EAD Analysis: Findings from the Building a National Finding Aid Network Project

Bruce Washburn
Principal Software Engineer

Merrilee Proffitt
Senior Manager

Chela Scott Weber
Senior Program Officer

May 2023

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 978-1-55653-290-0
DOI: 10.25333/atn7-qq32
OCLC Control Number: 1376802195

ORCID iDs
Bruce Washburn http://orcid.org/0000-0003-4396-7345
Merrilee Proffitt https://orcid.org/0000-0002-2322-8337
Chela Scott Weber https://orcid.org/0000-0002-6358-5128

This report represents work done on behalf of OCLC Research. When this report was written, Bruce Washburn was Principal Software Engineer at OCLC.

Please direct correspondence to:
OCLC Research
oclcresearch@oclc.org

Suggested citation:
Washburn, Bruce, Merrilee Proffitt, and Chela Scott Weber. 2023. *EAD Analysis: Findings from the Building a National Finding Aid Network Project*. Dublin, OH: OCLC Research. https://doi.org/10.25333/atn7-qq32.

This project was made possible in part by the Institute of Museum and Library Services, through grant LG-246349-OLS-20. The IMLS is the primary source of federal support for the nation's libraries and museums. To learn more, visit www.imls.gov.

The views, findings, conclusions or recommendations expressed in this project do not necessarily represent those of the Institute of Museum and Library Services.

CONTENTS

FIGURES

TABLES

INTRODUCTION

OCLC Research assembled an aggregation of Encoded Archival Description (EAD) finding aids for research purposes, and then analyzed EAD tag and attribute value usage patterns to address specific research questions. The analysis sought to uncover the promise in current EAD data as raw material for building a finding aid aggregation, looking for common elements and data structures already present in the data, as well as probing for gaps that could impede user discovery and use of archival collections.

OCLC Research conducted this research as one of the partners collaborating on the Building a National Archival Finding Aid Network (NAFAN) project alongside the University of Virginia, Shift Collective, Chain Bridge Group, and project lead California Digital Library. In 2020, IMLS awarded the California Digital Library (CDL) a National Leadership Grant to support Building a National Finding Aid Network (NAFAN), a two-year research and demonstration project to build the foundation for a national archival finding aid network to address the inconsistency and inequity of the current archival discovery landscape (LG-246349-OLS-20).

Work on the grant was done in parallel across multiple activities:

- Research investigating both end user and contributor needs in relation to finding aid aggregations
- Evaluation of the quality of existing finding aid data
- Technical assessments of potential systems to support network functions and formulating system requirements for a minimum viable product instantiation of the network
- Community building, sustainability planning, and governance modeling to support subsequent phases moving from a project to a program

OCLC Research has lead research in the first two areas of activity. This paper is focused on the first area, evaluating the quality of existing EAD finding aid data.

EAD encoded finding aids comprise the majority of data in current regional aggregators in the US.[1] These regional aggregators are likely to play a key role in contribution workflows to a national aggregation platform, and much of the EAD currently represented in these systems will likely flow into the NAFAN system, forming a metadata foundation for the platform to leverage. While the work represented in this report focuses on EAD, it is important to acknowledge that EAD is a small portion of what will ultimately be represented in NAFAN.

This project started with two research questions to inform our understanding of finding aid data quality:

- What are common data structures, and what elements occur consistently across finding aid data in current aggregations?
- Can the examined finding aid data support the needs identified in the user research phase of the NAFAN study? If so, how? If not, what are the gaps?

We used the same data set to answer other research questions related to user needs that were informed by the NAFAN pop-up survey.[2] Those questions were:

- Do EAD finding aids link to digital content?
- What is the completeness and consistency of the description of collections' physical characteristics and genre?
- Are content element values associated with controlled vocabularies, or can they be?
- Is institutional contact information in EAD finding aids consistent and reliable?
- How do EAD finding aids inform users about access to, use of, and reuse of materials in the described collections?

This report details the methods and findings from OCLC Research's quantitative analysis on a corpus of EAD encoded collection descriptions provided by current regional finding aid aggregator partners. This analysis of existing data quality in EAD encoded documents can help to scope the functionality that can be supported by a network made up of today's EAD finding aid data, as well as establish what is necessary for data remediation to support expanded network features. An analysis and discussion of the findings follow, including a series of recommendations for the NAFAN platform. These findings can help lay a foundation for building a nationwide aggregation that includes EAD finding aids as well as other forms of archival description.

A note to readers

This paper delves deeply into EAD elements and attributes and assumes at least a passing knowledge of the encoding standard. For those wishing to learn more about the definitions and structure, we recommend the official EAD website or the less official but highly readable and helpful EADiva site.[3]

Methodology

Data sources and ingest methods

Twelve regional aggregators of EAD finding aids participated in the NAFAN project and made their finding aids available to our OCLC team for quantitative analysis[4]:

- Archival Resources in Wisconsin
- Archives West
- Arizona Archives Online (AAO)

- Black Metropolis Research Consortium (BMRC)
- Chicago Collections Consortium
- Connecticut's Archives Online (CAO)
- Empire Archival Discovery Cooperative (EmpireADC)
- Online Archives of California (OAC)
- Philadelphia Area Archival Research Portal (PAARP)
- Rhode Island Archives and Manuscript Online (RIAMCO)
- Texas Archival Resources Online (TARO)
- Virginia Heritage

The resulting data set is composed of 145,673 EAD XML files, collectively representing 741 repositories. All of these EAD documents used the EAD 2002 DTD or Schema; no documents were provided that utilized the EAD3 Schema. Thirty-four files that could not be parsed as XML were excluded from the dataset. Though a few of the aggregator partners provided much of the content (see figure 1), the aggregation provides a useful mix from a wide variety of US locales and institution types.

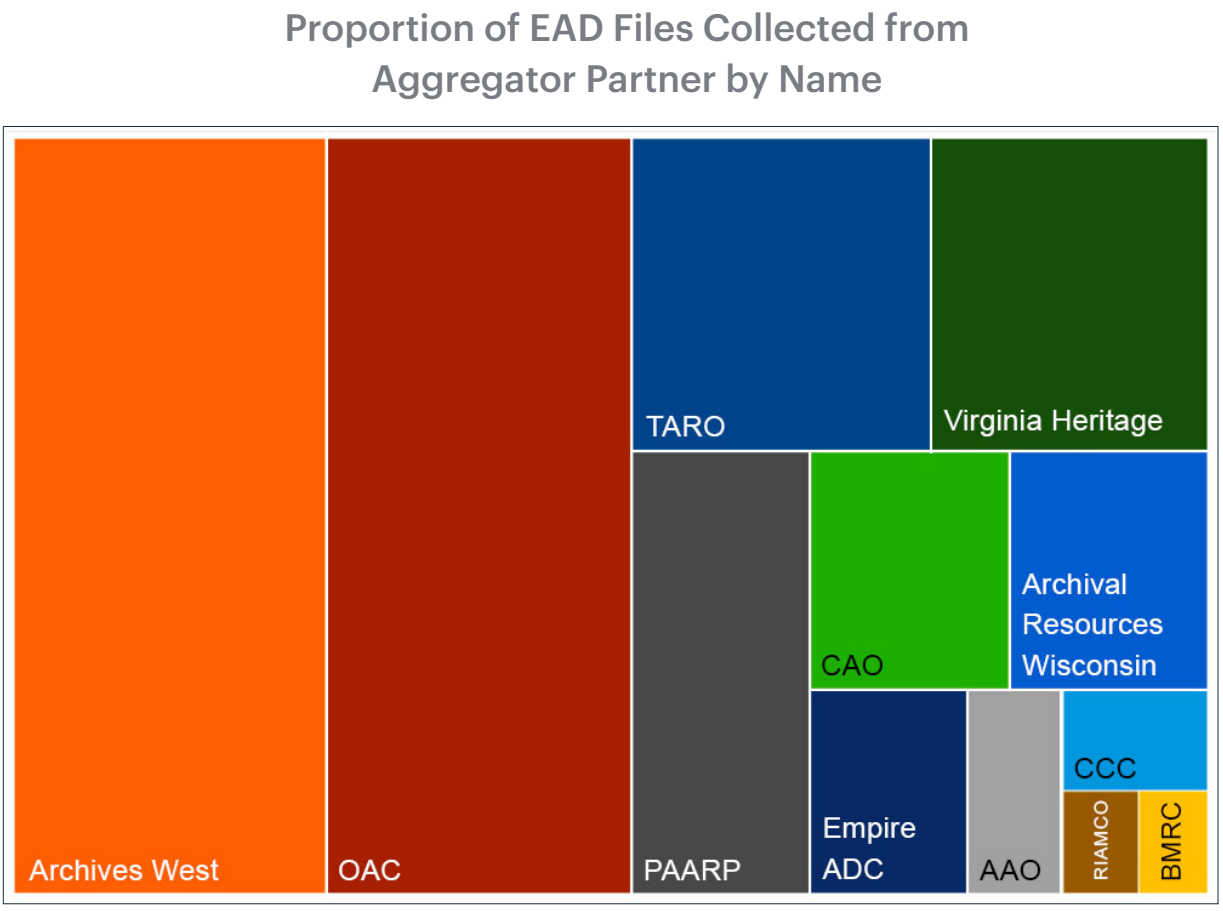

FIGURE 1. Proportion of EAD files collected from aggregator partner by name.

The mechanisms used to gather the EAD documents from the project participants varied. For some, their web server's file system was accessed by a web crawler to harvest their finding aid documents. A few made their EAD files available in the GitHub software version control system and accessible via a web crawler. Other institutions created a compressed archive file of all their finding aids and made them available for downloading.

Data set limitations

The extent to which the 12 NAFAN aggregator partners applied validation and data normalization to the EAD files made available for these studies varied. The different validation and normalization patterns summarized in the synopsis below demonstrate that variation. Thus, the data set may not accurately represent descriptive practices of individual institutions that contributed to aggregator partners. Rather, the data set should be understood to represent US regional aggregator data.

The regional aggregator partners support a mix of finding aids encoded in the XML Schema Definition (XSD) and Document Type Definition (DTD) versions of EAD. Once ingested into individual aggregator's systems, they may or may not preserve the DTD and/or XML Schema references if they are not needed within the context of their system. The EAD files shared with OCLC for this project may reflect that post-processing state.

The aggregator partners apply a variety of validation and normalization processes, which range from merely assuring that EAD documents are valid, to manual editing of EAD tags, to transforming files to meet aggregator system requirements. The variety of approaches to data validation and normalization could explain some of the differences we see in both the broad analysis of EAD tag use and in the additional focused analyses of content in elements and attributes. For example, it may impact the extent to which genreform element use varies across all aggregator partners and across repositories for a single aggregator, as described in the section on completeness and consistency in genre and physical characteristics section.

The source datasets were obtained during the period of November 2020 to January 2021 and were not refreshed, so this data set represents a snapshot in time. The finding aids represent a snapshot in another sense, as many may not have been updated by the originating institution in several years, and either may not reflect current best practices at the institution or may include outdated contact points or other information. We did not have file update dates for all of the NAFAN project participants' data. When examining data supplied by one contributing aggregator partner where files were created or updated between 2004 and 2020, two-thirds of files updated throughout that time period had last been updated since 2017 or earlier.

The gaps between when the finding aids were retrieved by OCLC for analysis and when they were last updated by their provider have implications for some of the data analysis projects described here, and potentially for others using these datasets for additional analyses. For the study on how EAD finding aids link to digital content, broken or unresponsive links may be over-reported if the related external resources have changed their locations and if the current versions of the finding aids are using the correct links. For the general EAD tag analysis reporting, the examination of genre and physical characteristics study, as well as our investigation of controlled vocabulary study, any additions of authority file numbers and sources or of EAD tags for content element strings made to the finding aids after retrieval would not be represented in the data. This gap could lead these studies to undercount the degree to which the finding aids support semantic searching and reconciliation to external vocabularies.

Tools and methods

The tools for retrieving, selecting, and analyzing EAD data elements and attributes included:

- Project-specific applications written in the Python programming language to open, read, and write to data files, and for data filtering and sorting
- The Python Requests software library was used for retrieving NAFAN participant finding aids, as well as to test links and return content linked to in finding aids for analysis
- XPath (XML Path Language) query language queries for selecting nodes from XML documents
- R software for statistical computing and graphics
- OpenRefine for cleaning, analyzing, and reconciling data
- Microsoft Excel for summarizing data and generating visualizations

Further details about analysis methods and how these tools were applied in the research studies are described in each subsection of the Analysis and Findings section.

Note regarding EAD element path patterns

The EAD standard[5] and best practice guidelines for its implementation[6] encourage consistency in the structure of finding aid data. At the same time, EAD allows for flexibility, which leads to variation across repositories and aggregators. This variation presents challenges when designing XPath queries to find relevant elements and attributes as in this research project. Queries written to analyze this corpus of EAD data attempted to take all potential nesting patterns and paths into account that might be used with a particular element.

Analysis and Findings

Dimensions for data quality analysis

The first research question informing this report is, "What are common data structures, and what elements occur consistently across finding aid data in current aggregations?" This question of data quality—understanding the structure and consistency across EAD finding aid data—is of primary importance for the NAFAN project, helping to establish a baseline for what EAD flowing into NAFAN would look like.

In undertaking this work, we followed a framework defined in 2013 for an investigation that approached EAD tag and attribute usage from a discovery perspective. That framework identified five high-level features that often are present in archival discovery systems.[7]

- **Search:** all discovery systems have a keyword search function; many also include the ability to search by a particular field or element.
- **Browse:** many discovery systems include the ability to browse finding aids by title, subject, dates, or other facets.
- **Results display:** once a user has done a search, the results display will return portions of the finding aid to help with further evaluation.

- **Sort:** once a user has done a search, they may have the option to reorder the results.
- **Facet:** once a user has done a search, they may have the option to narrow the results to only include results that fall within certain facets.

Analysis was then done to identify the EAD elements and attributes that, if present, could be accessed, indexed, and displayed to facilitate these high-level discovery features. Those EAD elements and attributes are:

- Dates: unitdate
- Extent data: extent
- Collection title sources: unittitle, titleproper/@type=filing
- Content tags in dsc: corpname, famname, function, genreform, geogname, name, occupation, persname, subject
- Content tags in origination: corpname, famname, name, persname
- Content tags in controlaccess: corpname, famname, function, geogname, name, occupation, persname, subject
- Material type: controlaccess, genreform
- Repository: repository
- Notes: abstract, bioghist, scopecontent

For example, dates could potentially be utilized as search terms, or leveraged for browsing or sorting. They may also be important for disambiguating similarly named collections in displays. Similarly, material types, represented by form and genre terms, could be important for narrowing a large result using a facet.

Having established these key elements and attributes necessary to drive a discovery apparatus, we sought to understand how often these key elements and attributes are used. To better characterize our usage findings, we describe them using threshold levels developed in the 2013 study. The threshold levels are:

- Low level (0-50%)
- Medium (51-80%)
- High (81-95%)
- Complete (96-100%)

Although we used these levels as a reference point, we recognized that usage as a proxy for discovery is an artificial construct. It is difficult to predefine thresholds for the level of usage of an element at which it becomes more or less useful for discovery. An element that is used 95% of the time does not become unusable if it is only used 94% of the time. The absence of an element does not directly lead to a breakdown in a discovery system. It is more like a gradual decay of its effectiveness. Though the NAFAN EAD aggregation used in 2021 is a different corpus of data provided by different contributing institutions at a different time, the results of the 2013 and 2021 data analysis projects are similar: some important elements are at a high or complete threshold, but many elements needed for discovery interfaces are at medium or low use (see table 1).

TABLE 1. EAD tag usage in 2021.

Discovery category	EAD element	Percentage of use	Threshold
Dates	unitdate	81.89	High
Extent data	extent	80.05	High
Collection title sources	unittitle	99.98	Complete
	titleproper @type=filing	00.08	Low
Content tags in dsc	corpname	01.56	Low
	famname	00.09	Low
	function	00.03	Low
	genreform	02.36	Low
	geogname	00.97	Low
	name	00.07	Low
	occupation	00.05	Low
	persname	03.53	Low
	subject	00.75	Low
Origination and content tags	origination	85.30	High
	corpname	25.25	Low
	famname	01.62	Low
	name	00.25	Low
	persname	54.34	Medium
Archdesc/controlaccess and content tags	controlaccess	85.05	High
	corpname	41.84	Low
	famname	06.09	Low
	function	00.66	Low
	geogname	34.78	Low
	name	00.04	Low
	occupation	06.97	Low
	persname	43.98	Low
	subject	75.41	Medium
Material type	genreform	38.75	Low
Repository	repository	99.68	Complete
Notes	scopecontent	91.51	High
	abstract	84.63	High
	bioghist	73.32	Medium

A minimum viable descriptive record

One of the goals of the NAFAN project is to define a subset of metadata fields that will be required of all records to be added to the system, or a NAFAN minimum viable descriptive record. Describing Archives: A Content Standard (DACS) is the widely adopted, Society of American Archivists-endorsed content standard for archival description in the United States,[8] which specifies required and suggested content for descriptive records to be compliant with the standard. The DACS single-level optimum record is a commonly used level of description across many archives and is therefore a good proxy for investigating what might be required by NAFAN for its minimal viable descriptive record.

Investigating the completeness of the fields included in a DACS single-level optimum record within the NAFAN data set can tell us how much the extant archival description is in alignment with this standard. It also can be taken as an indicator of common practice, helping to understand if requiring these fields might align with existing practice in archives or would mean asking many archives to change their practice.

METHODOLOGY

DACS is an output-neutral content standard that gives guidance on descriptive content and does not include encoding specifications for that description. In order to identify the subset of EAD elements that map to the required elements for a DACS single-level optimum record, we used the DACS to EAD to MARC crosswalk in the appendices of the DACS standard.[9]

Once the elements were identified, XPath queries were used to find the total occurrences of an element in the query results (some elements can occur more than once per document), unique occurrences, and the number of documents in which the element occurs at least once. The percentage of use was calculated by dividing unique occurrences by the total number of EAD documents in the NAFAN aggregated data set (145,639). A potential limitation of this approach is that the method only counted the occurrence of an element and did not discard empty elements. Empty elements can occur when EAD authoring is done by filling in a form or a template, so some elements may be overcounted.

The same thresholds defined in the Discovery analysis in the previous section are used to characterize the findings for this analysis.

FINDINGS

All but one of the elements included in a DACS single-level optimum record had complete or high use in our data set (see table 2).

The element that only attained a medium usage threshold was bioghist, used for the Biographic or Historical Note for a collection, which was used 72% of the time. In the discovery environment, data in this field is of most use for keyword searching. As it is an unstructured note, it cannot be used to build browse, facet, or sort features and so its lower usage would have limited impact on system functionality.

Though it still falls in the high usage category, unitdate was the next lowest usage of the elements analyzed. Date is of high interest to users, and the data in this field could be used to build browse, facet, sort, and display functionality, so the lower usage rate would have an impact on these functions.

TABLE 2. Usage of elements required in a DACS single-level optimum record.

Element	Total occurrences	Unique occurrences	Percentage of use	Threshold
unitid	144,381	141,590	97.22	Complete
repository	145,195	145,178	99.68	Complete
unittitle	147,040	145,615	99.98	Complete
unitdate	131,303	119,265	81.89	High
physdesc	162,172	144,902	99.49	Complete
origination	139,911	124,237	85.30	High
scopecontent	1,503,483	133,273	91.51	Complete
accessrestrict	256,182	132,406	90.91	Complete
langmaterial	148,787	143,297	98.39	Complete
bioghist	125,723	105,324	72.32	Medium
controlaccess	516,006	123,858	85.04	High

Links to digital content

In a survey conducted by OCLC Research and presented to users of archival aggregator systems in 2021,[10] nearly half of respondents (42.7%) indicated that they preferred online materials but were willing to use in-person materials, while 14.4% of respondents stated a strong preference for online materials only.

A possible avenue for presenting online material in a finding aid aggregation would be to provide a filter or search option that would limit results to those that include online content, or perhaps signal in a result set which items have online content associated with them. Some EAD attributes can be used to associate an element with an external resource as a link, and this can serve as an indicator that there is associated online material.

Investigating how EAD finding aids in the research aggregation link to digital content could help answer several questions, including:

- What is the average number of external links per finding aid?
- What EAD elements and attributes are most frequently used for external links?
- What types of digital objects are linked?
- How many relative URLs are present that rely on the finding aid to be accessed within its local context?
- What percentage of external links still resolve?

METHODOLOGY

To begin the analysis, the EAD attributes associated with links were extracted from all EAD finding aids provided by the aggregator partners; links were extracted from wherever they were present in the XML documents. Extraction was done via Python script with XPath queries.[11] Resulting values were gathered and analyzed in an Excel workbook, with a worksheet for each aggregator partner. Python scripts were then used to verify the link quality and the media types of the external resources. More than 600,000 unique external links were present in the NAFAN EAD dataset. A 25% sample (every fourth link in each aggregator worksheet) of those links was tested.

FINDINGS

The average number of external links per finding aid is about 5

In this study, an "external link" is one that includes the "http" or "https" protocol in the URL, in contrast to a "relative" that needs additional data or processing to be completed. For example, a relative link may include just the file name of the document ("ABC074_050.html") or a portion of the file system path and the file name ("../graphics/Box 1/A_GEN_VARI_001.jpg"). By this definition, on average there are 5.35 external links per finding aid across the NAFAN EADs, though that varies across aggregator partners where the average ranges from 0 to 9.

The same external link may be present in many finding aids from a NAFAN project participant, such as a link to the institution website, logos, or other shared resources. After deduplicating the external links for each NAFAN participant, the overall average number of links per finding aid drops to 4.13. The number of links that are referenced by more than one finding aid for each NAFAN participant varies considerably, with some having many and others having relatively few. See figure 2 for a comparison of the average number of external links for each of the NAFAN partner aggregations and the average number of links after deduplication.

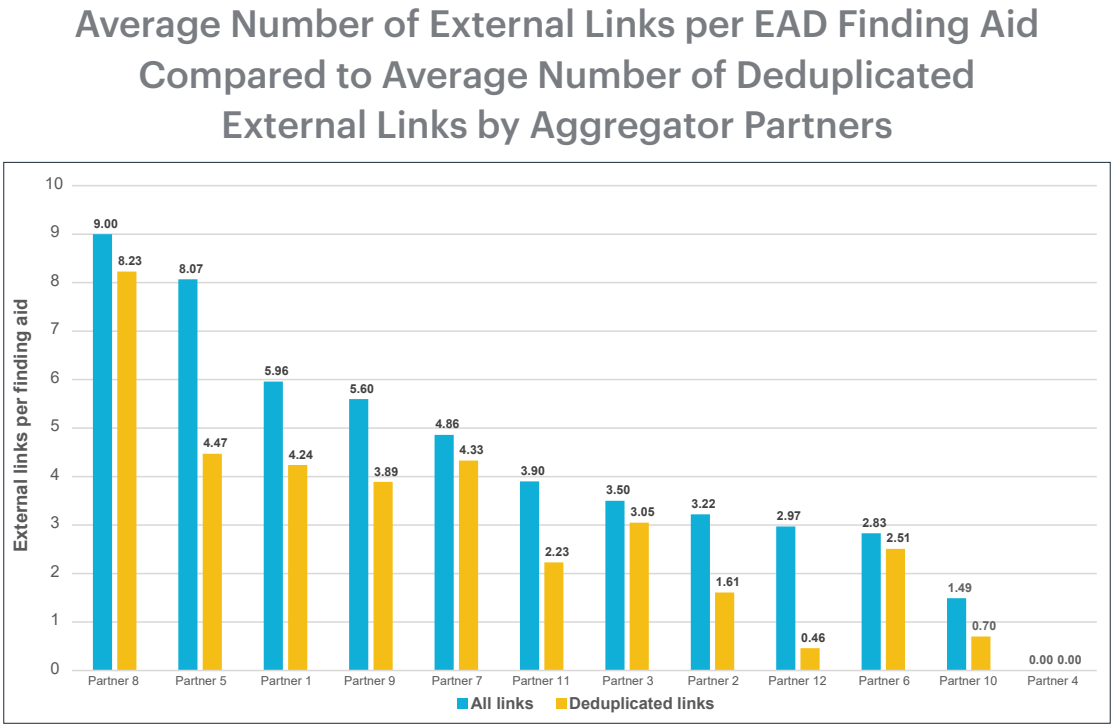

FIGURE 2. Average number of external links in each EAD finding aid compared to the average number of deduplicated external links by aggregator partners.

Relative URLs may impact EAD document syndication for some aggregator partners

For this study, entity references were included in the category of relative links, as they are codes that rely on a separately maintained dataset of full URL references to be completed. These datasets may not be available in all cases to other systems that are utilizing the EAD XML data. Relative URLs that had an audience attribute of "internal" were not counted, assuming that their purpose is local and not expected to be supported if the EAD is viewed in other systems.

The frequency of relative URLs and entity referencing is worthy of attention from the perspective of syndication and re-use of the EAD XML document. Since relative links can be expected to break when the EAD files are on a different web server, how much of the linking utility would potentially be lost when these files are moved onto a new file server?

The data analysis suggests that the use of relative links may present only a minor issue for most of the aggregator partners, especially after considering the deduplicated relative URL references. Across all partners, only two utilize a large number of relative links. The practice of recording relative URLs in finding aids should be revisited to maximize their linking potential when the EAD XML data is re-used in other web applications.

EAD external links are generally of higher quality than what other studies have reported

The decay of link quality is a common and pervasive issue on the web, and the issue tends to correlate with the amount of time that has passed since the document was last updated. For example, a 2021 study of external links in 1996-2019 New York Times articles found that only 75% of links were viable, while 25% of links were inaccessible.[12]

As noted in the section on EAD data sources and ingest methods for these projects, some documents in the NAFAN EAD dataset may not have been updated in several years, posing a question of whether a similar effect on link quality would be observed when testing their external links.

The majority of aggregator partners showed link decay well below that found in the 2021 New York Times study. Figure 3 shows that for nine of the 11 aggregator partners with links to digital content, 85%-95% of those links were viable. Overall, link quality is better than what might have been expected given what has been reported for web link quality in general, and the link quality is excellent for most aggregator partners.

The majority of aggregator partners showed link decay well below that found in the 2021 New York Times study.

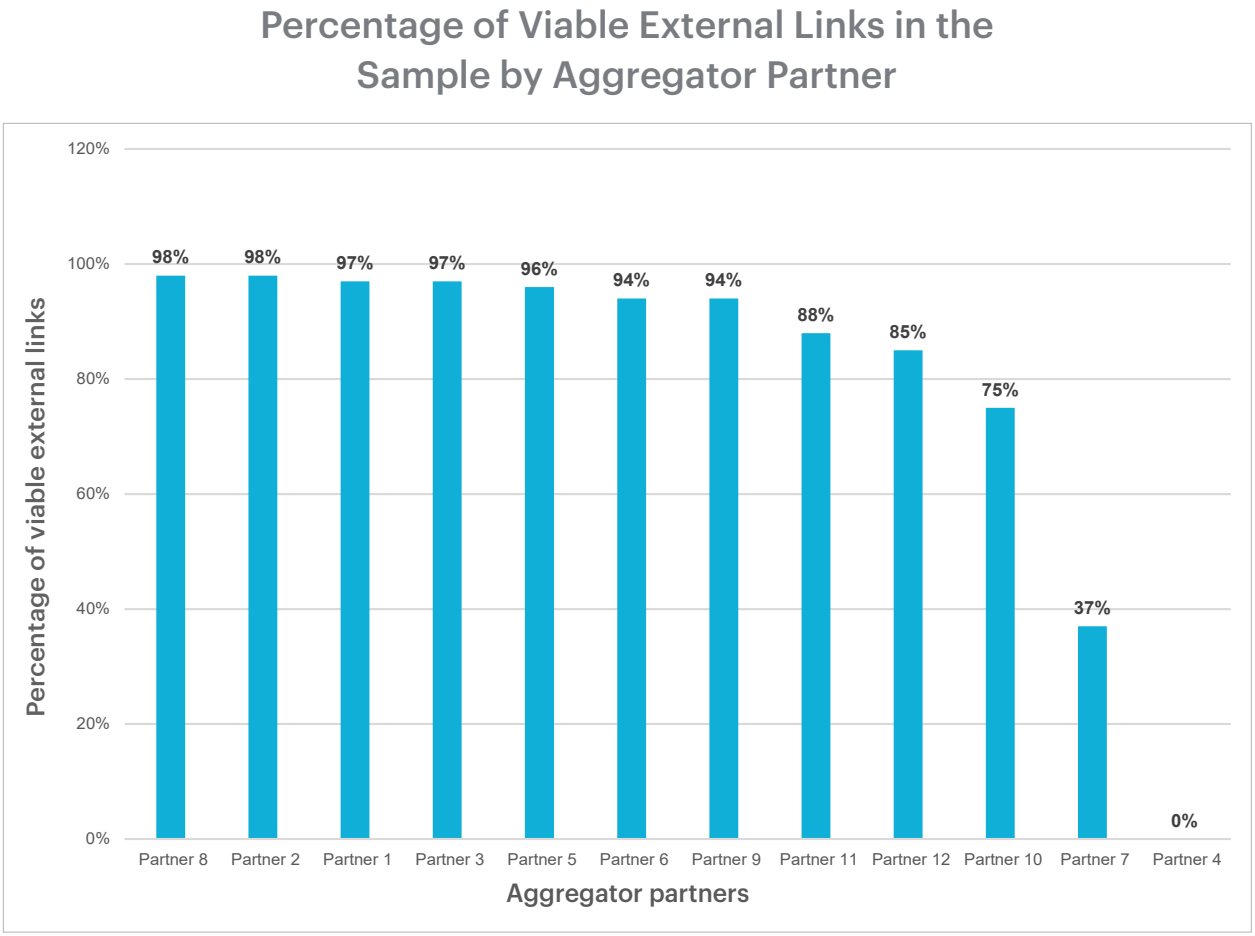

FIGURE 3. Percentage of viable links in the sample by aggregator partner.

It may be that the use of persistent identifiers and URLs in library and archive datasets is a more consistently followed practice than in general practice, and therefore has avoided the levels of decay in link quality over time that is found elsewhere on the web. For example, one NAFAN contributor used persistent URLs from at least five different URL providers for around half of their external links.

Statistics for external resource media types overstate the presence of HTML documents

Our study examined, categorized, and counted the media type (a two-part identifier for file formats and format contents) of the external resource so that we could better understand and characterize the types of media that are linked from finding aids.

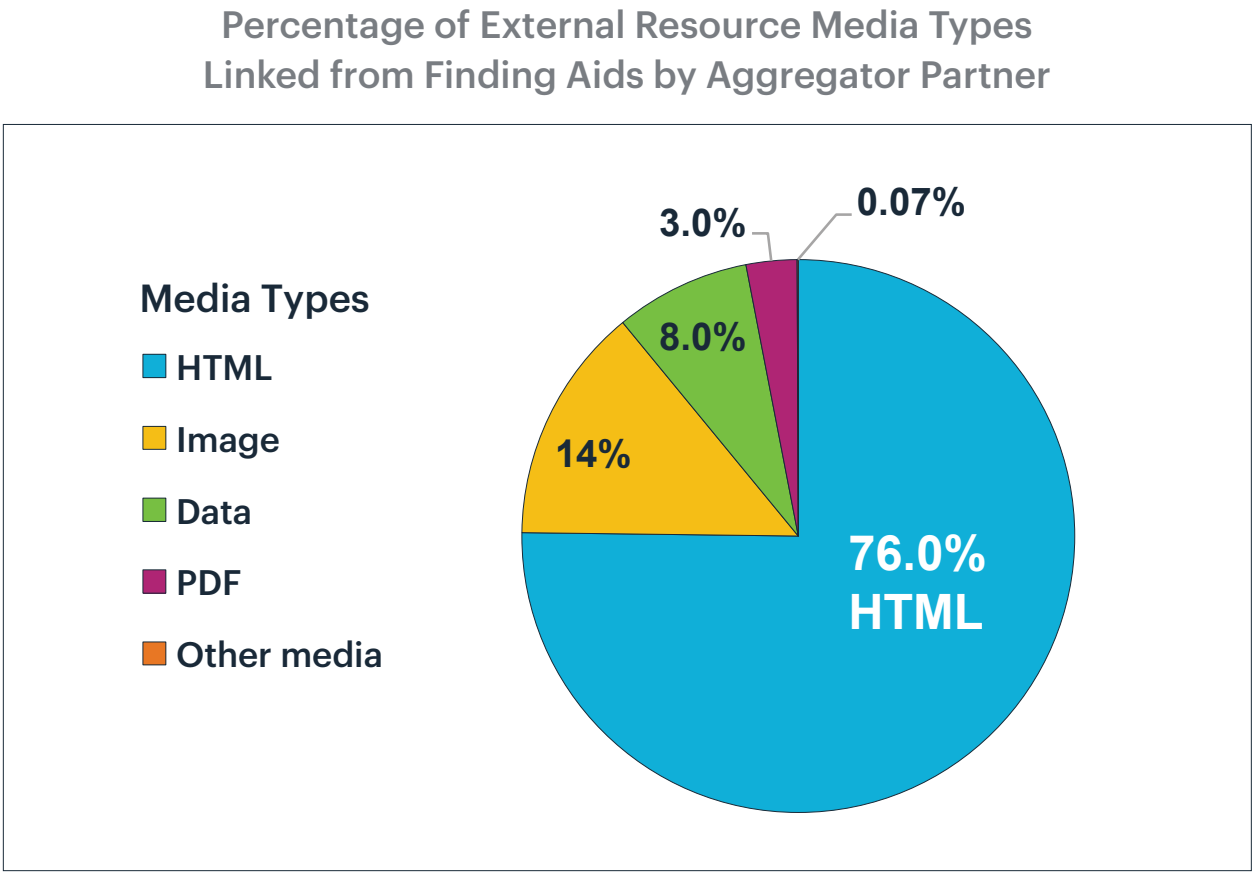

FIGURE 4. Percentage of external resource media types linked from finding aids by aggregator partner.

As shown in figure 4, 76% of files linked to from finding aids are HTML resources. However, for many HTML resources, the linked web page may be used to provide a framework for presenting core content that may be an image, streaming media, or an embedded PDF document. An in-depth review of the HTML resources would provide a fuller picture of linked media resources.

The different media types associated with the high-level groupings, and their associated counts, are listed in table 3 below. For the non-HTML resources (image, PDF, data, and other media), over half are images. Just over 30% are data; as with the HTML resources, it would take additional review to determine exactly what these resources represent but one can guess that they are an attempt to represent linked data. A little more than 11% are PDF documents—again, it would take resources to analyze these files to determine their nature and purpose.

TABLE 3. Media types and associate counts.

High level grouping	Media type	Count
HTML	---	105,791
Extent data	JPEG	18,791
	GIF	580
	PNG	46
PDF	---	3,884
Data	Excel	8
	JSON	9,055
	OpenXML	11
	Text/Plain	2
	Turtle	1,525
	XML	13
	ZIP	5
Other	AAC	1
	Octet-stream[13]	43
	MP4	10
	MPEG	39
	Quicktime	3

Completeness and consistency in genre and physical characteristics

A survey conducted by OCLC Research and presented to users of archival aggregation systems in 2021[14] revealed an interest in a broad range of materials, including some materials that may not be expected to be commonly included in the types of archival collections described by finding aids: periodicals, newspapers, and other published content. This prompted us to examine what EAD documents can tell us about the form and genre of the materials in their associated collections. This information could logically be recorded in the genreform element, so we focused investigation there.

Our tag usage investigation shows how widely the genreform element for Genre or Physical Characteristics is used within archdesc/controlaccess, where controlled access terms are expected to be representative of all or most of the collection. The analysis indicates that at least one genreform term is present in this upper level of controlaccess for about 38% of the NAFAN finding aids.

METHODOLOGY

We used a Python script with XPath queries to extract the genreform elements and their attributes where the element appeared in the finding aid within archdesc/controlaccess and above the dsc element to isolate terms that would apply to the collection as a whole. Terms that occurred in 10 or more finding aids were uploaded into the OpenRefine application for additional normalization,

sorting, and analysis. This threshold for inclusion was set to create a more manageable OpenRefine dataset, given the long tail of variant terms. Excel was used to generate visualizations of OpenRefine analysis results.

FINDINGS

Use of the genreform element varies widely across aggregator partners and repositories

The presence of a genreform term varies considerably across the range of NAFAN participant aggregations, with some making little use of this element directly in archdesc/controlaccess while several participants include it in 50% or more of their finding aids, as depicted in figure 5 below.

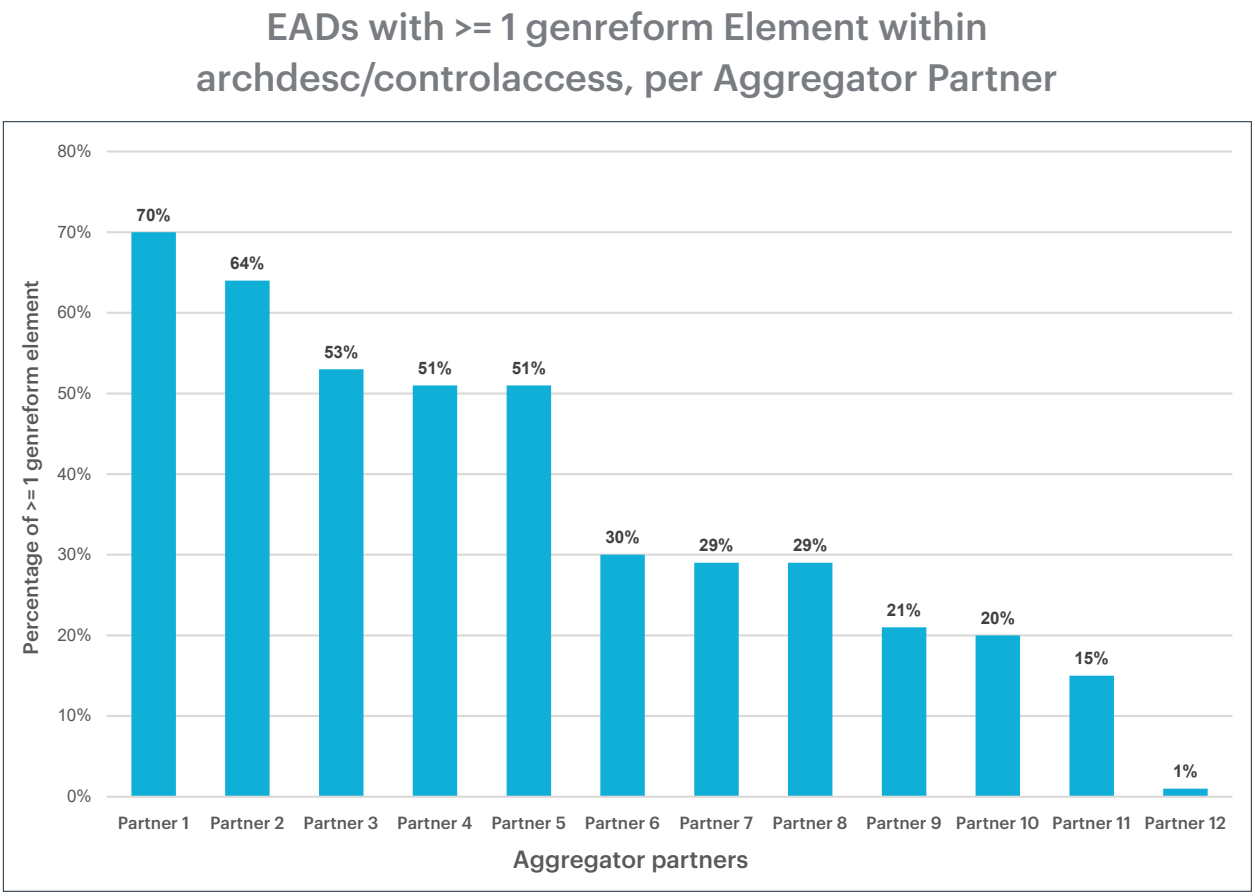

FIGURE 5. EADs with >= 1 genreform element within archdesc/controlaccess, per aggregator partner.

The variability of the presence of the genreform element is common within aggregations as well. In figure 6 below, the 54 repositories for a single aggregator partner are shown with the percentage of their EAD finding aids that include at least one genreform term, and the coverage ranges from 0–100%.

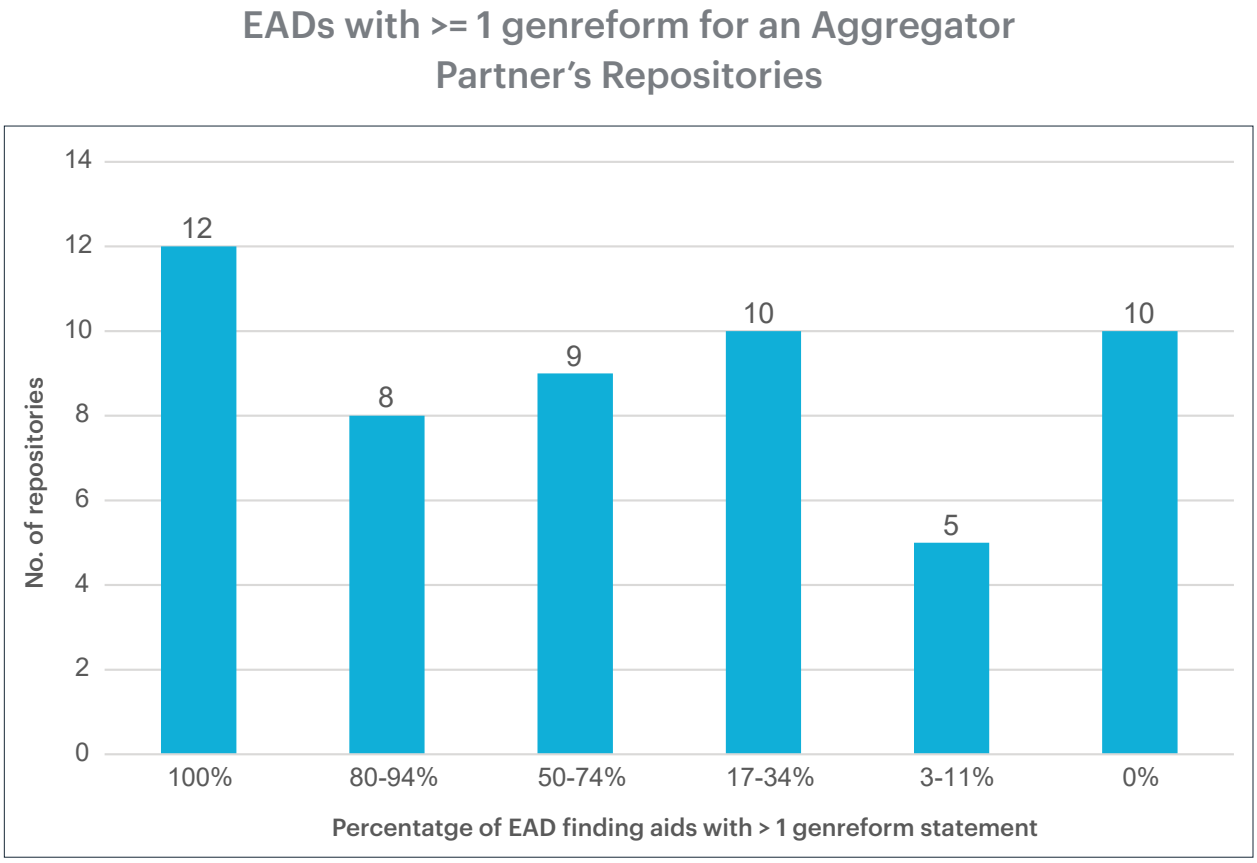

FIGURE 6. EADs with >= 1 genreform for an aggregator partner's repositories.

Genreform element value consistency is impacted by the use of many different controlled vocabulary sources

Analyzing the values and attributes found in the genreform element tells us something about the physical characteristics of this cross-institutional corpus of collection descriptions. After applying some string normalizations to cluster together typographically variant values for the same form or genre term, a tree map visualization (see figure 7) of the top 20 most frequently occurring terms suggest the types of materials at a very high level. Even with this limited number of frequently occurring terms, there is evidence present in the visualization of differing source vocabularies. For example, "manuscripts" from the Library of Congress Subject Headings in green appears on the bottom right corner of the chart and "manuscripts for publication" from the Getty Research Institute's Art and Architecture Thesaurus is just to the left in blue.

Treemap of the Top 20 Genreform Terms in the NAFAN EAD Corpus

FIGURE 7. Treemap of the top 20 genreform terms in the NAFAN EAD corpus.

The tree map includes the genreform term "Fieldwork project," which would not be expected to find its way into the top 20 terms across the aggregation given the specialized nature of this type of document. "Fieldwork project" appears in nearly all of the 3,700 finding aids for a single repository, but it is not used elsewhere across the NAFAN EAD dataset. This suggests a lot of variation in description practices. With more genreform terms from more finding aids, this type of over-representation of a specialized term in a NAFAN-wide view would be less common.

The source attribute for the genreform element indicates the controlled vocabulary in which the term can be found. When present, the source attribute could be used to determine if the vocabulary is local or shared, helping to be more selective when building visualizations or discovery mechanisms with consistency of source and practice in mind.

After normalizing and clustering the genreform element values, there were over 5,300 distinct terms, which seems high for an element that might be expected to draw from a more limited range of values. By comparison, there are about 3,500 Form and Genre headings in OCLC's FAST subject vocabulary. A wide range of controlled vocabularies, including local ones, were used as sources that would expand the set of distinct terms. And the variety of sources and terms could also pose challenges for providing consistent and predictable discovery across the aggregation by physical characteristics or genre.

Though the number of distinct values is high, their occurrence counts reflect a very long tail, with 84% of the distinct terms used in fewer than 10 different EAD documents. There are 228 distinct terms that were found in 100 or more different EAD documents; of these, two terms are used over 10,000 times, 21 terms are used between 1,004 and 3,600 times, and 203 terms are used between 100 and 982 times.

Only two attributes for the genreform element were widely used

Attributes for the genreform element had mostly limited use, though encoding and source were relatively common. The encodinganalog attribute value was typically "655" or an alternate representation of that string. This is not surprising since the MARC 655 field is used in bibliographic records to indicate the class, form, genre, and/or physical characteristics of the materials being described. There were over 60 different values in the source attribute, though some variation of either "aat," indicating the Getty Art and Architecture Thesaurus, or "lcsh," indicating the Library of Congress Subject Headings, were by far the most common.

When combined with the less frequently occurring authfilensumber attribute, the source and file number could be used to generate a persistent URL for the source. Therefore, some additional consistency and cleanup of the source attribute values could generate further enrichment of the data. For example, if a source URL is provided or can be generated for a genre term and other finding aids use that same term but haven't specified a source or authfilenumber attribute, they could borrow that enrichment from one finding aid to share more widely across the corpus.

TABLE 4. Attributes used in genreform elements within archdesc/controlaccess.

Attribute	No. used	Percentage of use
source	150,236	94.00%
encodinganalog	96,733	61.00%
authfilenumber	12,124	08.00%
altrender	4,035	03.00%
normal	1,578	01.00%
rules	661	00.40%
type	33	00.02%
audience	0	00.00%
id	0	00.00%

Controlled vocabulary analysis

In interviews with both archivists and end users, it became clear that a key value that NAFAN can contribute is surfacing relationships between archival collections held at different institutions. These relationships could be identifying collections on the same subject, collocating the writings of a single person, or showing relationships across collections by identifying correspondents. Leveraging controlled vocabularies within the data is an obvious way to build out such functionality, and we wanted to better understand if extant EAD data was up to the task.

For names of people, families, organizations, subjects, places, and genre forms, the data quality EAD tag usage analysis describes how frequently these content tags are used. A closer look at the content tags' element and attribute values can reveal how much work has already been done to associate those values with identifiers for the entity in a controlled vocabulary and provide a testbed for evaluating the potential for using automated tools to attempt further reconciliation.

METHODOLOGY

A Python script used XPath queries to extract element and attribute values from the persname, famname, corpname, subject, geogname, and genreform elements. The extracted data were analyzed in OpenRefine to determine the utilization of the authfilenumber (a number that identifies the authority file record for an access term) and source attributes (indicating the controlled vocabulary source of the heading). OpenRefine also was used to cluster and normalize matching headings and to evaluate the potential for further reconciliation to external vocabularies.

FINDINGS

Inclusion of authority file numbers is infrequent, but identification of controlled vocabulary sources is more common

The tabulation of elements that included values for the authfilenumber and source attributes indicates that fewer than 10% of these elements provided an authfilenumber, but 40% or more of the elements included a source value (see table 5). For example, in the extraction of 1,092,209 persname elements, 94,365 (8.6%) included an authfilenumber attribute value while 595,048 (54.5%) included a source attribute value.

TABLE 5. Tabulation of authfilenumber and persname attributes.

Element	No. of elements extracted	Percentage with authfilenumber	Percentage with source
corpname	504,438	5.2%	46.4%
famname	21,993	4.6%	79.2%
genreform	482,955	4.5%	67.1%
geogname	392,387	3.0%	57.9%
persname	1,092,209	8.6%	54.5%
subject	844,872	4.9%	83.2%

Establishing the source can improve the efficiency and accuracy of reconciling headings with a controlled vocabulary. A small number of widely used controlled vocabularies predominate. For example, the Library of Congress Name Authority File (LC NAF) and Virtual International Authority File (VIAF) controlled vocabulary sources were the most commonly referenced shared vocabularies in the aggregated EAD finding aids for personal names, as shown in figure 8 below depicting the top 10 source values identified.

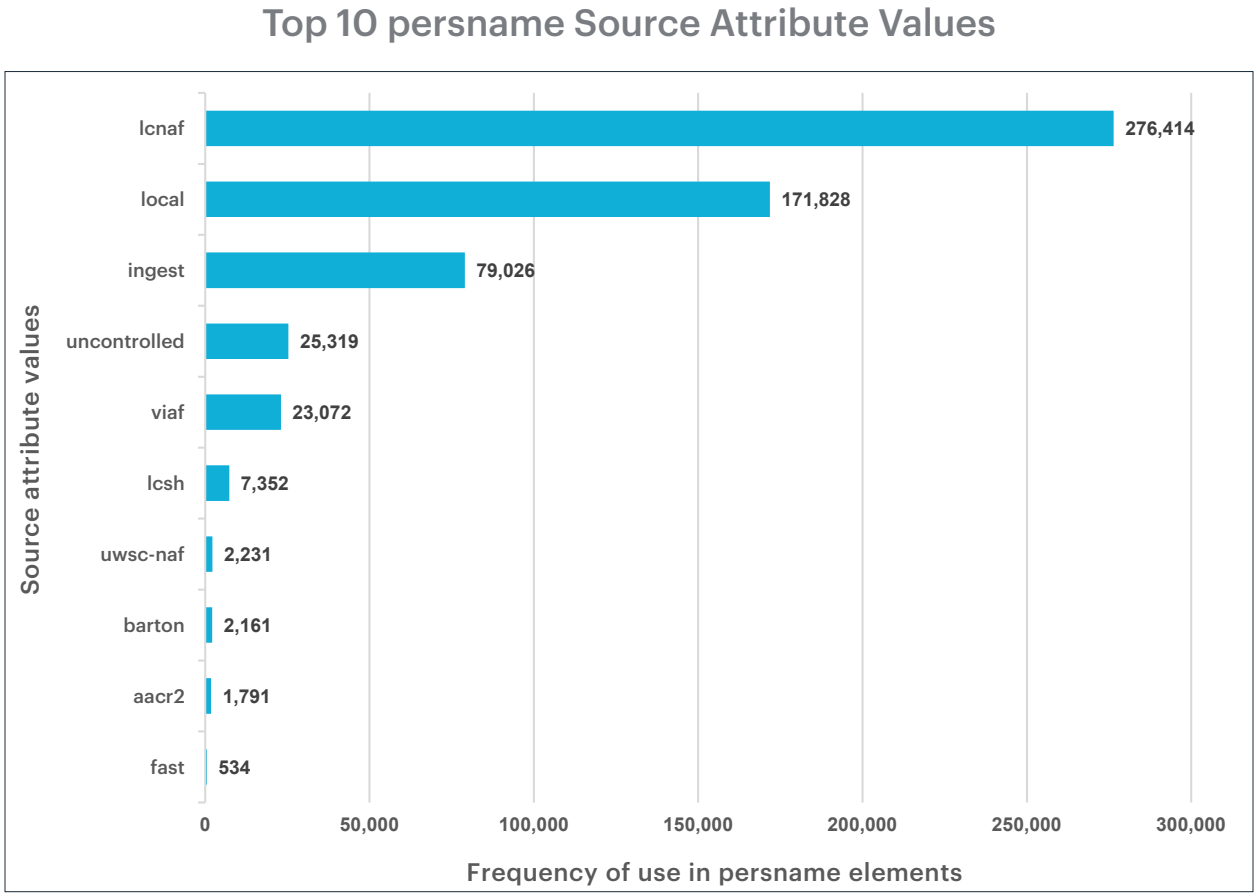

FIGURE 8. Top 10 persname source attribute values.

Clustering personal name elements could provide a path to enriching finding aids with controlled vocabulary identifiers

This study also focused on how identity can be established for personal names in EAD finding aids. The persname element describes people who owned or created the materials in collections and others who are noted in or related to the collection materials. It was expected that, in comparison with access terms for organizations, places, subjects, and genres, the people that finding aids refer to may not be as widely represented in shared controlled vocabularies from the library domain if they are not otherwise associated with published works as creators, contributors, or subjects. The Library of Congress EAD site provides guidance on the use of the persname element:

> All names in a finding aid do not have to be tagged. One option is to tag those names for which access other than basic, undifferentiated keyword retrieval is desired. Use of controlled vocabulary forms is recommended to facilitate access to names within and across finding aid systems.[15]

Authority control for personal names

The personal name investigation included extracting the 1,092,209 persname element values from the aggregation of NAFAN EAD documents. The persname element values were then normalized (converted to lower case, extraneous spaces and punctuation removed) and deduplicated into clusters of matching headings. To produce a more manageable dataset for cleanup and analysis, a cutoff point for the frequency of occurrence of the heading was applied to only include clusters in which the name occurred in five or more finding aids. This resulted in a dataset of 20,767 personal

name clusters, representing deduplicated and merged headings from 496,340 persname elements (45% of all the extracted persnames), and including their associated EAD attribute values. This dataset was further modified using the data cleanup tools available in OpenRefine. Those changes included resolving the variant forms of the LC NAF source attribute value (described below) to a single consistent form, converting LC NAF and VIAF identifier numbers in the authfilenumber attribute to a full URL, and deduplicating LC NAF and VIAF URLs.

As described in the section below on positive network effects that can be observed in an aggregation, the clustering of personal name elements amplifies the effect of some finding aids that used the source and authfilenumber attributes. In the OpenRefine project based on the 20,767 personal name clusters, 5,837 clusters (28%) included an authfilenumber for the LC NAF, VIAF, or both, while only 8.6% of the total extraction of persname elements included an authfilenumber attribute (see figure 9). Since the clusters are based on individual persname elements—some of which may not have had an authfilenumber provided in the source finding aid—the clustering process increased the potential availability of applicable authfilenumber values from 8% to 11%, without any additional reconciliation work.

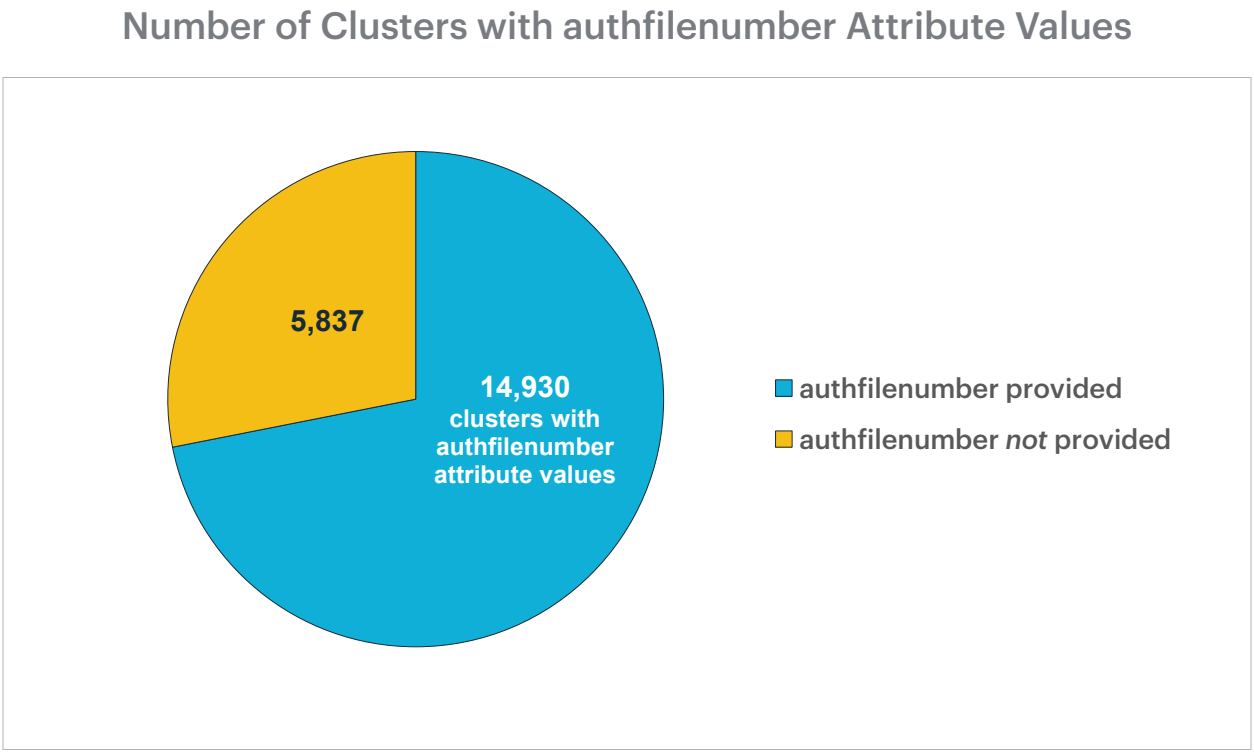

FIGURE 9. Number of clusters with authfilenumber attribute values.

Automated and manual reconciliation of personal names to a controlled vocabulary can further enrich the clustered elements

A key advantage of working with the OpenRefine tool is that, in addition to providing ways to clean up, transform, and sort data, it can connect to external controlled vocabulary systems and reconcile strings to matching authorized headings and their persistent identifiers. This OpenRefine reconciliation feature was used to look for matches in the LCNAF for the 11,439 persname cluster names that did not already have an LC NAF or VIAF authfilenumber attribute value in the cluster.

The OpenRefine reconciliation feature can be configured to point to a compatible "endpoint," which uses the OpenRefine Reconciliation API to convert requests into searches sent to the target-

controlled vocabulary's system, typically using that system's API or similar machine-readable data service. For this study, OCLC hosted an instance of a Library of Congress Reconciliation Service for OpenRefine, which is made available under an open-source BSD license in the GitHub software repository.[16] Its documentation provides more details on how it interacts with the LC Name Authority file and ranks its results.

The OpenRefine settings for reconciliation include an option for the system to automatically assign a match for any results that are returned from the endpoint with high confidence. With this setting, the first pass at reconciling the 11,439 persname clusters lacking an authfilenumber automatically matched 3,491 clusters to a LC NAF heading. The percentage of clusters with an authority file number increased after the automated reconciliation and matching from 28% to 44% (figure 10), and the total number of persname elements that have inherited or could inherit an authority file identifier from the cluster increased from 11% to 17%.

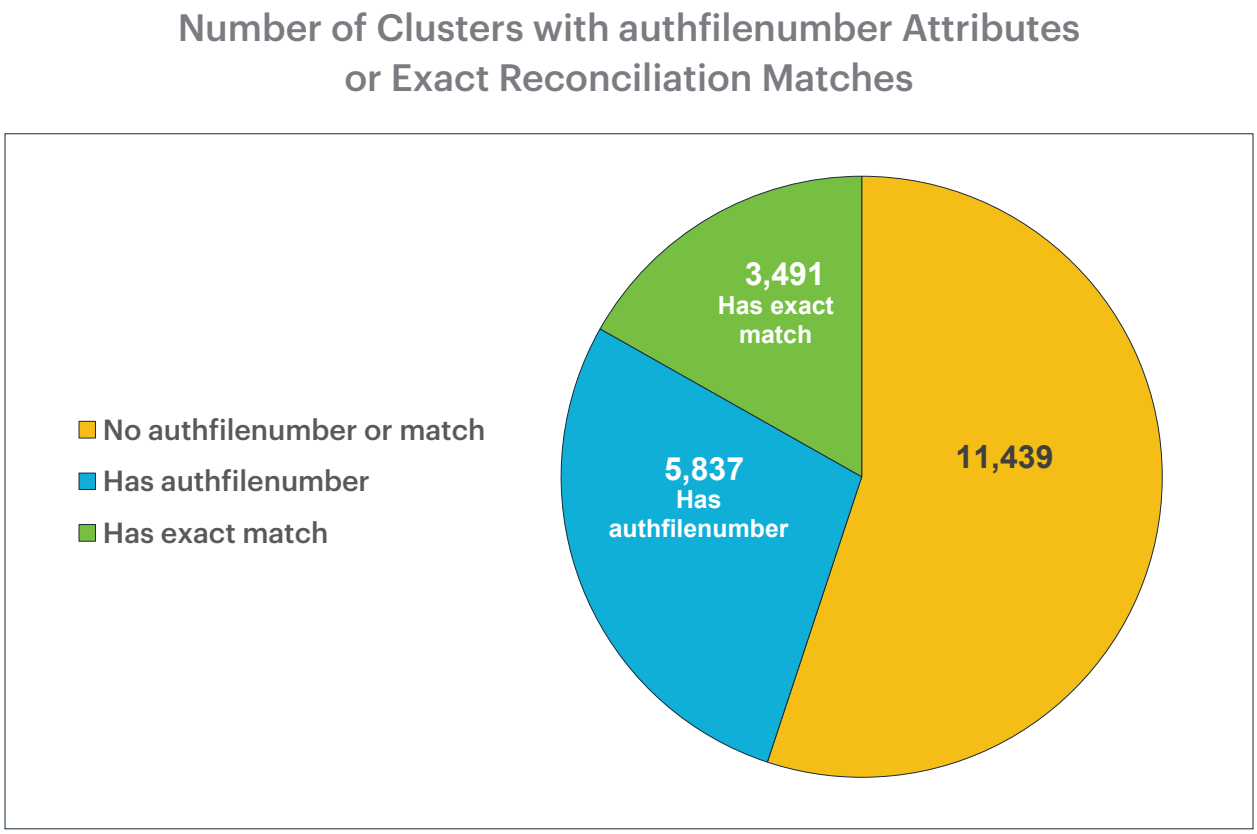

FIGURE 10. Number of clusters with authfilenumber attributes or exact reconciliation matches.

The real work of reconciliation is more painstaking and careful. The personal names that returned one or more potential matches from the LC NAF need to be evaluated—at times consulting their context within the original finding aid—to select or reject suggested matches from the authority file. This work requires diligence, time, and domain expertise. For this study, after clustering similar personal names, finding VIAF and LC NAF identifiers when available from one or more persname elements in the clusters, and looking for automated exact matches using the OpenRefine reconciliation service, there still were more than 11,000 persname clusters that would need manual reconciliation and review.

Researchers reviewed the top 500 persname clusters (ranked by the number of finding aids in which the personal name element was found) that lacked either an authfilenumber from the finding aid or an exact match from the first pass of the automated reconciliation process to evaluate the impact of manual reconciliation. Just 66 of those names were manually reconciled to corresponding LC NAF records, though matches were set only if there was very high confidence in the relationship without evaluating the name in its finding aid source to obtain more context. The tactic of working with clusters for names that appear many times in many finding aids meant that the 66 manual matches provided identifiers that had an outsized potential impact on finding aids. As the total number of persname elements that have inherited or could inherit from the reconciled cluster, an authority file identifier increased from 17% to 22%. See figure 11 below.

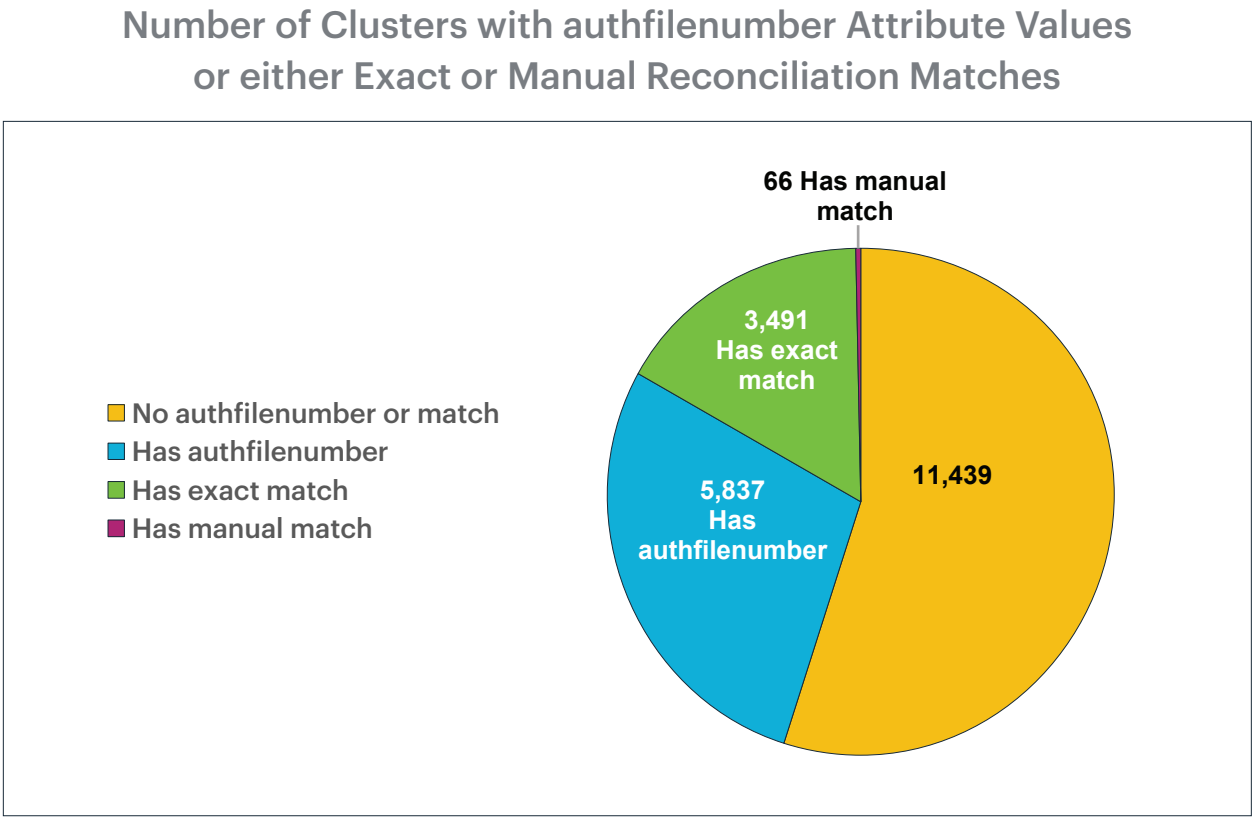

FIGURE 11. Number of clusters with authfilenumber attribute values or either exact or manual reconciliation matches.

Personal name elements in the aggregation represent a long tail and will require substantial resource commitments to establish their identity

This reconciliation study focused on a subset of the persname element values by working with clusters of names that occurred five or more times, ignoring 55% of the extracted values. There is a long tail of infrequently occurring personal names, some of which include too little data to support effective reconciliation (i.e., only providing a surname) and some representing people who are not likely to be found in authority files if they lack a type of literary warrant, not having been a creator of, contributor to, or subject of a published work. They may be accurately tagged as a personal name, but their authority and identity may either only be established locally or not at all. A source attribute value of "local" (or a similar designation) was found in 15.8% of the extracted persname elements.

Variations in source attribute values impede reconciliation

The source attribute is an optional method of identifying the controlled vocabulary source for an element value. When analyzing values in persname elements, 266 unique source attribute values were found, a surprisingly high number given the expected range of controlled vocabularies used for creating archival collection descriptions. But there can be multiple distinct representations of the same vocabulary. For example, the Library of Congress Name Authority File appears to be represented by these distinct source attribute values in the EAD finding aids evaluated for this study:

> lc, LC Name Authority File, lca, lcaf, lcanaf, lcanf, lccn, lchs, lcna, lcnaf, lcnaflocal, lcnag, LCNAH, lcnameauthorityfile, lcnat, lcnf, lcnnaf, lcsnaf, library of congress name authority file, library_of_congress_name_authority_file, Library_of_Congress_Name_Authority_File, lnaf, lncaf, lnnaf, naf.

Some of these variants may be the result of typographic data entry errors, while others may originate in the finding aid editing interfaces and conversion tools used to create the EAD.

After normalization and clustering of typographically different terms we found 119 unique source attribute values. However, only a handful of these values make up the make up the majority of the uses: there were over 270,000 occurrences of the term "lcnaf," over 160,000 occurances of the term "ingest," and over 70,000 occurrences of the term "viaf." The infrequently occurring sources may have important advantages for data management in a local context, but in cross-institution aggregation their functional benefits are less clear.

This level of variation can present a barrier to cross-document and cross-aggregation data analysis, as the source attribute value is important for determining what systems to use for reconciliation of headings to persistent identifiers. If a taxonomy of controlled vocabulary source values could be agreed upon and used across finding creation tools, the interoperability of this EAD attribute would improve.

There are potential network effects for name reconciliation in aggregated finding aids

The same person's name may be found within persname elements in finding aids from multiple repositories. While not all occurrences of that name will have been described with a source or an authfilenumber attribute, in some cases they may be. When many finding aid sources are brought together in a single aggregation by deduplicating and clustering persname values, there is the potential for enhancing all of the finding aids by inheriting the authfilenumber and source attributes from more completely described names, representing a positive network effect. For example, the personal name string "Obama, Barack" can be found in persname elements in 17 finding aids across the aggregation, but only a few occurrences make use of the authfilenumber attribute. By clustering these data, links to the LCNAF and VIAF can be derived and potentially applied to less fully described persname elements in other finding aids, avoiding duplicative or repetitive reconciliation of the access term by each repository.

A positive network effect created by aggregating multiple finding aid sources also can be seen when a personal name cluster is associated with more than one unique identifier from the same controlled vocabulary. For example, in the study of persname values, the cluster for the name "Hamilton, Alexander, 1757–1804" was associated with two LCNAF identifiers. One identifier was correct, the other was not.[17] Discrepancies like these can rise to the surface when multiple sources are aggregated, allowing for detection of the issue and potentially its correction.

The aggregation also can help to surface inconsistencies in the controlled vocabulary sources. In this persname study, the cluster for the personal name heading "Parker, Quanah, 1845?-1911" was found to be associated with two different VIAF clusters, which can be reported and likely merged.[18]

Reliability and consistency of repository contact information

The discovery system that the NAFAN project envisions will need to connect users with archive staff who can assist with on-site access and other research needs by supplying reliable contact information for institutions with holdings represented in the aggregation.

The NAFAN Technical Working Group is considering the administration and management requirements for contributing repository data, including identifying reliable sources for repository contact information. The repeatable EAD addressline element can be used within the address wrapper to indicate the location of a repository along with contact information in the form of phone numbers and email addresses. This investigation examines how widely the address information is used in the NAFAN EAD corpus, and how consistent and current the contact information is that it contains.

METHODOLOGY AND LIMITATIONS

This data analysis concentrated on the corpus of NAFAN EAD XML files to find unique address data values in these paths:

- /ead/eadheader/filedesc/publicationstmt/address
- /ead/archdesc/did/repository/address
- /ead/archdesc/did/repository/extref/address
- /ead/eadheader/filedesc/publicationstmt/xi:include/@href[19]

After extracting the data, unique address elements were evaluated in OpenRefine to determine if they held more granular and potentially useful details such as email addresses and website URLs.

There is an additional source of data, /ead/archdesc/did/unitid/@normal="repository code" which we overlooked in our initial analysis.

FINDINGS

EAD address element values could initialize a repository registry

The EAD address values can be a starting point for gathering data for a repository registry, since at least one address can be found for 93% of the repositories represented by the 12 NAFAN participants. Address element values include valuable information. Text mining of address values in OpenRefine indicates that the EAD XML also can supply phone or fax numbers (in 96% of addresses), email addresses (66%), and website URLs (56%).

To be useful as a repository registry for NAFAN, or to be more useful to end users, data remediation may be needed

This analysis revealed key ways that address data may need to be updated to be more actionable. Approximately 4% of addresses provide a PO Box instead of a street address. While still a viable contact point, these data do not lend themselves to automated processes for generating geolocation coordinates with building-level accuracy. Actionable geolocation data are increasingly important for supporting map visualizations, directions, and other features.

The analysis also found multiple addresses for the same repository. Variations between multiple unique addresses for the same repository were primarily typographic differences. This study did not normalize the address strings to eliminate these variations. When more than one address was found for the same repository, the most frequently occurring address may not be the most complete address. Other address variations may supply additional useful data elements, such as email addresses and website URLs.

Similarly, manual review would be needed to ensure the quality of the addresses, as contact elements may have changed and, in some cases, the differences can indicate data entry errors. For example, the difference may be two representations of the same phone number: "520-629-8699" vs. "520-629-8966."

The EAD documentation recommendation to use an entity reference for address information usually was not followed

The Library of Congress EAD documentation states that creators of finding aids should "consider using an entity reference to store address information that occurs in many finding aids, as it is easier to update the information when located in a single, shared file."[20] This approach rarely was present by NAFAN aggregator partners, as only two repositories in two different aggregator partners employed it. One aggregator made extensive use of a related technique, employing XML Inclusion elements to link to consistent address XML in an external XML file for each repository.

It may be that EAD XML normalizations applied prior to delivery of data to the NAFAN data analysis had used entity references or similar approaches but had then embedded the referenced data in the published EAD XML, making this practice appear to be relatively rare.

It's possible that the EAD is expected to be most frequently accessed in the context of the repository's website, which can disregard the EAD address values and external reference mechanisms and instead provide consistent contact information from a separately managed resource or registry (as is the case with the Online Archive of California and Archives West, two of the largest aggregator partners).

Externally referenced address information needs to be considered when EAD XML is syndicated and shared

When entity references or other mechanisms are used to improve the consistency of address data, applications that are consuming the shared EAD XML will need to take that into account and reflect any changes to the external referenced sources when refreshing the EAD XML corpus.

Use of ISO 15511 and the ISIL Code

Another possible approach for attaining address information is via /ead/archdesc/did/unitid@normal where the @repositorycode attribute can be used to record the ISO 15511 (International Identifier Standards for Libraries and Related Organizations) Code for a repository. We did not

assess this path in our initial data analysis. Due to time and resource limitations, we were unable to repeat our initial methodology but found that in our data set, close to 60% of the finding aids in our corpus follow this pattern, with over 1000 unique values represented in the repositorycode attribute. There are limitations to relying on the ISO codes for address information—the ISO codes are typically only associated with the name of a repository with a street address (or even a partial address) and the process for requesting a new code via a maintenance agency may be present a hurdle for some organizations.

Use, reuse, and access information

Our interviews with archival users for the NAFAN project indicate that they are interested in being able to easily find information about how to access archival collections, as well as terms of use and reuse for collections material.[21] Access restrictions are important when assessing what collections are immediately available to address a research question and in planning research trips to archives. Use restrictions are important when a user is interested in obtaining a reproduction for their own use in a publication or for making other uses of collection items (such as in digital scholarship).

Access and use restrictions vary by archival collection, and policies generally governing access and use vary by repository. In an aggregation environment, it will be especially important to communicate to users clearly and consistently what is known about these restrictions.

METHODOLOGY

A Python script used XPath queries to extract the element and attribute values associated with the EAD accessrestrict, userestrict, physloc, and phystech values, along with the EAD file names for their sources and the path to each element. OpenRefine was used to analyze the text from these fields, grouping and counting text that was repeated in multiple files. The resulting set of data was analyzed for content and categories were inductively developed based on that content. The categories were then applied to each grouping.

FINDINGS

Varying encoding practices presents challenges for machine interactions with the element values

Even when the accessrestrict field was used to indicate that there are no restrictions on access, the EAD data presented 40 distinct string patterns for categorizing those statements. The brief unstructured text found in these elements did not lend itself to out-of-the-box analysis using Natural Language Processing and Named Entity Recognition software. If limiting search results by access and use restrictions is a functionality that will be supported in the NAFAN aggregation, it would be beneficial if the EAD data employed more cross-collection and cross-institution consistency in the expressions of these restrictions.

Use and access notes often require the user to contact the archive but do not provide contact information

The userestrict data indicate that for 58% of collections, users need to request written permission to use the materials and/or verify that they have the rights for use from the creators or copyright holders of the content. A typical example of this language instructs the user to contact the library, but does not include contact information with which to do so:

> The Library holds copyright. The researcher must secure permission to publish. All requests for permission to publish or quote from manuscripts must be submitted to the Library. The researcher assumes full responsibility for complying with copyright, literary property rights, and libel laws.

The acessrestrict data indicates that 18% of finding aids state that an institutional contact is needed to arrange access to the collection.

While displaying information that explains terms of use is important, if users lack corresponding repository contact information, they cannot pursue their access and use needs without investing additional time and energy to track down the needed information to contact the repository. NAFAN should ensure that repository contact information is clearly available alongside use and access restriction information.

Information on access and use restrictions are found in multiple fields and levels

There are four tags that may include information about access and use restrictions: accessrestrict, userestrict, physloc, and phystech. Information alerting users that collections are stored off-site and require an advance request for access were found in the accessrestrict, userestrict, physloc, and phystech elements. The phystech element frequently contained information about audiovisual or electronic records formats and limitations on their access, or physical deterioration of items that prevent or require advance request to access.

The content of the accessrestrict and userestrict elements indicates that, for some finding aid creators, the documented purposes of the two elements may lead to confusion or uncertainty. In more than 500 finding aids, the use restrictions associated with copyright were supplied in the accessrestrict element, while some finding aids used userestrict to record information on materials that only could be accessed on-site or in a particular format.

The overall EAD tag analysis reported that the accessrestrict and userestrict element values are used within archdesc in most of the EADs (accessrestrict 91%, userestrict 77%), and less frequently within the c/c01-c12 elements (accessrestrict 5%, userestrict 1.5%). While the physloc and phystech elements are used less often, they follow a similar pattern of higher use at the archdesc (physloc 27%, phystech 2.4%) versus the c/c01-c12 levels (physloc 2%, phystech 1.7%). Though usage is lower in c/c01-c12, the information at these lower levels still can be important to the user's understanding of how they can access or use the collection materials.

Discussion and Recommendations

The NAFAN EAD research corpus yielded insights into the suitability of a large aggregation of EAD encoded finding aids for a national archival finding aid aggregation by measuring the data against the notion of a minimum viable descriptive record to support discovery functionality, as well as how well suited the data is for supporting the needs other needs end users. This data analysis also generated new opportunities for future action and investigation. Based on the research findings, we present the following recommendations for the NAFAN project to consider as part of its next steps.

Data remediation and enhancement

In considering EAD data as part of NAFAN, it is clear that there are numerous opportunities to enhance and enrich data in service of supporting a robust and rich discovery experience, as well as considerations for opportunities to return enriched data to contributors. The NAFAN project should consider which remediation or enrichment would offer most value to the aggregation users and participants. Other project research findings with archivists and end users should be used to guide and support these decisions.

The NAFAN project should consider where in the data creation chain enrichment might occur. If data enrichment or remediation occurs on the aggregation side, how might participants benefit? If data quality efforts need to occur on the content creation side, how can the NAFAN project support these efforts through tool creation, training, or through other means? How can the project appropriately invest in and support such activities?

As noted in this report, there are many activities that would add value in a discovery environment such as linking to vocabularies, disambiguating names, and cleaning up dates.

The NAFAN project also plays a role as a stakeholder in archival descriptive data creation and standards, and therefore has an opportunity to advocate for and offer training resources for improved descriptive practices.

Supporting a minimum viable descriptive record

As seen in this report, compliance with DACS in the NAFAN EAD research corpus is high. This echoes findings from NAFAN focus group interviews with archivists, who reported that they were attentive to DACS requirements in creating and maintaining archival description. Given the high level of DACS compliance, DACS may be a good starting point for defining minimal record requirements within the NAFAN project. However, given the need for additional elements to support resource discovery, consideration should be given to what other elements are needed to shape a functional minimum viable descriptive record that supports project goals.

Creating an archival registry to support connecting with materials

An emerging NAFAN project goal is to create a registry for archives that would help support the needs of users that have identified materials of interest and who are ready to take the next step to contact or visit an archive. The NAFAN EAD research corpus contains useful data that could help inform and populate such a registry.

Data analysis is valuable

The type of data analysis offered in this report would be a valuable feature for the NAFAN project to offer. It would provide participants information on data quality, support a means of measuring improvement resulting from data enrichment or remediation efforts, and offer the archival community insights into patterns in usage of descriptive practice and standards. The type of data analysis reflected in this report takes time and specialized skills and tools, and it should be budgeted for accordingly.

CONCLUSION

EAD is an important format for archival collection description. EAD files that currently are managed by archival aggregators will likely form the basis of the envisioned NAFAN platform. Through better understanding the quality and characteristics of this body of EAD data, a future stage of the NAFAN may be able to leverage existing data features, as well as plan to adapt or remediate data where needed to support the needs and expectations of end users.

A primary finding of this data analysis is that there is a high level of completeness in the EAD elements that comprise a DACS single level optimum record. Another finding is that there is a low level of completeness with fields that might be leveraged in a discovery system that would enhance advanced search, browse, sort, and facet functions. Other areas of investigation were inspired by insights into the needs of end users and their desires to find and use digital content; their wishes to find collections that are related by a common topic, person, or organization; the ability to locate specific materials based on genre or physical format; and their needs to know how collections can be used and how to contact a repository. In all of these areas, the EAD data have some ways to go in terms of meeting basic needs.

There are clear opportunities for data remediation, whether through data enhancement or tools for archivists. This study may provide the basis for a conversation within the profession about what constitutes a minimum viable archival record. Finally, there are ample opportunities for additional study, including undertaking a similar analysis of archival description in MARC or other data formats that will be included in a future NAFAN platform.

ACKNOWLEDGMENTS

This project would not have been possible without funding from the Institute for Library and Museum Services. The Building a National Finding Aid Network (NAFAN), was a two-year, IMLS-funded (LG-246349-OLS-20) research and demonstration project led by the California Digital Library.

OCLC provided significant co-investment for the qualitative and quantitative research that contributed to the success of this effort. Significant thanks are due to the California Digital Library for their overall coordination of the NAFAN project.

The NAFAN Research Advisory Working Group gave invaluable guidance and perspectives, asking smart questions, helping to review and shape draft project findings and reports, and positing additional areas for inquiry. Our work benefited from the blend of rich professional backgrounds and institutional types represented in this group.

- Hillel Arnold, Assistant Director for Digital Programs, Rockefeller Archive Center
- Rachael Hu, User Experience Manager, California Digital Library
- Bergis Jules, Shift Collective
- Holly Mengel, Archivist, University of Pennsylvania, Kislak Center for Special Collections, Rare Books and Manuscripts
- Derek Mosley, Archivist/Division Manager, Auburn Avenue Research Library on African American Culture and History, Fulton County Library System
- Molly Bruce Patterson, formerly Digital Archivist & Special Collections Librarian, Rhode Island College
- Ricky Punzalan, Associate Professor, University of Michigan
- William Stingone, Rabinowitz Director for Preservation and Collections Processing, New York Public Library
- Lydia Tang, Outreach and Engagement Coordinator, Lyrasis
- Adrian Turner, Senior Product Manager, Archives, California Digital Library
- Rachel Walton, Digital Archivist, Rollins College

Jodi Allison-Bunnell (Head of Archives and Special Collections, Montana State University) and María A. Matienzo (Engineering Manager, Core Experiences, Tome; formerly Stanford University). Both brought their deep expertise with EAD and knowledge about creating and sustaining aggregations as well as a passion for the people engaged with archives to their readings of drafts of this work.

Core members of the OCLC NAFAN team contributed substantially to this work by reviewing several rafts and offering helpful feedback, helping to dig into data, and thinking about the ways in which this work connects meaningfully to the other outputs across the NAFAN research projects. This work has been enriched by their insights and encouragement.

- Lesley A. Langa, Associate Research Scientist
- Lynn Silipigni Connaway, Executive Director Research
- Brooke Doyle, Senior Project Coordinator

Other OCLC colleagues contributed to this work. Thanks to Devon Smith (Technical Manager) and Chris Cyr (formerly Associated Research Science, OCLC). Both brought data science skills to the EAD and other data sets to support this project. We also recognize Kendra Morgan (Senior Program Manager) and Janet Mason (Executive Assistant to the Vice President) for their superior administrative support for this and many other projects.

No OCLC Research output would be possible without the efforts of the OCLC Research Communications Team who bring patience, persistence, and creativity to their work. Erica Melko (Senior Communications Coordinator/Editor) makes everything we write better by smoothing and adjusting text and ensuring that our words connect with the intended audience. Jeanette McNicol brings her expertise in layout, displaying gusto and courage in the face of tables. JD Shipengrover (Information Architect) uses her skills as a graphic designer to create meaningful and effective graphics that help readers connect with data.

Finally, this work was made possible by the senior leadership of OCLC; the authors would like to thank Rachel Frick, Executive Director, Research Partnerships and Engagement and Chip Nilges, Vice President, Business Development, Membership and Research, for their ongoing support.

NOTES

1. For a discussion about regional aggregators in the United States, please see Allison-Bunnell, Jodi, and Adrian Turner (ed.). 2019. *Finding Aid Aggregation at a Crossroads*. UC Office of the President: California Digital Library. https://escholarship.org/uc/item/5sp13112.

2. Langa, Lesley A., Chela Scott Weber, and Lynn Silipigni Connaway. 2023. *Pop-Up Survey: Findings from the Building a National Finding Aid Network Project*. Dublin, OH: OCLC Research. https://doi.org/10.25333/qfjb-h531.

3. Library of Congress. "EAD: Encoded Archival Description—Official Site." Standards. Updated 10 January 2023. https://www.loc.gov/ead/;

   EADiva: A plain-talking EAD tag library. "Home." Accessed 3 March 2023. https://eadiva.com/.

4. These are the names of the aggregator partners in 2021; since that time, some have changed their name or composition.

5. Library of Congress. "EAD: Encoded Archival Description—Official Site." Standards. Updated 10 January 2023. https://www.loc.gov/ead/.

6. Society of American Archivists. "Encoded Archival Description (EAD)." Standards. Accessed March 3, 2023. https://www2.archivists.org/groups/technical-subcommittee-on-encoded-archival-standards-ts-eas/encoded-archival-description-ead.

7. Bron, Marc, Merrilee Proffitt, and Bruce Washburn. 2013. "Thresholds for Discovery: EAD Tag Analysis in ArchiveGrid, and Implications for Discovery Systems." *The Code4Lib Journal*, no. 22 (October). https://journal.code4lib.org/articles/8956.

8. Society of American Archivists' Technical Subcommittee. n.d. "(DACS) Describing Archives: A Content Standard." Github. Accessed 3 March 2023. https://github.com/saa-ts-dacs/dacs.

9. Society of American Archivists' Technical Subcommittee. 2022. "Crosswalks: DACS to ISAAR(CPF) to EAC(CPF) Permalink." Appendix C in "Describing Archives: A Content Standard." Society of American Archivists. GitHub. https://saa-ts-dacs.github.io/dacs/09_appendices/03_appendix_c_crosswalks.html#dacs-to-ead-and-marc.

10. Langa, Lesley A., Chela Scott Weber, and Lynn Silipigni Connaway. 2023. *Pop-Up Survey: Findings from the Building a National Finding Aid Network Project*. Dublin, OH: OCLC Research. https://doi.org/10.25333/qfjb-h531.

11. A list of linking attributes for EAD Version 2002 can be found here:

    Library of Congress. "Encoded Archival Description Tag Library, Version 2002: Linking Attributes." In "EAD: Encoded Archival Description Official Site." Standards. Last updated 29 March 2022. https://www.loc.gov/ead/tglib/att_link.html.

12. Zittrain, Jonathan, John Bowers, and Clare Stanton. 2021. "The Paper of Record Meets an Ephemeral Web: An Examination of Linkrot and Content Drift within the New York Times," 4. *Library Innovation Lab*, Harvard Law School. (Harvard University DASH Repository). https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37367405.

13. Octet-stream is an unspecified binary file.

14. Langa, Lesley A., Chela Scott Weber, and Lynn Silipigni Connaway. 2023. *Pop-Up Survey: Findings from the Building a National Finding Aid Network Project.* Dublin, OH: OCLC Research. https://doi.org/10.25333/qfjb-h531.

15. Library of Congress. "Encoded Archival Description Tag Library, Version 2002: EAD Elements—<persname> Personal Name." In "EAD: Encoded Archival Description Official Site." Standards, EAD. Last updated 26 May 2006. https://www.loc.gov/ead/tglib/elements/persname.html.

16. "Library of Congress Reconciliation Service for OpenRefine (LCNAF, LCSH)." Github. Accessed 3 March 2023. https://github.com/mphilli/LoC-reconcile.

17. Library of Congress. n.d. "Hamilton, Alexander, 1757–1804." LC Name Authority File (LCNAF). Linked Data Service. Authorities and Vocabularies. (Web page). Accessed 8 March 2023. https://id.loc.gov/authorities/names/n79021633.html;

    Library of Congress. n.d. "Hamilton, Alexander." LC Name Authority File (LCNAF). Linked Data Service. Authorities and Vocabularies. (Web page). Accessed 8 March 2023. https://id.loc.gov/authorities/names/n87120589.html.

18. "Parker, Quana." VIAF ID: 69729198. VIAF (Virtual International Authority File). OCLC. Accessed 3 March 2023. https://viaf.org/viaf/314807983 or https://viaf.org/viaf/69729198.

19 Note that for this path, the @href attribute's value is a link to an external XML file containing the address XML. This study followed those links and included the address XML in the analysis.

20. Library of Congress. "Encoded Archival Description Tag Library, Version 2002: EAD Elements—<address> Address." In "EAD: Encoded Archival Description Official Site." Standards, EAD. Last updated 25 May 2006. https://www.loc.gov/ead/tglib/elements/address.html.

21 Weber, Chela Scott, Itza Carbajal, Lesley A. Langa, Lynn Silipigni Connaway, Brooke Doyle, Brittany Brannon, and Merrilee Proffitt. 2023. *User Interviews: Findings from the Building a National Finding Aid Network Project*. Dublin, OH: OCLC Research. https://doi.org/10.25333/xq53-yv76.