Kerstin Arnold: EAD3 and the consequences of the new version

Abstract:
Encoded Archival Description (EAD) has been undergoing a major revision during the last three and a half years. The current article is trying to shed a light on the aims of this revision, on the steps undertaken as well as on some of the changes envisaged with EAD3. The article thereby covers the activities of the Technical Subcommittee on EAD (TS-EAD) of the Society of American Archivists (SAA) until early February 2014 when the EAD3 gamma has been made available for final comments by the EAD community. The last part of the article also provides a first evaluation of the expected impact of a possible move from EAD 2002 to EAD3 with the example of the Archives Portal Europe.

Introduction

Encoded Archival Description (EAD) – the oldest and probably best established of the international archival descriptive standards – has been undergoing a major revision during the last three and a half years. In early February 2014, the third draft of the new EAD format – EAD3 – was released by the Technical Subcommittee on EAD (TS-EAD) of the Society of American Archivists (SAA), which has maintained the standard since its first version back in the late 1990s. This EAD3 gamma release is now available for final comments from the EAD community until 1 March 2014.

The latest status of the schema development can be followed at GitHub where it is also possible for GitHub users to have a look at the pending issues that are still under discussion and to flag additional issues found in EAD3 gamma. Anyone who is not registered on GitHub can provide feedback via a form.

Furthermore, there is a draft EAD3 Tag Library (gamma release) available and the EAD community is invited to provide test instances encoded in EAD3 gamma at GoogleDrive. These test instances of course also can include comments (<!-- [text] -->) in order to provide feedback to the TS-EAD on aspects that might not be clear enough yet or that may require further revision and detailing.

Aims of revising EAD

When the TS-EAD set out to revise EAD after a first call for comments between October 2010 and February 2011, several general aims for the revision were defined:

Improve the interoperability with other archival standards

This refers especially to EAC-CPF (Encoded Archival Context – Corporate Bodies, Persons, and Families)[1], the communication standard for describing persons, families and corporate bodies that have created, maintained
or otherwise worked with certain archival material and for exchanging these descriptions between different source and target systems. EAC-CPF had previously undergone revision and was published in its current version in March 2010 i.e. about half a year before the EAD revision process started. One goal for the latter therefore was to orientate a new version of EAD to solutions found when creating EAC-CPF with regard to encoding certain pieces of information such as dates and control data on the EAD instances themselves. Other international standards used by the archival community that formed the focus of this aim were METS (Metadata Encoding and Transmission Standard)\(^2\) and PREMIS (Preservation Metadata: Implementation Strategies)\(^3\) among others.

Alternatively, to look at the way in which other standards use and encode certain information, an option for improving the interoperability of EAD with these standards could also (have) been to include corresponding namespaces and to thereby allow for the possibility to include certain data encoded in another standard within an EAD document. A third aspect of discussion in this context was to improve interoperability by defining more clearly and more precisely than was done in the current version EAD 2002\(^4\), where and how documents in other standards can be linked i.e. in which element such links should be included, if an attribute should be used with it for specifications and if there should be a fixed list of values for such corresponding attributes.

### Improve the usability and exchange of EAD within itself

The second aim was seen in relation to various aspects that had been commented on by members of the EAD community already in the early stages of the EAD revision. With EAD being a rather flexible framework that allows for adaptation to one's own circumstances, needs and purposes, the EAD instance from one institution does not necessarily equal the EAD instance of another institution. This can complicate an exchange of EAD instances between different institutions, especially when it comes to cases where the same – or at least similar – type of information can be encoded in different EAD elements.

So, one request in this context has been to be more definite on how to encode certain information, in tandem with the aspect of possibly minimising the amount of elements in total. This concerned not only the question of referencing and linking elements such as the subelements of `<controlaccess>` and their relation to external authority files for names, subject headings, place names, etc, but also the use of elements like the numbered and unnumbered version of the `<c>` elements representing the different subcomponents within the hierarchy of an archival fonds, collection or other type of records group.

Other aspects of this aim have been to concentrate on encoding the data instead of their presentation i.e. the question whether formatting elements – which also have corresponding counterparts in HTML like `<head>`, `<table>`, `<list>` or `<emph>` – should be maintained in EAD directly or rather left to Extensible Stylesheet Language Transformations (XSLT) for online presentation.

Thinking along similar lines though with a different emphasis, there furthermore has been the request to minimise the options for mixed content i.e. EAD elements that allow for having content themselves next to having subelements for additional tagging of parts of their content. An example could be the encoding of language information regarding the material described in the format of:

```xml
<langmaterial>
  <language langcode="eng">English</language>
  <language langcode="fre">French</language>
  <language langcode="lat">Latin</language>
</langmaterial>
```

opposed to only having:
The latter would pose less of a problem with regard to importing such information into a locally used database.

**Improve the use of EAD in other contexts**

With regard to current developments in the field of providing online access to archival descriptions not only on one's own institution website but also in domain-specific, cross-domain or thematic projects and portals eg the Archives Portal Europe, aspects such as having EAD more interoperable for multilingual environments has been one aspect under consideration for the EAD revision. Furthermore, the TS-EAD intended to examine the question of either including or at least referencing user-generated content, providing guidelines on how to do this with EAD.

**Workflow for the revision process**

For all possible changes envisaged, the main directive has been to not only create and define a new version of EAD but to also define and ensure the migration from previous versions of EAD to this new one. As EAD has been used in the archives domain for nearly two decades, there are a lot of users of EAD worldwide for whom the new EAD shall provide additional and improved possibilities for encoding archival descriptions that would by far outweigh the challenge of moving from EAD 2002 to EAD3.

**Steps so far**

```
Image 1: Workflow from first call for comments to releasing EAD3 beta
```

As mentioned above, the revision process started with a call for comments in October 2010. By the deadline of February 2011, few more than 120 comments were received and grouped into five sections: schema and
documentation, general nature and purpose, relationships to related resources and entities, EAD header and archival description. The latter was furthermore structured into seven subsections dealing with questions and proposals related to hierarchical elements, wrapper elements, controlled access terms, digital archival objects, date encoding, languages and scripts and other elements.

The outcomes of the commenting phase were presented at the EAD Roundtable during the Annual Meeting of the SAA in Chicago in August 2011. Comments and further feedback gathered at this presentation were taken on at a working meeting of TS-EAD in March 2012, followed by several telephone conferences throughout the year and a work meeting of the Schema Development Team (SDT) of the SAA in October 2012. February 2013 then saw the alpha release of the new EAD and a working meeting of the Tag Library team for EAD based on this. A second commenting phase until April 2013 and further telephone and small circle face-to-face meetings led to the beta release of the new EAD in August 2013, right in time for the next Annual Meeting of the SAA in New Orleans.

After the beta release

The beta release once more was followed by a two-month feedback period for comments and a joint meeting of the TS-EAD with the Technical Subcommittee on EAC-CPF (TS-EAC-CPF) in New Orleans in August 2013 also finally produced a name for the new version: EAD3. In addition to the presentation of the current status and the then estimated timeline of the remaining tasks for completing the revision process, the joint meeting also discussed some pending issues such as licencing of the schema and the tag library, the use of the element group <relations> etc.

After that – and after the call for comments on the beta release had been closed – the TS-EAD and its subgroups continued working on reviewing and discussing all bugs and comments received from the EAD-community, the EAD mailing list and during the Tag Library meeting(s) in order to decide on solutions for pending issues and to prepare for the gamma release. With EAD3 gamma now being out for comments, the remaining timeline is concise with regard to finalising EAD3 and having it approved by the relevant committees of SAA for publication during the second half of 2014.

Image 2: Workflow for remaining steps

Going into the details: decisions taken so far

The decisions taken so far are numerous as could be expected after more than three years of work, detailed feedback and comments sent in by the EAD community along the way, including the beta release of EAD3.

They can be grouped as being:
decisions taken with regard to certain elements
decisions taken with regard to certain data types
decisions taken with regard to a reconciliation with EAC-CPF
decisions taken on general aspects
decisions taken on further aspects
decisions taken with regard to new elements and to elements recommended to no longer use

Decisions on certain elements

One of the elements that was discussed in more detail is the wrapper element `<c>` for the description of subcomponents of an archival unit such as series, files or items within a fonds or collection. In EAD 2002, this element can be used in numbered form with options `<c01>` to `<c12>`, thereby allowing a classification scheme of 13 levels including `<archdesc>` as the level describing the highest level unit in a certain setting. However, `<c>` can also be used unnumbered which would allow for an even more deeply structured hierarchy. As both options are used widely – and more or less in equal measure – deciding on one or the other options would have resulted in a major challenge for a good part of the EAD community when it comes to the question of moving from EAD 2002 to EAD3. Therefore, it has been decided that both will be maintained.

Another big topic during the revision process has been the use of indexing or controlled access elements such as `<persname>`, `<geogname>`, `<corpname>` etc and their use for referencing external national and international authority files for certain groups of terms. The alpha release in February 2013 tried to make a distinction between elements for encoding names eg:

```xml
<p>Letter of President</p>
<p>  <persname>Harry S Truman</persname>
</p>
<p>  to</p>
<p>  <persname>Winston Churchill</persname>
</p>
<p>  July 10, 1948</p>
```

and elements for referencing external vocabularies of names eg:

```xml
<controlaccess>
  <persnameentry identifier="http://viaf.org/viaf/61551827/#Truman_Harry_S__1884-1972" >
    <part localtype="full_name">Harry S. Truman</part>
  </persnameentry>
  <persnameentry="http://viaf.org/viaf/94507588/#Churchill_Winston__1874-1965">
    <part localtype="full_name">Winston Churchill</part>
  </persnameentry>
</controlaccess>
```

This however, proved to be another case of offering two elements for somewhat similar purposes so that the beta release reversed this and decided on general use of `<name>` elements in both cases. In any location within EAD3 where `<persname>`, `<geogname>` etc are provided, they now include a mandatory subelement `<part>` which would allow to e.g distinguish between first and last names via the attribute `@localtype` and allow for referencing external vocabularies.

With the gamma release in order

“to better support lightweight Linked Data applications, the `<controlaccess>` elements (persname, corpname, famname, name, subject, geogname, function, occupation, and title) were all given the following attributes:

```xml
@source [URI or string for name of vocabulary from which term is taken]
```
Decision on certain data types

The content model has been discussed for several elements during the revision, especially with regard to the possibilities of having structured/unstructured pairs for encoding certain data. This should on the one hand allow for more interoperability with databases while on the other hand, should enable more direct retrieval of certain information. The decisions taken in this respect refer to the following elements (among others):

- `<unitdate>` for encoding the date of creation of the archival unit described will get a counterpart in `<unitdatestructured>`. The latter will follow a similar structure as dates in EAC-CPF, providing subelements such as `<datesingle>` and `<daterange>` with `<fromdate>` and `<todate>` in order to encode single dates as well as date ranges. Furthermore, the element `<dateset>` will be available permitting various combinations of the aforementioned two elements. Both ways of encoding dates will be enabled to also encompass normalised date formats as `<unitdate>` already does in EAD 2002.

With the gamma release, "pattern validation for date normalization [sic] attributes was removed. A separate Schematron schema will be provided for validating date attributes against ISO 8601. Patterns were removed to allow the use of alternate date normalisation standards."[6]

- Similar to that, `<physdesc>` will now be accompanied by `<physdescstructured>`. While the first is meant to be used with plain textual information, the latter includes the already known – and some new – subelements for encoding the physical description of an archival unit. While the elements `<dimensions>` and `<physfacet>` remain more or less unchanged, `<extent>` and `<genreform>` as subelements of `<physdescstructured>` will both be replaced by a combination of the mandatory elements `<quantity>` (which can be left empty and – in EAD3 gamma – allows for the attribute `@approximate`) and `<unittype>.

Depending on whether one would intend encoding the extent of an archival unit i.e. the space this unit occupies within an institution’s repository or the genre or physical characteristics of the unit, the new element `<physdescstructured>` is to be typed as either “spaceoccupied” or “materialtype”. Several statements on the physical description i.e. two or more `<physdescstructured>` elements can furthermore be grouped with another new element: `<paralellphysdescset>`.

- Information about the language of the material can now be encoded with a more distinctive approach for languages and scripts. `<langmaterial>` permits concentration on the language information with the repeatable subelement `<language>` that now only includes an attribute for language encoding following ISO 639-2b, or to provide information on language(s) and script(s) present in the archival unit described. The latter option, combining the elements `<language>` and `<script>` in a new wrapper element `<languageset>`, also enables the user to encode cases where there may be several languages but just one script – or the other way round, such as:

```
<langmaterial>
  <language langcode="eng">English</language>
  <language langcode="fre">French</language>
  <language langcode="lat">Latin</language>
</langmaterial>
```

While `<langmaterial>` itself does not allow for content anymore, there is a new subelement `<descriptivenote>` which provides the option to give information about the language(s) and script(s) in a more illustrative way.
- The elements to encode data about the records creator (<origination>) and the institution holding and maintaining the material (<repository>) also have changed from their mixed content models to clearer data models, allowing for references to external sources and vocabularies providing additional information. Both elements now require a subelement out of the group of <corpname>, <famname>, <name>, and <persname> with the latter only being applicable for <origination>. The element <repository> furthermore includes the subelement <address> for contact details. The subelement <descriptivenote>, which had been enabled for both in the EAD3 beta, “was removed from <origination> and <repository> for lack of a use case”[7] in EAD3 gamma.

Decisions for reconciliation with EAC-CPF

Reconciliation with EAC-CPF has – apart from relatively small changes like the additional date encoding options for <unitdatestructured> as mentioned above – also brought the two probably biggest changes to EAD3. Firstly, the element <eadheader> for bibliographic and other information about the finding aid (printed and encoded versions) will be replaced by <control> for administrative data on the EAD file itself. Compared to EAC-CPF, EAD3 will however maintain the element <filedesc> within <control> for bibliographic metadata such as title of the finding aid, author, publisher, publication date and place. EAD3 gamma now also has “restored @langencoding, @scriptencoding, @dateencoding, @countryencoding, and @repositoryencoding to <control>. Each of those attributes may have either their default values from EAD 2002 or “other”. “[8]

Apart from <filedesc>, the <control> element encompasses various subelements for maintenance information related to the EAD instance. While EAC-CPF uses fixed lists of values for some of these elements, EAD3 “[moved] enumerated element values for <publicationstatus>, <maintenancestatus>, <agenttype>, and <eventtype> […] to a “value” attribute to allow for alternate languages to be used in the elements themselves.”[9]. This change also might be fed back to the TS-EAC-CPF for consideration.

Another addition coming from EAC-CPF is the element <relations> with its various subelements to encode relations between the archival unit described in the EAD instance and eg persons, corporate bodies and families described in an EAC-CPF instance as well as other resources.

“For EAD3 <relations> will be considered an "experimental" element, and not included in the official schema. A separate schema is available that includes <relations> for those EAD implementations that choose to use it. This arrangement is a compromise between the desire to provide the more robust Linked Data functionality introduced by <relations> in EAC-CPF and the concerns that <relations> may duplicate functionality already available within <controlaccess> and that the successful implementation of <relations> may be dependent upon as yet undeveloped external vocabularies.”[10]

Decisions on general aspects

As for general aspects of the revision, the question of spelling elements’ and attributes’ names and the issue of documentation shall be detailed here as two examples. The alpha release of the new EAD included a change in spelling towards „camel notation“ ie all elements and attributes which are comprised by two or more terms would have had the second and third of these starting with a capital letter. Eg <unitdate> would have become <unitDate>. Though this adheres to current best practices in XML definition and also was used when defining EAC-CPF, the history that EAD has contrary to eg EAC-CPF in terms of usage and user base, led to the decision to reverse that already in the beta release. Otherwise, the barrier for backwards compatibility might have been too high.

Another request received as feedback on the alpha release was to emphasise the documentation of changes a bit more; that is, to not only rely on the Tag Library for describing and summarising the changes made from...
EAD 2002 to EAD3, but to also document changes within the new schema itself. Beta and gamma releases therefore have both been published together with an updated Tag Library as a first step.

Decisions on further aspects

In order to better support the exchange of EAD instances especially in multilingual environments, all elements in EAD3 that can contain textual information allow for the attributes @lang and @script to be used. With these attributes the language and the script applying to an element's content can be encoded. Furthermore, the new element <foreign> has been introduced for inline tagging of language and script of only part of a text eg:

```xml
<unititle lang="eng">Letter by Alexander von Humboldt about being awarded with the
<foreign lang="ger">Orden</foreign>
<foreign lang="fra">Pour le Mérite</foreign>
<foreign lang="ger">für Wissenschaften und Künste</foreign></unititle>
```

With regard to the aim of clarifying how and where to use certain elements, EAD3 is furthermore un-nesting elements such as <arrangement> (only sibling, not subelement of <scopecontent> now), <legalstatus> (new sibling, not subelement of <accessrestrict> now) and <unitdate> (only sibling, not subelement of <unititle> now). Similarly, the different <note> elements have been renamed according to their context to eg <didnote> when being used as part of the descriptive identification <did>; <controlnote> when being used along with controlled vocabularies; <descriptivenote> when being used as additional descriptive information specifically in the <control> element; <footnote> when being used as such.

The last change to be mentioned here has been made between the beta and the gamma releases and refers to the element <daogrp> being one of the elements that are no longer recommended for use.

"In response to beta feedback requesting the restoration of <daogrp>, [...] <daoset> [has been created], which wraps two or more <dao> elements. This acknowledges the need to bind multiple <dao> elements, but still eliminates the burdensome and little used extended linking model of <daogrp>. Within <dao>, <daodesc> was replaced with <descriptivenote>. This was done for simplicity: <descriptivenote> was already available elsewhere in the schema and it allows one or more <p>, rather than <daodesc>, which allows <head>, <list>, <table>, and other block elements, and for which [...] nearly zero evidence of use [could be found. Furthermore] @coverage (with values "whole" or "part") [was added] to <dao>."[11]

Decisions on elements no longer recommended for use and on new elements

As in all revision processes of EAD so far, some new elements and attributes have been added and some of the current EAD 2002 elements will be recommended to not be used any longer. The first group includes additions such as:

@containerid with <container> to provide an identifier eg of the box or folder in which the material is housed
@instanceurl with <recordid> to include the URL of the EAD-XML itself
<representation> as sibling element of <recordid> to include the URL of any kind of representation (HTML, PDF, etc) of the EAD instance
<geographiccoordinates> as a subelement of <geogname> to allow the encoding of geographic coordinates of a place or region following encoding systems such as WGS84, OSGB36, or ED50

Among the elements that will be deprecated are <frontmatter> for formatted title page information, the grouping elements <eadgrp>, <archdescgrp>, <dscgrp>, <descgrp>, <daogrp> and <linkgrp> as well as <runner> and <note>. Furthermore, the subelements of <bibref> (ie <bibseries> and <imprint>) will not be recommended for...
use anymore.

**Delivery package for EAD3**

When EAD3 is published in the second half of 2014, the updated schema will be accompanied by a migration stylesheet for a general transformation of EAD 2002 to EAD3. Both of these documents will be made available under the [CC0 licence](http://creativecommons.org/publicdomain/zero/1.0/) while the Tag Library, as the third part of the publication, will adopt [CC-BY](http://creativecommons.org/licenses/by/4.0/). The latter will summarise the revision process and final outcomes, define elements and attributes of EAD3 and provide crosswalks to other standards as well as encoding examples. Additionally, Schematron schemas will be provided for use with certain normalised data as eg dates according to ISO 8601.

**Impact of the EAD revision on the Archives Portal Europe**

As a Best Network Project, APEx (Archives Portal Europe network of excellence) has been following the revision process closely. There are several project members who are also members of the TS-EAD, the TS-EAC-CPF and the Schema Development Team and the project's work package on standards and guidelines has been active in providing feedback on the alpha and beta releases so far. It remains to be seen when and to which extent EAD3 will be adopted by APEx and the Archives Portal Europe, as this is currently highly dependent on the final schedule for the revision process in correlation with the lifetime of the project.

**Use of EAD in the Archives Portal Europe**

A central EAD profile has been defined for the Archives Portal Europe called [apeEAD](http://apex-project.eu/index.php/articles/149-ead3-and-the-consequenc...). This profile has been based on a comparison of existing use and usage of EAD within the partner countries and institutions of APEx's predecessor, the APEnet project (2009-2012). It is being adapted along with other developments for the Archives Portal Europe ie with regard to new central functionalities or as a result of expanding the network of content providers and thereby the number of original data formats to deal with.

Aiming at central presentation and functionalities, apeEAD is used as a target format for conversion from local data formats and for validation, as a basis for the central presentation and search index and as a basis for further conversion eg to the Europeana Data Model (EDM) for exchange of archival data with the cross-domain portal [Europeana](http://apex-project.eu/index.php/articles/149-ead3-and-the-consequenc...).

With EAD playing such a central role in the system of the Archives Portal Europe for exchange and communication of archival descriptions, the impact analysis of moving to EAD3 must bear the following main aspects in mind:

- The advantages and opportunities offered by the new version need to be balanced with the workload resulting from its adaptation
- The possible impact not only on the central installation itself but also on the portal's content providers and national aggregators needs to be taken into account
Necessary changes and aspects to consider

Possibly the main challenge when moving from EAD 2002 to EAD3 will be the change from `<eadheader> <control>` as this to some extent affects the maintenance data that is used in the back-end of the Archives Portal Europe for data management. This specifically refers to the change from `<eadid>` as the identifier of the EAD instance to `<recordid>` as the addition of some mandatory elements in this part as eg `<maintenanceresources>` can probably be dealt with by using generally applicable default values.

Furthermore, some subelements of `<did>` have changed as mentioned above and so display and possibly search index would need to be adapted as well as perhaps some parts of the conversion stylesheets which have been developed. This refers eg to elements such as `<originatign>` and `<repository>` getting mandatory subelements or to `<physdesc>` and `<physdescstructured>`. With regard to `<unitdate>` it would first be necessary to evaluate the possible advantages of using its new structured version; otherwise, sticking with the current format of `<unitdate>` with @normal is most likely. Smaller adaptations such as `<note>` within `<did>` becoming `<didnote>` should not pose too big a problem; the same might go for the fact that `<extref>` and `<extptr>` are not used anymore and will have to be changed to `<ref>` and `<ptr>`. The siblings of `<did>` (eg `<scopecontent>`, `<bioghist>`, etc) remain mainly unchanged.

In any case, the interesting aspect for the Archives Portal Europe will be the possibilities in EAD3 for encoding multilingual content with the @lang and @script attributes having been introduced. Also the more structured relation(ship) encoding which is offered with the element `<relations>` might be a plus with regard to creating a more flexible corpus of different types of data within the Archives Portal Europe as well as relating to external resources. With regard to the possibility of a more structured encoding of certain data it remains to be seen in which use cases this can be of advantage for the Archives Portal Europe.
The options are many and there is a lot that could and can be done with EAD3 so that the APEX standards and guidelines work package will surely analyse the new version of EAD in more detail once it is officially adopted and published, not least defining a possible updated apeEAD based on EAD3 in preparation for the actual technical changes.

4. See information on EAD 2002 as well as on the previous version EAD 1.0 at http://www.loc.gov/ead/ (viewed February 2014).
5. TS-EAD co-chair Michael Rush, Yale University's Beinecke Rare Book and Manuscript Library, in his mail to the EAD mailing list (EAD@listserv.loc.gov) on 6 February 2014 about the release of EAD3 gamma.
6. Mail by TS-EAD co-chair Michael Rush, 6 February 2014, about the release of EAD3 gamma.
8. Idem.
10. Idem.
11. Idem.

Tags: apeEAD Archives Portal Europe EAD3 finding aids standards

About the Author

KERSTIN ARNOLD

Kerstin Arnold holds a Master each in Communication Studies and Information Management. She is currently doing her doctoral thesis next to the work at the Federal Archives of Germany for the Archives Portal Europe - network of excellence (APEX) project. She has been working at the Federal Archives in several projects on standardisation of (encoded) archival descriptions and online access to archival material. In the project’s context she is leading the work package on Interoperability and she acts as the APEX Technical Coordinator.

More articles from this author