EAD Tag Usage: Community Analysis of the Use of Encoded Archival Description Elements

Katherine M. Wisser and Jackie Dean

ABSTRACT

Encoded Archival Description has been actively implemented for more than 15 years. Research around EAD has focused on implementation and user interaction. Encoding behavior, through the analysis of a sample of finding aids, is presented here to describe the general use of elements and attributes across the EAD structure. In total, 108 repositories submitted up to 15 finding aids for the analysis; 1,136 finding aids comprise the entire sample. Descriptive statistics on element and attribute usage are presented as well as commentary on the overall picture of EAD encoding evident in the sample.

KEY WORDS

Description, Encoded Archival Description (EAD), Metadata

Encoded Archival Description was last revised in 2002. The Technical Subcommittee-Encoded Archival Description (TS-EAD) began planning the revision process for the standard at the Society of American Archivists (SAA) annual meeting in August 2010. At the meeting, the need to understand encoding behavior was discussed as an important contribution to the revision process. This research seeks to address that need. The researchers requested repositories to submit 12 to 15 finding aids to be analyzed for tag usage. No effort was made to evaluate the quality of the encoding; rather, the intent was to report what tags were being used and in what way. In total, 108 repositories submitted finding aids for this project.[1] Individual repositories were not identified in terms of encoding behaviors.

The analysis sought to understand the use of elements and attributes throughout the EAD structure. In addition, the researchers considered a global perspective to understand the EAD community's interpretation of the standard. They also undertook additional research into specific uses of elements.

Findings reveal that elements are used across the entire structure, leveraging the flexibility and the iterative nature of the EAD structure. Finding aids in the sample used both the Document Type Definition (DTD) and the schema. Finding aids generated through the Archivists' Toolkit represented a small percentage of the sample. In analyzing the tag usage, researchers split the documents into the various sections of EAD and examined them separately. The <eadheader>, <frontmatter>, and <archdesc> were all considered. The <archdesc> element was further divided based on elements above the <dsc> or within the <dsc>. Formatting elements, digital archival objects, and general attributes were also considered. Results demonstrate significant variability across the EAD encoding community. While there is some consensus, this research demonstrates that the flexibility of EAD's structure is employed both within repositories and across repositories.[2]

Literature Review

Since the release of EAD in 1995, many articles have discussed various aspects of the standard. *The American Archivist* devoted two special issues (volume 60, numbers 3 and 4, 1997) that compiled various aspects of the standard in anticipation of the release of Version 1.0 in 1998. Articles covered the need for the standard, its development, its general structure and design, and implementation considerations, and included case studies of implementation that highlighted the decisions that needed to be made in terms of the application of elements and attributes. These articles were brought together as a monograph in 1998.[3] Many implementation articles followed this introductory research, as well as articles that looked into the impact of EAD on the usability of finding

aids. For example, Eidson critically examined the design of EAD by outlining what he saw as the assumptions made in the design of the standard:

> First, by identifying the structure and marking the relationships, the results will be a multi-tiered environment that maintains the inherent qualities of the original document. This will in turn advance the function of the finding aid by creating a greater opportunity for better quality and more precise results in its searching capabilities. Second, it presumes that by creating this opportunity for better functionality in finding aids, user needs will be met with great access to archival materials via the World Wide Web.[4]

Eidson claimed that the problems with EAD result from a failure to reconceptualize the traditional structure of a finding aid. As he noted, "The struggle was with the whole concept of a finding aid and although it must have been disturbing for the project leaders to know that the selected group of experts from the archival community could not come up with clear definitions or even describe the countless boundaries of a typical finding aid, the problem was ignored."[5] Lisa R. Coats's "Users of EAD Finding Aids: Who Are They and Are They Satisfied?" included a comprehensive literature review of user studies that focus on EAD finding aids.[6] The problem with these kinds of treatments is a general misunderstanding of what EAD purports to be. While many user studies have uncovered a significant number of problems with the delivery of finding aids online, the structure of EAD does not dictate the usability of the finding aid. In fact, how the finding aid is actually displayed and what navigational techniques are employed are incredibly important areas in need of continued research. They do not, though, tell the archival community much about EAD.

EAD research, instead, should focus on encoding practices and how the standard is influencing content decisions that are being made in terms of descriptive weight. Hannah Frost's article, "Guidelines Counseling: A Comparative Analysis and Evaluation of EAD Implementation Guidelines," gets closest to doing this. She analyzed encoding guidelines to discern "the choices and decisions made by archivists as they adapt EAD to suit their data and descriptive practices."[7] Frost's findings are but a good start to understanding encoding behaviors. As she noted, "The guidelines document how finding aids are actually being encoded, or at least how institutions *think* they should be encoding their finding aids."[8] Guidelines only tell us about intentions rather than about actual encoding practice. This research seeks to address the practice.

Methodology

A total of 1,136 finding aids were submitted for analysis. The analysis process was challenging for several reasons. First, the finding aid sample was large, and several XML documents within the sample were huge. Second, the flexible

design of EAD made direct analysis difficult. Third, the iterative design of EAD also created challenges to looking at encoding behavior at various levels of the structure. After several false starts, the researchers split the EAD documents: one set covered all the elements above the <dsc> and another set covered all the elements within the <dsc>. This process provided a more manageable set of tags to analyze. It also suggested the structure of this article, as each section is dissected.

The researchers calculated raw numbers of tag occurrences. A challenge that emerged from this process was the occurrence of tags in XML comments; these were separated and extracted from the raw numbers. While the raw numbers are useful in terms of providing a view of how encoding has been constructed, the number of unique finding aids was an important metric to consider. For instance, there were 6 instances of the <abbr> tag; but those 6 appeared in only 2 finding aids. Both raw numbers and instances in unique finding aids are reported and considered in the discussions accompanying the descriptive statistics.

Results

This tag research focused on the presence of elements, attributes, and values in the sample of 1,136 finding aids. The smallest file in the sample was 3KB and the largest was 12,033KB. The size of the largest file was an anomaly; the second largest file was 5,292KB. The average file size was 273.22KB, and the median was 50KB. The mode was 11KB.

Before examining the individual sections of the structure, the general characteristics of the sample need to be considered. It is important to assess how finding aids are being encoded. In particular, the researchers were interested in the impact that Archivists' Toolkit had on the sample. Within the sample, 80 (7.0%) finding aids included the standard Archivists' Toolkit in the <creation> element: "This finding aid was produced using the Archivists' Toolkit." Those 80 represent only 10 (9.3% of 108) repositories. This may represent the encoding community in general, or it could be an aberration of the sample. For instance, some repositories using the Archivists' Toolkit may have considered their participation undesirable. The researchers made every effort to assuage that concern, but it possibly explains their low representation. It is clear, however, that a majority of repositories may be generating EAD documents using a template or other tools rather than creating them with the Archivists' Toolkit. Whether or not they are loaded into the Archivists' Toolkit to take advantage of its other functionality could not be determined, but it is important to consider the issues surrounding the ways in which encoded finding aids are generated as

the standard is revised. The cascade of changes that will be needed, aside from the transformation of existing finding aids, remains to be explored.

An examination of the version of EAD that comprises the finding aids in the sample netted some interesting results. In the sample, 164 (14.4%) finding aids reference the schema, while 840 (73.9%) reference EAD Version 2002 DTD. In addition, 132 (11.6%) did not reference either the schema or the DTD. It is unclear how these repositories ensure valid EAD documents, although no evidence indicates that these documents are not valid. There is no such evidence in the sample of encoding EAD Version 1.0.

Table 1. General Statistics for EAD Finding Aids (*n*=1,136)

Element	Number in Sample	% in Sample
eadheader	1,136	100.0%
frontmatter	279	24.6%
archdesc	1,136	100.0%
archdescgrp	0	0.0%
eadgrp	0	0.0%
dsc	1,050	93.0%
dscgrp	0	0.0%

Given that <eadheader> and <archdesc> are required elements in EAD, it is not surprising to see a comprehensive number in the sample (see Table 1). The <frontmatter> element is optional, employed in under a quarter of the documents in the sample. The two "grp" elements at this level were not used in the sample. It is interesting to note that 7.0% of the finding aids were single-level descriptions, not employing the <dsc> element structure that EAD specifically offers.

The <eadheader> Element

The required <eadheader> element provides information about the creation of the electronic document. It has two required elements, <eadid> and <filedesc>, and two optional elements, <profiledesc> and <revisiondesc>. There are few surprises in the element usage within the <eadheader>.

Table 2. Elements Used within <eadheader> (*n*=1,136)

Element	Number in Sample	Number in Unique Finding Aids	% in Sample
eadid	1,136	1,136	100.0%
filedesc	1,136	1,136	100.0%
profiledesc	1,114	1,114	98.1%
revisiondesc	372	372	32.7%

Table 3. Attributes Used with <eadheader> (*n*=1,136)

Attribute	Number in Sample	Number in Unique Finding Aids	% in Sample
countryencoding	1,017	1,017	89.5%
dateencoding	1,002	1,002	88.2%
findaidstatus	316	316	27.8%
langencoding	1,079	1,079	95.0%
repositoryencoding	997	997	87.8%
scriptencoding	822	822	77.6%

On the whole, the values for the encoding attributes followed the recommendations of the tag library: the @countryencoding had a uniform value of "iso3166-1"; the @dateencoding value was "iso8601"; and the @scriptencoding value was iso15924. Deviations from this include a variation in referencing the standard for @langencoding. Some referenced "iso639-2," while others referenced "iso639-2b." With @repositoryencoding, 994 (99.7%) had the value "iso15511," but 3 had the value "Archiv." For @findaidstatus, the values were much more diverse. Many of the values reflected those constrained in Version 1.0: "unverified-partial-draft," "unverified-full-draft," "edited-full-draft," and "edited-partial-draft." With the change to EAD 2002, when @findaidstatus became an uncontrolled value list, additional values were used, such as "processed," "partial," "completed," "In_process," "For_supervisor_review," "proviso ire," "published," "originaldraft," "publish," "unprocessed," and "approved." In addition, capitalization varied with several of the values.

Table 4. Attributes Used with <eadid> (*n*=1,136)

Attribute	Number in Sample	Number in Unique Finding Aids	% in Sample
countrycode	1,071	1,071	94.3%
mainagencycode	1,052	1,052	92.6%
publicid	353	353	31.1%
url	481	481	42.3%
urn	44	44	3.9%
identifier	560	560	49.3%

Again, the use of attributes on the <eadid> element are not terribly surprising. The @countrycode and @mainagencycode attributes were nearly universal, whereas the other attributes were used with significantly less frequency. It is interesting that 54 finding aids used @countrycode, more than those that referenced the standard "iso3166-1." In addition, @countrycode and @identifier are attributes available in other elements. The number in the sample represents only those present in the <eadid> element.

Table 5. Elements within <filedesc> (*n*=1,136)

Element	Number in Sample	Number in Unique Finding Aids	% in Sample
editionstmt	45	45	4.0%
notestmt	103	103	9.1%
publicationstmt	1,081	1,081	95.2%
seriesstmt	1	1	0.1%
titlestmt	1,136	1,136	100.0%

The <fildesc> element is required, and, within it, <titlestmt> is required. In addition to the use of <titlestmt>, a majority of the finding aids used <publicationstmt>. Other elements, such as <editionstmt> and <notestmt>, were used rarely, and <seriesstmt> appeared in only one finding aid.

Table 6. Elements within <profiledesc> (*n*=1,114)

Element	Number in Sample	Number in Unique Finding Aids	% in Sample
creation	1,076	1,076	96.6%
descrules	486	486	43.6%
langusage	1,047	1,047	94.0%

The <profiledesc> element provides information about the creation of the EAD document. It is not surprising to see the pervasive use of the <creation> element. In addition, the strong recommendation to include language information about documents can account for the wide use of <langusage> in the <profiledesc>. Surprising, therefore, is the less uniform use of the <descrules> element, which refers to the descriptive rules used to create the finding aid. This might, perhaps, be accounted for by the appearance in 2004 of *Describing Archives: A Content Standard* (*DACS*), six years after the initial publication the EAD Version 1.0 and two years after the release of EAD 2002.

Table 7. Elements within <revisiondesc> (*n*=372)

Element	Number in Sample	Number in Unique Finding Aids	% in Sample
change	552	345	92.7%
list	7,575	27	7.3%

The formatting element <list> is available across the structure of EAD in 39 elements. Therefore, the number of <list> elements above does not necessarily represent multiple lists within <revisiondesc>, while the <change> element is only available in <revisiondesc> and represents multiple <change> elements in the <revisiondesc>. The usefulness of <list> here is to demonstrate that the option is being used for <revisiondesc>.

The <frontmatter> Element

The <frontmatter> element represents a formal title page and is understood to be a bridge between legacy finding aid structures and new presentations in the online environment. Only 279 (24.6%) of the finding aids in the sample used it. Analysis of the structure of <frontmatter> as the finding aids used it reveals that, in general, it followed the <titlepage> structure.

Table 8. Elements within <frontmatter> (*n*=279)

Element	Number in Sample	% in Sample
titlepage	259	92.8%
div	6	2.2%
empty	14	5.0%

The *empty* value in Table 8 represents encoding <frontmatter></frontmatter> or <frontmatter/> within those 14 finding aids. The researchers assumed that the element was part of a template but is not being used any longer.

The <archdesc> Element

The <archdesc> element contains the descriptive components of the finding aid for the materials being described. It has only a few requirements, including the @level attribute and the use of a <did> element. All other components of the finding aid are optional and, as demonstrated in Table 12, represent a wide variety of implementation choices. The two divergent approaches to archival description stand out in this analysis. The first relies to a large measure on the description at the <archdesc> level; description at lower levels inherits the information from the <archdesc>-level description and is therefore more streamlined. The alternative is sparse description at the <archdesc> level and relies more heavily on description at lower levels (within the <dsc>).

Table 9. Values for @level within <archdesc> (*n*=1,136)

Level Value	Number in Sample	% in Sample
collection	1,033	90.9%
fonds	55	4.8%
class	3	0.3%
recordgrp	16	1.4%
series	7	0.6%
subfonds	3	0.3%
subgrp	11	1.0%
subseries	0	0.0%
file	4	0.4%
item	3	0.3%
otherlevel	1	0.1%
TOTAL	1,136	100.0%

The single instance of "otherlevel" is paired with an "otherlevel" attribute with a value "accession." As can be seen in Table 9, EAD finding aids are being created at many different starting points. Only the value "subseries" was not represented in the sample. While this is the case, it is not surprising that "collection," "fonds," and "recordgrp" represented the largest number of level values (1,104 of 1,136 or 97.2%). These three represented the highest level of description in various level structures.

The <archdesc>/<did> Element

The <did> element represents the core of description at any level. It contains descriptive identifying information, such as title and date, creator, and so on. All elements in the <did> take PCDATA directly; none of them includes elements such as <p>; all elements are unique to the <did>. Therefore, it was relatively easy to analyze the structure of the <did> across the sample.

Table 10. Elements within <archdesc>/<did> (*n*=1,136)

Element	Number in Sample	Number in Unique Finding Aids	% of Unique Finding Aids in Sample
abstract	1,085	984	86.6%
container	10	4	0.4%
langmaterial	1,042	1,021	89.9%
materialspec	18	18	1.6%
origination	1,216	1,011	89.0%
physdesc	1,176	1,104	97.2%
physloc	557	316	27.8%
repository	1,141	1,132	99.6%
unitdate	1,651	1,102	97.0%
unitid	1,151	1,024	90.1%
unittitle	1,582	1,136	100.0%

The occurrences of elements noted above do not represent a significant departure from the perceptions of encoding as Frost noted.[9] Some surprises include a finding aid with 108 <origination> elements and another with 39. The most common count above 1 <origination> element was 2, perhaps 3. There were an additional 446 <unittitle> elements, indicating that multiple <unittitle> elements are being used. The low occurrences of <container> are not surprising as traditionally that element is employed at lower levels of description, but the low use of <materialspec> does indicate that that element may not be much used or well understood.

Table 11. Elements within <archdesc>/<did>/<physdesc> (*n*=1,136)

Element	Number in Sample	Number in Unique Finding Aids	% of Unique Finding Aids in Sample
dimensions	28	20	1.8%
extent	1,383	867	76.3%
physfacet	27	19	1.7%

Drilling into the physical description component of the <did>, the researchers found the <extent> element to be the dominant subelement. It often occurred in multiple instances. For example, 3 finding aids had 5 instances of <extent>, 3 had 6, and 1 had 7. For both <dimensions> and <physfacet>, 4 finding aids included 3 instances of the tags.

The <archdesc> Element: above the <dsc>

Other elements that may occur within the <archdesc> element above the <dsc> include such descriptive components as administrative information, biographical and historical information, scope and content notes, bibliography, index, related materials, separated materials, and so on. To gain an overall picture of the use of these elements at the high-level description, general descriptive statistics were derived. Instances and unique finding aids are reported.

Table 12. Other Elements within <archdesc> above the <dsc>

Element	Number in Sample	Number in Unique Finding Aids	% in Sample
accessrestrict	991	979	86.2%
accruals	81	81	7.1%
acqinfo	796	772	68.0%
altformavail	152	144	12.7%
appraisal	57	54	4.8%
arrangement	761	744	65.5%
bibliography	134	115	10.1%
bioghist	1,118	992	87.3%
controlaccess	3,543	966	85.0%
custodhist	163	160	14.1%
descgrp	557	363	32.0%
fileplan	7	7	0.6%
index	48	14	1.2%
odd	214	110	9.7%
originalsloc	39	39	3.4%
otherfindaid	146	135	11.9%
phystech	48	48	4.2%
prefercite	970	970	85.4%
relatedmaterial	494	458	40.3%
runner	12	12	1.1%
scopecontent	1,111	1,061	93.4%
separatedmaterial	178	168	14.8%
userestrict	808	776	68.3%

With the exception of <fileplan> and <runner>, the use of elements is relatively widespread. Many of the elements represent information that would only be employed if relevant to a collection (e.g., <originalsloc>, <phystech>, and <custodhist>), while other elements represent particular descriptive styles (e.g., <bibliography> and <index>). Elements that appear to be relatively universal (e.g., <bioghist>, <scopecontent>, <controlaccess>, and a handful of the administrative descriptive elements) are not surprising, based on what Frost discovered in her analysis of encoding guidelines.[10] A comparison with the appearance of these elements within the <dsc> is noted in Table 20.

The <dsc> Element

Within the sample, 1,053 finding aids contained at least one <dsc> element. A corresponding 83 finding aids in the sample did not contain a <dsc> element, although some did include <dsc></dsc>.

Table 13. The Inclusion of <dsc> in Finding Aids (*n*=1,136)

Element	Number in Sample	% in Sample
One <dsc>	1,026	90.3%
Multiple <dsc>s	27	2.4%
No <dsc>s	83	7.3%

The 1,136 finding aids in the sample contained 1,105 <dsc> sections. In total, 1,026 finding aids contained one <dsc> element, 20 contained 2 <dsc> elements, and 2 contained 3 <dsc>s. The highest number of <dsc> elements in a single finding aid was 9. Of the 1,105 <dsc>s, by far the most common value of the type attribute was "combined."

Table 14. <dsc> Type Attributes (*n*=1,105)

@type Values	Number in Sample	% in Sample
Total <dsc>s	1,105	97.2% (*n*=1,136)
no type attribute	90	8.1%
analyticover	56	5.1%
combined	735	66.5%
in-depth	185	16.7%
othertype	39	3.5%

For those @type values that were "othertype," 11 included an @othertype. The values for that attribute include "containerlist-inmagic," "listoutput," "unprocessed," "segregated," and "sonsttypen."

Table 15. <c>–<c12> Tags (*n*=1,053)

Element	Number in Sample	Number in Unique Finding Aids	% in Sample
<c>	113,133	117	11.1%
<c01>	31,792	927	88.0%
<c02>	189,148	763	72.5%
<c03>	239,029	440	41.8%
<c04>	104,161	217	20.6%
<c05>	31,306	113	10.7%
<c06>	10,820	48	4.6%
<c07>	3,127	21	2.0%
<c08>	1,546	7	0.7%
<c09>	485	3	0.3%
<c10>	10	1	0.1%
<c11>	0	0	0.0%
<c12>	0	0	0.0%
Total # of <c>–<c12>	724,557		

Nine finding aids contained <dsc> sections that did not include <c> elements. These typically consist of a note in a <p> tag that carries information about the collection being unprocessed. No finding aids included both numbered and unnumbered <c> tags.

Drilling down to the lower level of descriptions presents an interesting view of the use of subordinate components. For instance, 82.3% of the finding aids that included <c01> also included a <c02>; similarly, 57.7% of the finding aids with <c02> also included <c03>. The percentages hover at this range for <c04> through <c07> (49.3% of <c03> finding aids contained <c04>; 52.1% of <c04> finding aids contained <c05>; 42.5% of <c05> finding aids contained <c06>; 43.8% of <c06> finding aids contained <c07>). After <c07>, rates drop significantly: 33.3% of finding aids with <c07> contained <c08>; 42.9% of finding aids with <c08> contained <c09>; and 33.3% of finding aids with <c09> contained <c10> (see Figure 1).

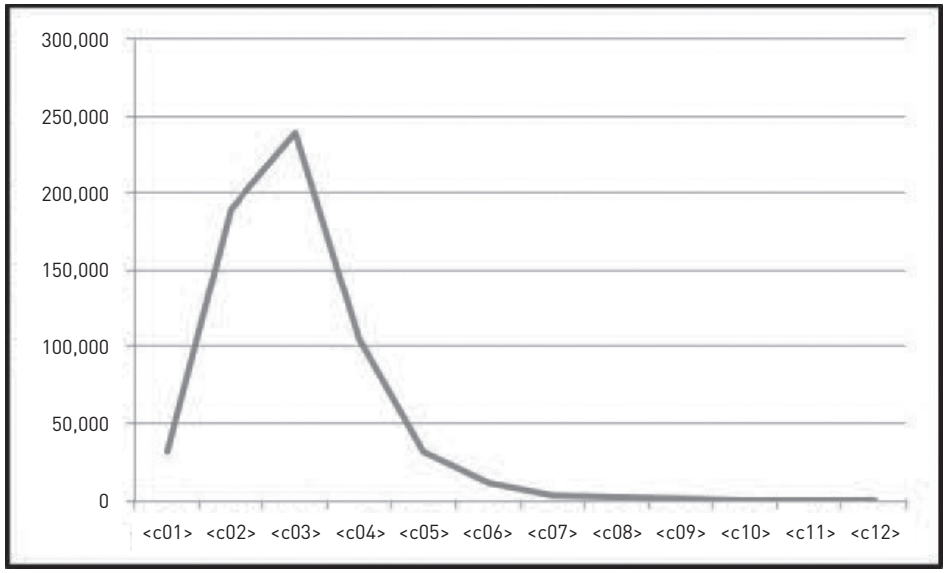

FIGURE 1. Distribution of <c01>–<c12> tags.

The utility of the percentages here is questionable, given that there are a total of 724,557 total component elements (<c> through <c12>) and a total of 539,214 level attributes used (74.4% of the component elements contained a level attribute). The use of @level is not required for the component elements. In addition, the sample included 14,880 <c> tags (2.1% of component elements) without the level attribute. For elements <c01> through <c12>, 170,463 (23.5% of component elements) did not contain @level.

Table 16. Level Attribute on <c> (*n*=1,053)

Value of @level	Number in Sample	Number in Unique Finding Aids	% in Sample
collection	509	22	2.1%
fonds	12	7	0.7%
class	1,535	13	1.2%
recordgrp	31	7	0.7%
series	8,390	818	77.7%
subfonds	119	18	1.7%
subgrp	339	33	3.1%
subseries	14,962	372	35.3%
file	357,262	599	56.9%
item	130,178	255	24.2%
otherlevel	25,877	96	9.1%

The "otherlevel" value can be accompanied by an @otherlevel. For the 25,877 level values that were "otherlevel," 25,452 included an @otherlevel. Values for @otherlevel include: "accession"; "box"; "container"; "content_list"; "cross-reference"; "cumulative"; "event"; "filegrp"; "group"; "groupe-de-notices"; "interview"; "item"; "MAD3.50"; "namegroup"; "notice"; "section"; "segment; "série"; "series"; "sleeve"; "sous-notice"; "sub"; "sub-fonds"; "sub-series"; "sub-sub_series"; "sub-sub-sub-subseries"; "sub-sub-subseries"; "sub-subseries"; "sub-subgrp"; "sub-subseries"; "sub-subsubseries"; "subfile"; "subitem"; "subseries"; "subsubseries"; and "topicalgrouping."

Table 17. <c>–<c12>/<did> Elements (*n*=1,053)

Element	Number in Sample	Number in Unique Finding Aids	% in Sample
abstract	7,128	26	2.5%
container	704,884	869	82.5%
langmaterial	10,078	64	6.1%
materialspec	8,395	14	1.3%
origination	20,549	85	8.1%
physdesc	124,763	573	54.4%
physloc	11,354	61	5.8%
repository	2,651	3	0.3%
unitdate	470,673	954	90.6%
unitid	233,952	486	46.2%
unittitle	697,246	1,041	98.9%

The total number of component levels in all finding aids in the sample is 724,557; the number of <unittitle>s is 697,246, meaning that there are 27,311 component levels that did not use <unittitle>. Combining the unique finding aids with <c> or <c01> (1,044) and comparing that to the total finding aids that included <unittitle>, only 3 finding aids did not include <unittitle> in the highest-level subordinate component.

Table 18. <c>–<c12>/<did>/<physdesc> Elements (*n*=1,053)

Element	Number in Sample	Number in Unique Finding Aids	% of Unique Finding Aids in Sample
dimensions	17,717	55	5.2%
extent	76,171	385	36.6%
physfacet	29,426	72	6.8%

A larger number of <dimensions> and <physfacet> occurrences appeared in the <dsc> than at the <archdesc> level of the finding aids, but the <extent> information remained the dominant subelement. It should be noted that at least 61 <physdesc> elements did not use any subelements (10.6% of 573, the number of unique finding aids with <physdesc> in the <dsc>). If compared with the use of subelements at the <archdesc>-level encoding, it is surprising to note that that practice was greater at the collection level (906 or 17.9% of 1,104, the number of unique finding aids with <phsydesc> in the <archdesc>-level <did>). The researchers did not anticipate increased encoding precision within the <dsc>, but it is clear both here and in the tables below that significant encoding is taking place within the <dsc> as well as at the <archdesc> level.

Table 19. Other Elements Found in <c>–<c12> (*n*=1,053)

Element	Number in Sample	Number in Unique Finding Aids	% in Sample
accessrestrict	6,727	113	10.7%
accruals	0	0	0.0%
acqinfo	3,613	47	4.5%
altformavail	5,518	28	2.7%
appraisal	43	7	0.7%
arrangement	2,128	200	19.0%
bibliography	1,435	16	1.5%
bioghist	1,099	48	4.6%
controlaccess	58,366	54	5.1%
custodhist	5,504	23	2.2%
descgrp	785	19	1.8%
index	2,943	7	0.7%
note	25,961	214	20.3%
odd	16,525	76	7.2%
originalsloc	150	11	1.0%
otherfindaid	463	24	2.3%
phystech	570	16	1.5%
prefercite	2	1	0.1%
processinfo	740	40	3.8%
relatedmaterial	1,701	46	4.4%
runner	0	0	0.0%
scopecontent	110,648	645	61.3%
separatedmaterial	0	0	0.0%
userestrict	848	34	3.2%

While none of the descriptive elements in the table above is pervasive across all the 1,056 <dsc> elements, 4 made an appearance in more than 10% of the subordinate component lists, with <scopecontent> being the most frequently used. With only 3 elements not appearing anywhere (<separatedmaterial>, <runner>, and <accruals>), this demonstrates the variability that is the hallmark of the <dsc> structure and the information that it represents.

Table 20. A Comparison between the <archdesc>-level and Component-level Use of Elements outside <did>

Element	Number in Sample by Level		Number in Unique Finding Aids		% in Sample	
	<archdesc> (*n*=1,136)	component (*n*=1,053)	<archdesc> (*n*=1,136)	component (*n*=1,053)	<archdesc> (*n*=1,136)	component (*n*=1,053)
accessrestrict	991	6,727	979	113	86.2%	10.7%
accruals	81	0	81	0	7.1%	0.0%
acqinfo	796	3,613	772	47	68.0%	4.5%
altformavail	152	5,518	144	28	12.7%	2.7%
appraisal	57	43	54	7	4.8%	0.7%
arrangement	761	2,128	744	200	65.5%	19.0%
bibliography	134	1,435	115	16	10.1%	1.5%
bioghist	1,118	1,099	992	48	87.3%	4.6%
controlaccess	3,543	58,366	966	54	85.0%	5.1%
custodhist	163	5,504	160	23	14.1%	2.2%
descgrp	557	785	363	19	32.0%	1.8%
index	7	2,943	7	7	0.6%	0.7%
note	48	25,961	14	214	1.2%	20.3%
odd	214	16,525	110	76	9.7%	7.2%
originalsloc	39	150	39	11	3.4%	1.0%
otherfindaid	146	463	135	24	11.9%	2.3%
phystech	48	570	48	16	4.2%	1.5%
prefercite	970	2	970	1	85.4%	0.1%
processinfo	494	740	458	40	40.3%	3.8%
relatedmaterial	12	1,701	12	46	1.1%	4.4%
runner	1,111	0	1,061	0	93.4%	0.0%
scopecontent	178	110,648	168	645	14.8%	61.3%
separatedmaterial	808	0	776	0	68.3%	0.0%
userestrict	991	848	979	34	86.2%	3.2%

Table 21. Content Tags in dsc (*n*=1,053)

Element	Number in Sample	Number in Unique Finding Aids	% in Sample
corpname	11,384	88	8.4%
famname	251	18	1.7%
function	0	0	0.0%
genreform	25,893	54	5.1%
geogname	28,082	66	6.3%
name	23,555	15	1.4%
occupation	204	4	0.4%
persname	66,197	136	12.9%
subject	28,767	49	4.7%

While the counts of individual uses of the tags were high, the use of content tags, elements used to designate the content of inline text, were limited to a small number of finding aids in the <dsc>.

Formatting Elements

Table 22. Formatting Elements

Element	Number in Sample
blockquote	732
defitem	6,622
div	53
emph	48,826
head	28,039
head01	9
head02	9
item	55,417
label	6,622
lb	18,295
list	7,575
listhead	9
p	248,907
runner	12

Table 22 illustrates elements that are primarily used for formatting text. Many of these elements are so pervasive throughout EAD that unique finding aid occurrence was difficult to obtain. The raw numbers do provide an overview of the elements in use, but they can be deceptive. For instance, <blockquote> has 732 occurrences but appeared in no more than 20 finding aids; the element , with 53 occurences, appeared in only 8 finding aids. Note the 12 instances of <runner> appeared in finding aids from one repository.

Table 23. <dsc> Type Attributes

Attribute	Number in Sample
continuation="continues"	0
mark	36
numeration	99
numeration="arabic"	85
numeration="upperalpha"	2
numeration="loweralpha"	12
numeration="upperroman"	0
numeration="lowerroman"	0
type	4,120
type="simple"	2,756
type="deflist"	987
type="marked"	238
type="ordered"	139

The <list> attributes demonstrate some variable encoding. For example, of the 7,575 <list> elements in the sample, only 4,120 had an identified @type. The marked list had only 36 instances that identified the type of mark using the mark attribute.

Table 24. Values for @render

Value	Number in Sample
altrender	0
bold	4,423
doublequote	9,709
bolddoublequote	0
bolditalic	1,201
boldsinglequote	0
boldsmcaps	378
boldunderline	29
italic	42,309

Value	Number in Sample
nonproport	4
singlequote	12
smcaps	60
sub	7
super	18,344
underline	668

Table 25. Table Elements

Element	Number in Sample
colspec	75
row	22,706
table	123
tbody	114
tgroup	114
thead	414

Table 26. Attributes Associated with Table Elements

Attribute	Number in Sample
align	37
char	0
charoff	0
colname	6,704
cols	114
colsep	13
colwidth	70
frame	15
morerows	5,898
nameend	0
namest	66
pgwide	0
rowsep	11
tpattern	0
valign	666

Tables 24, 25, and 26 look at specific formatting possibilities available in the EAD structure. @render pervaded the sample finding aids, although not all values were used.

Digital Archival Objects Elements

Table 27. <dao> Elements (*n*=1,136)

Element	Number in Sample	Number in Unique Finding Aids	% in Sample
dao	24,997	87	7.7%
daodesc	2,217	136	12.0%
daogrp	5,503	106	9.3%
daoloc	12,193	123	10.8%

Table 28. Linking Attributes (*n*=1,136)

Attribute	Number in Sample
actuate	10,803
arcrole	48
entityref	68
from	2,564
href	22,650
linktype	23,258
parent	37,610
role	55,595
show	10,811
target	38,881
title	1.727
to	2.564
xpointer	57
xlink attributes	69,920

The xlink attributes used included xlink:type, xlink:href, xlink:title, xlink:actuate, xlink:show, xlink:from, and xlink:to, with xlink:type being the most prevalent (47,719 occurrences).

General Attributes

Table 29. General Attributes

Attribute	Number in Sample
normal	176,665
source	43,019
unit	3,424
rules	10,107
langcode	11,241
scriptcode	695
id	259,500
altrender	15,360

Table 30. Date Attributes

Attribute	Number in Sample
calendar	193,658
certainty	5,101
datechar	4,229
era	193,669
type	111,125
type="inclusive"	110,744
type="simple"	381

Table 31. @relatedencoding and @encodinganalog

Attribute and Value	Standard	Number in Sample	% of All @relatedencoding
relatedencoding		1,622	
	MARC	1,079	66.5%
	Dublin Core	521	32.1%
	ISAD(G)v2	19	1.2%
	MidosaXML	3	0.2%
encodinganalog		358,165	

@relatedencoding is mapped to 4 standards. The representation of those standards varies, however. For instance, the MARC standard is represented as "MARC 21," "MARC," "MARC21," and "USMARC" with variations in capitalizations. Dublin Core, likewise, is represented as "Dublin Core," "dublincore," and "dc."

Table 32. Multiple @relatedencoding

Number of @relatedencoding	Number in Sample	% with @relatedencoding (*n*=885)	% in Sample (*n*=1,136)
one	222	25.1%	19.5%
two	569	64.3%	50.1%
three	94	10.6%	8.3%
TOTAL	885		77.9%

Finding aids with multiple relatedencoding attributes either included a combination of standards or referenced the same standard in the various sections of the EAD document. Combinations found include MARC and Dublin Core or MARC and ISAD(G) version 2. When it was the same standard, it was MARC. Relatedencoding attributes were placed on <ead>, <eadheader>, and/or <archdesc>, with the 94 finding aids with three @relatedencoding having that attribute on each of those elements.

Discussion

The analysis of tag usage presented here indicates that little uniformity exists in encoding practices. While Frost made a compelling argument in her introduction for the need for uniform encoding practices, even though many guidelines and best practices have been established, the flexibility of the EAD structure is being taken full advantage of in practice. With the exception of those aspects of EAD that the schema requires, encoding varies widely. Not surprisingly, variability is more striking within the <dsc> section of the finding aid than above the <dsc>. Encoding behaviors in the <eadheader> and <frontmatter> elements appear to be the most uniform, but even within those relatively static aspects of the EAD structure there is inconsistency.

The sample size and the number of contributing institutions in this study allow the researchers to derive some characterstics of encoding behavior. One caveat is that elements with 0 or low appearance should not be dismissed outright. The researchers believe that in a survey of the use of specific elements, repositories engaged in the use of EAD would supply examples of those tags. In addition, the number of instances of an element across the sample versus the number of unique finding aids that have at least one instance of an element are significantly different perspectives. For example, the <c07> appears over 1,500 times, but only in 7 finding aids. The component levels particularly demonstrate this phenomenon, but it is also the case with other elements. In terms of attribute use, the variability is even greater. It appears that the application of

attributes varies in the context of encoding practice: some repositories put great store in attributes, while others seek a more streamlined and simple encoding procedure and forego the use of attributes.

This research was not intended to construct best practices; rather, it was intended to provide a view of existing practices to initiate discussion within the encoding community. The contributing institutions will likely be interested in comparing their own practices with those represented as general trends in the results presented here. Nonetheless, the researchers feel that despite the variability, the problems do not necessarily lie with individual institution implementation, but with a design that is perhaps overly flexible or with a lack of discussion about the impact of encoding decisions.

Variability in implementation of encoding standards has the potential to diminish the ability to aggregate records and effectively leverage structures for management and retrievability, but it is not the sole cause of these problems. With appropriate measures taken with the data, there are ways to resolve variability. That being said, variability does add a layer of effort that may diminish the advantages of structuring the data in the first place.

Conclusion

This study demonstrates that with the exception of the required elements (<eadheader>, <eadid>, <filedesc>, <archdesc>, and <did>), no uniform encoding template is in use. While there is some agreement on general elements, particularly at the <archdesc> level, ultimately, the flexible design and iterative nature of the standard enable broadly varying encoding behaviors throughout the community. This study, however, represents only the start of potential research in this area. As mentioned, specific tag analysis could provide help in understanding the purpose of certain elements. Such likely candidates include the use of administrative elements (<accessrestrict>, <userestrict>, <altformavail>, etc.) at various levels within finding aids. The use of attributes that seek to normalize or standardize information for machine manipulation (@normal, @langcode, @scriptcode, etc.) also deserves more attention. Finally, and perhaps most importantly, the interplay between content and encoding is an important extention of this research.

Notes

[1] A list of repositories may be acquired through a request to the authors.

[2] The raw number of elements and attributes that appear in the sample, a list of contributing institutions, and a table of element contents and attributes can be obtained from the authors upon request.

[3] *The American Archivist* 60, no. 3 (1997) and *The American Archivist* 60, no. 4 (1997). Republished as Jackie M. Dooley, ed., *Encoded Archival Description: Context, Theory and Case Studies* (Chicago: Society of American Archivists, 1998).

[4] Matthew Young Eidson, "Describing Anything that Walks: The Problem behind the Problem of EAD," *Journal of Archival Organization* 1, no. 4 (2002): 10.

[5] Eidson, "Describing Anything that Walks," 11.

[6] Lisa R. Coats, "Users of EAD Finding Aids: Who Are They and Are They Satisfied?," *Journal of Archival Organization* 2 (2004): 25–39. In particular, Coats provides a comprehensive review and bibliography of user studies that look at online finding aids, primarily in conjunction with the implementation of EAD.

[7] Hannah C. Frost, "Guidelines Counseling: A Comparative Analysis and Evaluation of EAD Implementation Guidelines," *Journal of Archival Organization* 1, no. 3 (2002): 74. Emphasis in the original.

[8] Frost, "Guidelines Counseling," 74.

[9] Frost, "Guidelines Counseling," 79–81.

[10] Frost, "Guidelines Counseling," 80–81.

ABOUT THE AUTHORS

Katherine M. Wisser is assistant professor and codirector of the dual degree program in archives and history at the Graduate School of Library and Information Science at Simmons College. She has previously served as the director of Instructional Services at the School of Information and Library Science (SILS), UNC-Chapel Hill and worked professionally in New Hampshire and North Carolina. She was a teaching fellow at the SILS from 2001 to 2009. She teaches courses on bibliographic cataloging, archival description, abstracting and indexing, metadata, and the history of libraries. She has an MA in American history from the University of New Hampshire and an MSLS and a PhD in information science from UNC. She served as chair of the EAC Working Group, which released Encoded Archival Context–Corporate Bodies, Persons and Families (EAC-CPF) in March 2010 and currently serves as cochair of the Technical Subcommittee for Encoded Archival Context, which maintains the standard.

Jacqueline M. Dean is the manuscripts processing coordinator at the Louis Round Special Collections Library at the University of North Carolina at Chapel Hill. In this position, she manages activities relating to processing of manuscript materials and archival description, including training and supervising graduate student processors and maintaining documentation. Dean is currently serving on SAA's Technical Subcommittee on Describing Archives: A Content Standard (TS-DACS). She teaches archival description at the School of Information and Library Science at UNC-Chapel Hill. Dean previously held positions in the North Carolina Exploring Cultural Heritage Online project at the State Library of North Carolina, at the Special Collections Research Center at North Carolina State University, and at Houghton Library, Harvard University.