcate and label certain component parts of the finding aid and, as such, allowed for the isolation of data fields for particular purposes such as indexing and display. In 1983, with the publication of Archives, Personal Papers, and Manuscripts (Hensen, 1983), the archival community put into place a set of companion rules (content standard) for formulating data within archival catalog records.

Archivists, however, were still searching for a way to provide remote union access to all of the descriptive information in the finding aid, including robust support for the multilevel hierarchy or level information. A potential solution emerged with the advent of the internet and the Gopher protocol at the University of Minnesota (Ruth, 2001; Hensen, 1997). As systems predating the World Wide Web, gophers allowed archivists to mount ASCII files on the internet, and users to preform keyword searches in and across the finding aids. However, there were problems with the gopher systems, as they did not support the various typographical conventions that were used to demarcate and add meaning to information within the finding aids (e.g., boldface and italics), and there was no way to logically or dynamically link from the MARC catalog record to the gopher finding aid (Ruth, 2001; Hensen, 1997).

In the early 1990s, Daniel Pitti at the University of California Berkeley Library, and a team of archivists and librarians at cooperating institutions revisited the idea of developing a non-proprietary archival encoding standard for machine-readable finding aids. Various options were considered and rejected by the Berkeley Finding Aid Project (BFAP) team, including gopher presentation of ASCII text, ASCII text marked up using HyperText Markup Language (HTML) tags, and MARC tagging using either existing or new implementations of MARC. Instead, Pitti settled on Standard Generalized Markup Language (SGML), a descriptive markup language used to facilitate the authoring of structured documentation. As a meta-language, SGML provided the grammar or rules that allowed the BFAP team to write a draft DTD (called FindAid) specifically for archival finding aids.

During the mid to late 1990s, Pitti and a team of experts in archival description (the so-called Bentley Fellowship) further reviewed, tested, and refined the DTD. This work involved analyzing existing finding aids, as well as evaluating existing archival descriptive standards (including the MARC AMC format and an international data standard that defined a list of elements and structure for archival description, General International Standard Archival Description (ISAD(G)) (Ruth, 2001). An alpha version of the DTD (renamed Encoded Archival Description) was released in February 1996 and at this juncture, the Society of American Archivists' EAD Working Group took on responsibility for the development of the DTD and associated documentation.

A beta version of EAD followed in mid-September 1996, accompanied by a beta version tag library and application guidelines. In the summer of 1998, version 1.0 of the EAD DTD was released in conjunction with an updated EAD tag library. In July 2000, an EAD cookbook (containing an encoding protocol, template files, and stylesheets) provided archivists with practical help to implement EAD in their repositories. With the release of EAD 2002, the EAD DTD was designed to function as both an SGML and XML DTD. In 2004, and again in 2013, the Society of American Archivists released versions of a companion content standard, Describing Archives a Content Standard (DACS). An EAD schema was released in 2007 in two syntaxes: Relax NG Schema (RNG) and W3C Schema (XSD). At the time of writing this paper, EAD 2002 is slated to be replaced by EAD 3 sometime in 2014 or early 2015.

In concert with the past twenty years of the development of EAD, has come an associated rise of institutions that have adopted this standard as the main mechanism for delivering finding aids on the web, although its implementation has not been universal (Yakel & Kim, 2005; Gracy & Lambert, 2014). Beginning in the 1990s, EAD consortial websites have also been developed to aggregate and provide a wider distribution channel for collection descriptions. These sites now include the Online Archive of California (OAC), Northwest Digital Archives (NWDA), and Texas Archival Resources Online (TARO) (Combs et al., 2010). In the case of the TARO project, this involved an initial conversion of legacy finding aids via a vendor service, with individual institutions responsible for quality control and additional mark-up of the returned files. Encoding of new TARO finding aids is now handled by individual repositories.

EAD Main Components

At the heart of EAD is a diverse set of data elements that can be used to describe the whole of an archival collection

(Pitti, 1999). EAD 2002 is composed of 146 elements (and associated attributes) that describe the finding aid itself, and a body of archival materials and its constituent components. These elements exist within three high-level containers:

<eadheader>
(Required). The EAD header is the prologue for the XML document and is a wrapper element for information about the electronic finding aid document.

<frontmatter>
(Optional). The front matter element bundles prefatory matter for the formal publication of the finding aid, with a focus on the creation, publication or use of the finding aid (rather than information about the archival materials being described).

<archdesc>
(Required). The archival description element is a wrapper element that assembles the information *about* the whole collection by describing "the content, context, and extent of a body of archival materials, including administrative and supplemental information that facilitates use of the materials" (Society of American Archivists, 2002).

Within <archdesc> there are two important elements:

<did>
As the only required element in <archdesc>, the descriptive identification element brings together the core elements that constitute a good basic description of an archival unit. This ensures that "the same data elements and structure are available at every level of description within the EAD hierarchy" (Society of American Archivists, 2002)

<dsc>
The description of subordinate components element allows for the description of component parts in the form of hierarchical groupings (typically: subgroups, series, subseries, files, and items).

Although the available units of information or elements are clearly delineated in EAD, there is a complicated relationship between EAD and its various companion standards (DACS and ISAD(G)). While EAD defines the structure of finding aids, DACS and ISAD(G) define what content should be included in a finding aid.

One key issue is the lack of equivalency in terms of what data elements should be present, and are considered most important, for archival description. In EAD only a handful of elements are required for a finding aid to be validated against the specifications of the DTD or schema. Instead, it is up to content standards and best practice guidelines to establish minimal elements for encoding. ISAD(G)v2 lists 26 specific data elements that "may be combined to constitute the description of an archival entity," (ISAD(G)v2, 2000) while DACSv2 (DACS, 2013) lists 25. ISAD(G)v2 lists six elements as essential for the exchange of archival information (reference code, title, creator, date(s), extent of the unit of description and level of description). DACS handles this issue by articulating a 'minimum,' 'optimum,' and 'added value' usage of elements.

EAD also provides for tagging elements that are not included or not differentiated in DACS or ISAD(G). For example, DACS makes the name and location of a repository a required element (and EAD provides a <repository> tag for this purpose), while ISAD(G) does not require this information (Dow, 2005).

EAD CHALLENGES

The process of putting EAD finding aids online is the result of two main activities: encoding (requiring knowledge of XML markup and text encoding software) and publishing (generally requiring knowledge of servers, style sheets and scripting) (Yakel & Kim, 2005). Research studies on EAD have generally focused on studying implementation issues, as well as studying subsequent user interaction with online finding aids and online finding aid Web sites (Wisser & Dean, 2013). However, one important research question has remained largely unanswered, namely, how the design of EAD and current encoding practices impact the quality, precision and usefulness of finding aid data throughout its lifecycle.

To date, only a handful of studies have examined the issue of the quality of EAD finding aid data and metadata, and the encoding practices by which finding aids are created. The impetus for these studies were wide ranging - including concerns about the degree to which archivists follow descriptive standards, about the need to provide a sense of uniformity to users of finding aids, about concerns for the retrievability of EAD data in non-EAD environments, and about discovery and semantic interoperability across EAD digital repositories.

Drawing from a sample of finding aids from 57 institutions, Prom looked at "whether EAD finding aids are consistently structured so that they might be searched alongside descriptions of cultural heritage materials that are drawn from other metadata formats" (2002, pp. 52-53). In doing so, encoding patterns were studied in order to uncover potential machine handling problems that might arise from the scope, structure, and consistency of EAD. Frost (2002) examined nine EAD encoding guidelines and established places where there were divergent and convergent approaches to encoding archival descriptive data. Carpenter and Park (2009) looked at metadata use and quality via an in-depth study of the frequency, completeness, and consistency of the wrapper element <eadheader>, and its associated sub-elements. This study was carried out using a random sample of finding aids from six digital archival repositories, coupled with an analysis of three best practice guidelines for EAD. In the three studies, the authors found that localized modifications of EAD had the potential to impede interoperability across repositories.

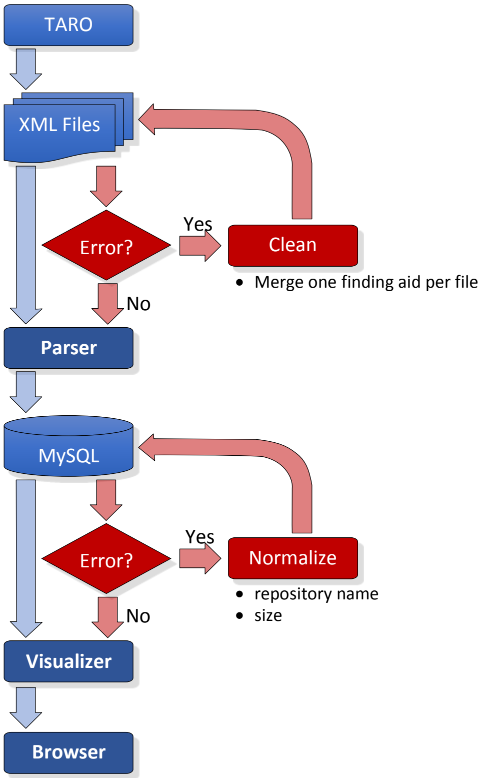


Figure 2. Ideal workflow (blue) and actual workflow (red)

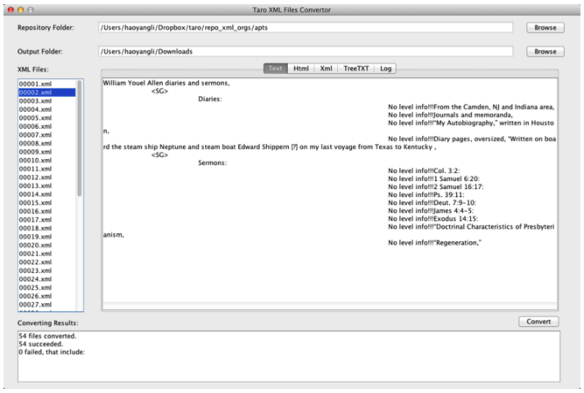
More recently, two studies have examined large corpora of EAD documents. Wisser and Dean (2013) analyzed a sample of 1,136 finding aids from 108 repositories with the goal of analyzing what elements and attributes were being used throughout the EAD structure and in what ways. The authors found substantial variability in element and attribute usage both within and across repositories, with variability more prevalent within the <dsc> section of the finding aid (Wisser & Dean, 2013). Bron, Proffitt, and Washburn (2013) examined over 129,000 finding aids from OCLC Research's ArchiveGrid, in order to determine how well select EAD tags supported aspects of online discovery (search, browse, results display, sort, limit by). In order to do this, the researchers established the levels of usage necessary for each element to be useful for discovery, and then determined whether the elements met the optimal threshold. Overall, the authors found that "the picture for archival discovery and EAD is decidedly mixed" (Bron, Proffitt, & Washburn, 2013).

QUALITY OF ARCHIVAL DATA

In order to analyze the quality of archival data for aggregation and visualization purposes, the project secured all the XML files from the Texas Archival Repository Online (TARO) as they were on February 28, 2013. In total, the data included 8729 finding aids from 35 repositories in the TARO database.

The project used a custom-made analytic tool named VADA (Visualizing Archival DAta system) that parses and visualizes aggregations of finding aids in order to help humans analyze potentially interesting trends and patterns in the data. The work documented in this paper relied on one specific functionality of the VADA system, namely its ability to flag and visualize variations, anomalies, and issues associated with the finding aid data, as a crucial first step in the overall process of aggregating and visualizing archival data. This functionality is handled by two components of VADA, the parser and the problem report tool.

The parser reads all the XML files in the aggregation, separates them into their constituent parts and, when found, flags anomalies and ambiguities in the tags, or content within the tags. While one might think that parsing XML files should be simple and could follow a linear approach (shown in blue on the left in Figure 2), in actuality, parsing EAD finding aids is more complex, requiring an iterative approach (shown in red on the right in Figure 2). For example, TARO finding aids that had been split into multiple files had to be converted back into one file. This parsing process often necessitated a human-in-the-loop (an expert archivist) to review the source files manually and interpret them before an automatic approach could be devised. In order to facilitate this process, a user interface was built for the parser (see Figure 3). In addition, the problem report tool provides a visualization of the flagged anomalies and ambiguities (see Figure 4) to allow users to examine them.



Repository Folder: /Users/haoyangli/Dropbox/taro/repo_xml_orgs/apts

Output Folder: /Users/haoyangli/Downloads

XML Files:

```
00001.xml
00002.xml
00003.xml
00004.xml
00005.xml
00006.xml
00007.xml
00008.xml
00009.xml
00010.xml
00011.xml
00012.xml
00013.xml
00014.xml
00015.xml
00016.xml
00017.xml
00018.xml
00019.xml
00020.xml
00021.xml
00022.xml
00023.xml
00024.xml
00025.xml
00026.xml
00027.xml
```

```
William Youel Allen diaries and sermons,
            <SG>
                        Diaries:
                                                                      No level info!!!From the Camden, NJ and Indiana area,
                                                                      No level info!!!Journals and memoranda,
                                                                      No level info!!!"My Autobiography," written in Housto
n,
                                                                      No level info!!!Diary pages, oversized, "Written on boa
rd the steam ship Neptune and steam boat Edward Shippern [?] on my last voyage from Texas to Kentucky ,
            <SG>
                        Sermons:
                                                                      No level info!!!Col. 3:2:
                                                                      No level info!!!1 Samuel 6:20:
                                                                      No level info!!!2 Samuel 16:17:
                                                                      No level info!!!Ps. 39:11:
                                                                      No level info!!!Deut. 7:9-10:
                                                                      No level info!!!James 4:4-5:
                                                                      No level info!!!Exodus 14:15:
                                                                      No level info!!!"Doctrinal Characteristics of Presbyteri
anism,
                                                                      No level info!!!"Regeneration,"
```

Converting Results:

```
54 files converted.
54 succeeded.
0 failed, that include:
```

Figure 3. Parser's User Interface

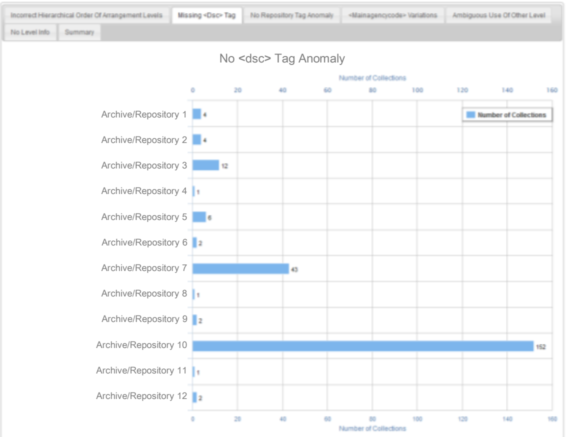


Figure 4. EAD Problem Report Tool

ID	Problem	Description	# Col.	# Rep.
1.	Missing <repository> tag	The <repository> tag was not supplied.	305	5
2.	Non-standard representation of data for physical size	Data content within the <physdesc> element is not normalized. The units and quantity for physical size vary greatly.	ALL	ALL
3.	Ambiguous use of "Other Level"	Sometimes otherlevel is not defined, and sometimes it is defined outside the standard arrangement levels.	146	10
4.	Missing <dsc> tag	The <dsc> tag was not present, making it hard to obtain the level information.	230	12

Table 1. Problems with EAD Standard

The problem report tool is a Web-based application. The aggregation and computation operations are performed on the server, and the output is visualized on a Web browser. The backend is built in Java and MySQL. The frontend is built in Javascript using libraries such as HighCharts JS and JavaScript InfoVis ToolKit.

QUALITY OF EAD AGGREGATIONS

The development of VADA allowed the research team to visualize and analyze the quality of the finding aid data in TARO, specifically the data that related to archival arrangement. This effort revealed a number of problems related to the design of EAD, the actual encoding practices for EAD finding aids, and the overall archival workflow.

As Figure 4 shows, VADA facilitates the visualization and analysis of the problems in the finding aids. VADA users can see additional information for each type of problem or go to the TARO page to inspect the original finding aid.

Overall, the problems identified by VADA can be grouped into two main categories (1) problems associated with the EAD standard, and (2) problems associated with the actual encoding practices.

Problems Associated with the EAD Standard

The first class of problems is those associated with the EAD standard itself. Creating EAD as a flexible data model in order to facilitate the encoding of legacy finding aids and to accommodate the significant variations in procedure across archives has resulted in a number of key issues that impede large-scale aggregation and analysis of data – permissiveness of the data model, and ambiguous or redundant encoding.

The first issue speaks to the fact that the EAD standard requires few elements to be present in order to create a valid EAD document, and that values and units can be encoded in multiple non-standard ways (e.g., two boxes vs. 2 ln.ft.). Problems 1-4 in Table 1 illustrates this issue.

The second issue refers to the fact that some information in EAD can be encoded in different places (e.g., information about who controls the collection can be found in multiple elements and attributes). Problem 1 in Table 1 illustrates this situation.

Initial attempts to aggregate and visualize TARO data at the repository level highlighted these problems. In EAD the <repository> element, and the MAINAGENCYCODE[4] and REPOSITORYCODE attributes can be used to identify unique archival institutions. However, this element and these attributes are not used consistently in the TARO data. E.g., while the repository element is required by DACS, the tag was missing in 305 collections across five repositories (in two repositories the tag was absent altogether).

Another place where these issues were made manifest was in attempts to aggregate and visualize data about the extent of collections in TARO. In EAD, <physdesc> (physical description) is a wrapper element for bundling information about the appearance or construction of the described materials. Although the data can be divided into sub-elements (<dimension>, <extent>, <genreform>, and <physfacet>), the information is also commonly presented as plain text (e.g., "9 boxes plus oversize and artifacts (5 linear feet)"). In the TARO data obtained in February 2013, there is wide variation in how this extent data is articulated (including linear feet, cubic feet, inches, boxes, items, and volumes) with the result that data normalization must occur before any meaningful analysis can be done.

A third example is that of the LEVEL attribute, which indicates the hierarchical level of the materials being described by the element. Ten TARO repositories utilized an OTHERLEVEL attribute, in place of the tags naming set levels in the arrangement (subgroup, series, subseries, file etc.). This makes the parsing process more difficult and makes it hard to compare arrangement practices across repositories.

Finally, some xml files did not include a <dsc> tag. The <dsc> is a required wrapper if individual components of a collection are going to be described. However, it was difficult to neatly extract from the data why the <dsc> was missing (e.g., the finding aid was simply a collection overview or there were no subordinate components because the collection consisted of only one item, etc.).

[4] MAINAGENCYCODE provides the ISO 15511 code for the institution that maintains the finding aid, while REPOSITORYCODE is a unique code in <unitid> that indicates the repository responsible for intellectual control of the materials being described.

The issues associated with the permissiveness of the data model and the encoding ambiguousness or redundancy, are not technically errors, since the EAD standard legally allows for them. Indeed, these characteristics of the design of EAD facilitate upstream activities in the archival workflow (e.g., the creation and encoding of finding aids, including legacy finding aids). However, these same characteristics are problematic for downstream activities in the archival workflow (e.g., aggregation, summarization of federated sets of finding aids).

Problems Associated with Encoding Practices

The second class of problems was those associated with the actual encoding practices within archival repositories (see Table 2). With humans in the mix, issues with the quality of the encoding can be expected.

One clear example is the variations of repository name in the <repository> element. In all but four of the TARO repositories, the repository name was inconsistent across multiple finding aids. The nature of these inconsistencies range from the completeness of the full name for a given repository to variations in the mailing address provided for the repository.

Related to the variations in the repository name, the problem of the varying mainagencycode attribute presents challenges for aggregating and visualizing finding aids across institutions. Problem 6 in Table 2 shows that the same mainagencycode can correspond to two different repository names. This might result from institutional name changes, or institutional merges. While VADA revealed this problem for only four mainagencycode variants, in one case there are 1325 records described under the institution's former name, while 3155 records are described under its current name. Problem 7 in Table 2 shows the converse situation, where finding aids from the same repository have different mainagencycode values. These variations range from typos and punctuation differences, to completely different codes.

While in most cases these variations refer to the same repository (even when the code is completely different), in one case, the repository had finding aids with mainagencycodes of other repositories. This last anomaly could be the result of a repository maintaining overall access to collections that are physically stored at other repositories.

Another problem with finding aid encoding relates to the nesting of components and sub-component levels. Initially, the parsing of the data was based on an assumption that the component level tags <c0X> followed an incremental order. However, the parser uncovered some nesting errors in the XML (e.g., <c03></c03> nested within <c03></c03>). Other finding aids, though they followed the incremental sequence of component level tags <c01><c02>…</c02></c01>, presented an unexpected nesting order of the archival arrangement levels: subgroup, series, subseries, file, and item.

VADA also revealed great variations in the use of different arrangement levels. Problem 9 in Table 2 illustrates this problem at the top level of the arrangement hierarchy.

ID	Problem	Description	# Col.	# Rep.
5.	Repository name variations	Within the same repository, its repository name was inconsistent across multiple files. Inspection of these reveals human error in keying in the repository's mailing address, including the expression of street names and ZIP codes. The data was normalized using the most frequent name, and only 4 repositories did not have this type of error.	NA	31
6.	One mainagencycode correlates to two repositories (including old and new names)	A single mainagencycode can correspond to two repository names. This anomaly results when a repository changes name.	NA	4
7.	One repository has finding aids with different mainagencycode	This is an anomaly where one repository has finding aids listing different mainagencycode. The variations range from small typos to completely different codes.	NA	13
8.	Incorrect hierarchical order of: subgroup, series, subseries, file, and item	For some collections, though they follow the incremental sequence of <c0X></c0X> in their XML file, there remains incorrect nesting order of the levels: subgroup, series, subseries, file, and item.	298	20
9.	Multiple instances of recordgroup, collection and fonds	There were multiple instances of *recordgroup*, *collection* or *fonds*. This is problematic, since the top level of the arrangement hierarchy should generally be a single node.	27	6
10.	No level information in the XML markup	No level information was provided in the XML markup from 31 repositories. E.g., there is a <c01>, but it does not specify the level)	6535	31

Table 2. Problems with Encoding Practices

While the top level of the hierarchy (encoded as *recordgrp*, *collection* or *fonds*) should generally be a single node, 27 collections in 6 repositories have more than one instance of this level in a finding aid. This is due either to the fact that multiple collections are described within one finding aid, or that the designation recordgrp/fonds/collection is used in lieu of more appropriate lower levels of the hierarchy (subgroup, series etc).

Problem 10 in Table 2 show that many finding aids have at least one component level (e.g., <c02>) without a defined type (e.g., level="series"). This is problematic because it is impossible to automatically infer the arrangement level associated with that component level (e.g., does <c02> refer to a "series" or a "file").

While these encoding problems are usually easily interpreted by humans, they create unnecessary complexity for automatic mechanisms aiming at parsing, aggregating and visualizing finding aids. The following section elaborates on the implications of these challenges, and suggests possible measures to alleviate these problems.

DISCUSSION AND CONCLUSION

While the application of visualization techniques is commonly used in areas such as data mining and data warehousing (Allen, 2005), this remains a nascent area for the archival profession. As archivists consider new ways of facilitating access, discovery, and exploration of collections of archival materials, traditional search oriented HTML-based Web form interfaces are gradually being joined by more advanced information visualization techniques that support additional ways of engaging with collections, such as browsing (Allen, 2005; Whitelaw, 2012).

Information visualization, in particular, has been seen as a way to support analytical reasoning for archivists (Lemieux, 2013). In this model, computational methods for extracting data are coupled with visualization techniques with the goal of aiding archivists in their analysis and decision-making activities during the archival curation workflow. This notion of visual analytics is seen as applicable to archival work such as appraisal (Xu et al., 2010), preservation and storage assessment (Xu et al., 2010; Xu et al., 2011; Esteva et al, 2011), and processing (arrangement and description) of collections (Lemieux, 2013). In addition, in cases where archivists are working with born-digital material, the metadata being visualized can also include structural and technical information directly mined from digital files and file directories (Xu et al., 2010; Xu et al., 2011; Heard & Marciano, 2011; Lemieux, 2013).

Particular value has been placed on visualization techniques that: facilitate the navigation of a document structure (whether hierarchical or multidimensional); illustrate networks of relationships; model work processes; and organize people, subjects, and events in time and space (Allen, 2005; Kramer-Smyth et al., 2007; Anderson, 2009; Xu et al., 2010; Esteva et al., 2011; Xu et al., 2011; Whitelaw, 2012). The underlying data used for such visualizations is largely drawn from the descriptive metadata extracted from EAD finding aids (Kramer-Smyth et al., 2007; Anderson, 2009).

The work of this research team is geared toward the use of visual analytics as a tool to study the archival practice of arrangement. However, this study has helped to establish that the quality of EAD finding aid data may result in the need for significant additional data cleanup before any such large-scale data aggregation and analysis is possible. Many of the problems identified in this work are the result of a mismatch between the needs and practices of the people working upstream in the archival workflow (e.g., arranging collections and encoding finding aids), and the people working downstream (e.g., aggregating and visualizing large sets of finding aids). Specifically, EAD seems to be biased towards the creation and encoding of finding aids, and away from the needs of those that subsequently wish to use them.

This work highlights the situated nature of standards for metadata description. As this work reveals, EAD allows great flexibility in the encoding of finding aids. This is a positive factor for encoding legacy data, and accommodating the practices of multiple different archival repositories. However, the resulting encoding is unnecessarily hard to process by automatic mechanisms. Visual analytics tools often have to iteratively parse and process the data in order to account for anomalies and variations.

Finding and documenting such problems with the EAD encoding is a key first step in instituting more rigorous control over descriptive and encoding practices that facilitate the aggregation, visualization and analysis of archival data. Such control requires tackling the issues of data completeness, accuracy, ambiguity, and consistency (semantic consistency in terms of data value and structural consistency in terms of data format (Carpenter & Park, 2009)). Specific recommendations for achieving this level of control include:

- Harmonize descriptive traditions and use standard encoding guidelines that facilitate cross-repository agreement.
- Require the use of core elements and attributes across all finding aids and repositories.
- Make the job of the human-in-the-loop easier through encoding resources such as software tools and data entry forms.
- Remove conditions that permit the encoding of the same data in multiple places, as this makes it difficult to aggregate and allows for data inconsistencies. Create a stricter data model that constrains the encoding of certain elements and attributes to specific parts of the EAD structure.
- Have finding aids define explicitly the correct hierarchical order of the component levels, and enforce that order.

- Enforce explicit tags for all values (e.g., size and dates), requiring that all values are written following a standardized format and explicitly specify the units (e.g., "ln.ft.").
- Mandate that all general values for a collection are explicitly tagged (e.g., date of accession, date of processing). Unknown values could be tagged as ‘unknown’ but should not be omitted.
- Enforce strict validation of constant/official values (e.g., mainagencycodes, repository name).

As TARO and other consortial sites look to the future and seek to improve search features and usability, and the archival profession works on revising standards such as EAD, archivists should also engage in a larger dialogue about the purposes and values of both metadata standards and the archival practices that are captured within them. By focusing on past practices, and the practices of here and now, the potential of EAD will remain unrealized and archivists may miss the opportunity to create robust encoding practices and tools that allow us to aggregate, visualize and analyze our finding aid data in new and transformative ways.

ACKNOWLEDGMENTS

This research was partially funded by a University of Texas at Austin, School of Information, Alumni Fellowship. The authors wish to thank Donna Coates, Jennifer Hecker and the ASIST reviewers for their constructive feedback on the paper.

REFERENCES

Allen, R.B. (2005). Using information visualization to support access to archival records. *Journal of Archival Organization* 3, 1, 37-49.

Anderson, I. G. (2009). From ZigZag to BigBag: Seeing the wood and the trees. In CEUR-WS, *Proceedings of the 1st Workshop on New Forms of Xanalogical Storage and Function, held as part of the ACM Hypertext 2009* (Vol. 508, pp. 12-17). CEUR-WS. org.

Bron, M., Proffitt, M. & Washburn, B. (2013). Thresholds for discovery: EAD tag analysis in ArchiveGrid, and implications for discovery systems. *Code4lib Journal* 22.

Carpenter, B., & Park, J.R. (2009). Encoded Archival Description (EAD) metadata scheme: An analysis of use of the EAD headers. *Journal of Library Metadata* 9, 1-2, 134-152.

Combs, M., & OCLC Research. (2010). Over, under, around, and through: Getting around barriers to EAD implementation. Dublin, Ohio: OCLC Research.

Dow, E.H. (2005). *Creating EAD-Compatible Finding Guides on Paper*. Lanham, Md.: Scarecrow Press.

Esteva, M., Xu, W., Jain, S.D., Lee, J.L., & Martin, W.K. (2011). Assessing the preservation condition of large and heterogeneous electronic records collections with visualization. *International Journal of Digital Curation* 6, 1, 45-57.

Frost, H.C. (2002). Guidelines counseling: A comparative analysis and evaluation of EAD implementation guidelines. *Journal of Archival Organization* 1, 3 73-86.

Gracy, K.F. & Lambert. F. (2014). Who’s ready to surf the next wave? A study of perceived challenges to implementing new and revised standards for archival description. *The American Archivist* 77, 1, 96-132.

Heard, J.R., & Marciano, R.J. (2011). A system for scalable visualization of geographic archival records. In *LDAV 2011 Symposium on Large-Scale Data Analysis and Visualization* (pp. 121-122). IEEE.

Hensen, S.L. (1983). *Archives, Personal Papers, and Manuscripts: A Cataloging Manual for Archival Repositories, Historical Societies, and Manuscript Libraries*. Washington D.C.: Manuscript Division, Library of Congress.

Hensen, S.L. (1997). “NISTF II” and EAD: The evolution of archival description. *American Archivist* 60, 3, 284-296.

International Council on Archives, & Committee on Descriptive Standards. (2000). *ISAD(G): General International Standard Archival Description*. Ottawa: International Council on Archives.

Kramer-Smyth, J., Nishigaki, M., & Anglade, T. (2007). ArchivesZ: Visualizing archival collections. Retrieved April 14, 2014 from http://archivesz.com/ArchivesZ.pdf

Lemieux, V.L. (2013). Visual analytics, cognition and archival arrangement and description: studying archivists' cognitive tasks to leverage visual thinking for a sustainable archival future. *Archival Science*, 1-25.

MacNeil, H. (1995). Metadata strategies and archival description: Comparing apples to oranges. *Archivaria* 39, 22-32.

Pitti, D.V. (1999). Encoded archival description: An introduction and overview. *D-Lib Magazine* 5, 11, 61-69.

Prom, C.J. (2002). Does EAD play well with other metadata standards? Searching and retrieving EAD using the OAI protocols. *Journal of Archival Organization* 1, 3, 51-72.

Ruth, J.E. (2001). The development and structure of the Encoded Archival Description (EAD) document type definition. In D. V. Pitti and W. M. Duff (Eds), *Encoded Archival Description on the Internet*. New York: The Haworth Information Press, 27-59.

Shaw. E.J. (2001). Rethinking EAD: Balancing flexibility and interoperability. *New Review of Information Networking* 7, 1, 117-131.

Society of American Archivists. (2002). *Encoded Archival Description: Tag Library*. Chicago: The Society of American Archivists.

Society of American Archivists (2013). *Describing Archives: A Content Standard*. Chicago: The Society of American Archivists.

Trace, C.B. & Dillon, A. (2012). The evolution of the finding aid in the United States: From physical to digital document genre. *Archival Science* 12, 4, 501-519.

Trace, C.B. & Francisco-Revilla, L. (2014). The value and complexity of collection arrangement for evidentiary work. Forthcoming in the *Journal of the Association for Information Science and Technology*.

Whitelaw, M. (2012). Towards generous interfaces for archival collections. In *Proceedings of International Council on Archives Congress*. Retrieved from http://ica2012.ica.org/files/pdf/Full%20papers%20upload/ica12Final00423.pdf

Wisser, K. & Dean. J. (2013). EAD tag usage: Community analysis of the use of Encoded Archival Description elements. *American Archivist* 76, 2, 542-566.

Yakel, B. & Kim, J. (2005). Adoption and diffusion of EAD. *Journal of the American Society for Information Science and Technology* 56, 13, 1427-1437.

Xu, W., Esteva, M., & Dott, S. J. (2010). Visualization for archival appraisal of large digital collections. In *Archiving Conference* (Vol. 2010, No. 1, pp. 157-162). Society for Imaging Science and Technology.

Xu, W., Esteva, M., Jain, S.D., & Jain, V. (2011, October). Analysis of large digital collections with interactive visualization. In *IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 241-250). IEEE.