Mike Ferrando
IT Specialist
Library of Congress
Washington, DC
7-4454

April 4, 2016

Critique and Analysis : EAD 3 Workflow Proposal

Introduction:

Abandoning the present EAD 2002 workflow to implement EAD 3 would be fatal to the progress and achievements of the current fidning aids EAD 2002 project (over 10 years of success). The proposal to replace the current EAD 2002 schema with the new EAD 3 schema has been termed an "Upgrade." However, EAD 3 schema offers enhancements that will not be used by divisions creating finding aids, contrary to best practices creating finding aids, require laborous and verbose tagging of container level bibliographic description, and will conflict with current LC guidelines and practices for metadata creation and use (searching) by current LC applications.

The Finding Aid Created EAD:

The Library of Congress uses the Society of American Archivist finding aid document to represent processed special collection materials. The Finding Aid document format captures general information as well as brief bibliographic description of these special collections materials. The Library of Congress guidlines and practices for creating these brief descriptions have undergone only slight changes since the 1990s (RDA, DACS, AARC3). For the greater majority of this time EAD 2002 has successfully been used to wrap and tag finding aid data. EAD was created specifically to capture and wrap the SAA finding aid document format. Only a change in the creation of finding aids (i.e., the detail and level of bibliographic description) here at the Library of Congress could justify any demand to disrupt the current EAD 2002 workflow replacing it with a different EAD tag set.

The Finding Aid Document Format:

The Finding Aid is a document format representing special collections materials. The Society of American Archivist has guidelines for creating these documents. The Library of Congress uses the SAA guidelines to create finding aids as guides to represent special collections. The document consists of branding of the institution, front matter (scope content, bibliographic information, etc.), administrative information (provenance, acquisitions, etc.), and the container list. The Container list presents the collection materials in groupings usually called "series" (i.e., Correspondence, Music, Diaries, Manuscripts, Journals, etc.). Each designated item is given a brief bibliographic description which is patterned after the bibliographic MARC format. The item level description is often considered "low level cataloging".

Chief cook and bottle washer:

I was hired by the Music Division in the Library of Congress in 1991 to process special collections. During that time, I was introduced to finding aids and bibliographic description. I worked with Ray White and other music specialists to create a procedures manual for creating finding aids for the Music Division (which other institutions used as in creating their own manual : University of Maryland College Park). I also was solely responsible for creating a centralized workflow for creating and compiling finding aids into a established format using Word Perfect. This word processor format became the basis a uniform published PDF presentation of the finding aids for the Music Division where I also created metadata (drawn from the MARC record and the Finding Aid) in the PDFs, internal table of contents linking, and registered handles for these WordPerfect generated PDFs. Google was quick to add these finding aids to their search results as the very first hit.

Not long after developing the Word Perfect to PDF workflow, I began to work with EAD 1.0 in 2001 with my first EAD coding class (1 week) at the Rare Book School under the training of Daniel Pitty. It was at that point I realized the need to create a conversion workflow for our word processing documents into EAD XML UTF-8. Shortly thereafter, in my study of XSLT/XML transformations (via Michael Kay's books) I was able to establish this workflow. At this time the Library of Congress (and for sometime afterwards) had no XML tools available for staff. Part of my experience was to find and test XML tools that were Unicode compliant. As a result I was able to successfully convert our finding aid word processor documents into XML UTF-8 capturing and retaining all diacritic characters. However, it was my work in EAD using XSLT/XML that enabled me to create and process finding aids quickly and efficiently into EAD. When the Library of Congress moved from EAD 1.0 to EAD 2002, I created a conversion toolkit that became a standard in the EAD community not only doing the conversion, but also generating a report of all converted elements with a brief description of line number and any resolved issues assumed by the stylesheet. This toolkit was made available on the Library of Congress web site for the EAD community. Moving finding aid into EAD 2002 was successful, yet there was no hard copy/publication format for the finding aids at that time. I began to develop a XSL-FO stylesheet to move the EAD XML into PDF via XSLT-FO. This took some time and required a proprietary rendered (XSL Formatter by Antenna House), but the purchase was made and now finding aids could be round-tripped from word processor, to XML, to PDF.

At this time the Library of Congress was using Enquiry to index and display finding aids for public search and use. Approaching the end of my time in the Music Division, I was asked to create a tutorial for converting word processing finding aids. I developed a tutorial on the Music Division staff pages to indicate how EAD tagging could be done and how word processing formats (list, table, etc) could be tagged in EAD. Upon my promotion to ITS, the move began from Enquery to eXist. The new open source application used xQuery scripting and Lucene index tokenizing (Java). I refactored the eXist web application code to conform to Security requirements and created the xQuery scripts to transform EAD into HTML actuated by the eXist web application. In sum, I have been part of very fabric of not only processing special collection materials, but also the creation of finding aids and their movement into EAD. My perspective, experience, and knowledge of the workflow is broad and extensive over this 15-20 year period. I am in a unique position to declare an opinion on the current proposal to move the EAD 2002 workflow into EAD 3. In my judgement, the move is a mistake and will be over all detrimental and negative. At best, there will be no discernible difference in the products and no benefit to the divisions and/coders. Moreover, as the chief cook and bottle washer, the burden of the conversion will be on me to implement resulting in a dramatic increase in my work starting point to square one.

The XML Schema Document Format:
The XML Schema is an international standard format and serves as a blueprint, rule based protocol, and set of requirements that dictate and control the structure of the data (hierarchical) as well as the content. These components are defined and constrained by Unicode compliance, controlled vocabulary, and syntax validation. the schema's ability to bind all of these entities into a functioning map creates a stable, strong and enduring foundation sttone as well as architectural framework upon which the data can be built. This has been demonstrated by the varied and wide range of EAD related applications created, built, and current in the EAD 2002 workflow today. In addition, the EAD 2002 Schema offers elements that can be mapped to MARC field and subfield values (EAD 2002 to MARC Crosswalk) which has been a guide for the Library of Congress EAD group to establish best practices for bibliographic description tagging.

EAD 2002 Schema has enabled the creation of toolkits (MARC to EAD Controlaccess elements, EAD to HTML Finding Aid Pages for Reading Room Web Sites, etc.), EAD XML creation/editor templates, EAD document validation, scripts, data typing of bibliographic information, PDF publishing events, RSS feed, XML metadata documents, interface with MARC records and data types, including the METS object as the final instance of the EAD/MARC XML object. These toolkits are used by all coders of EAD XML finding aids at the Library of Congress. Furthermore, all scripts and stylesheets are built on the EAD 2002 XML Schema blueprint that generate, process, and create all EAD related products (PDF, RSS, MARC to EAD conversion, RSS, etc.). The edifice is extensive and broad in scope covering all aspects of finding aids from creation to publication. This is all possible because we have invested in the strengths of the EAD 2002 Schema. Removing this schema will be devastating in the extreme, toppling the structure, and necessitate abandoning all current construction across the board from soup to nuts. But this is just the beginning, because individuals in all divisions will need to create new workflows and require training to use the new schema, EAD 3. The prospect of dual workflows and multi tier document file systems (EAD 2002, and EAD 3) has virtually no willing adherents in the Library of Congress coding community. I once expressed the wish that EAD 3 would just go away and was quickly affirmed by a long time group member processing specialist. It is more than obvious that the impetus for abandoning the EAD 2002 workflow (which took so much work to implement and establish) has nothing to do with every day work of capturing finding aid information in the EAD tag set.

The EAD 2002 Schema workflow has not been idle, but rather using the foundation of EAD 2002 Schema we have been able to build and expand the use of EAD finding aid bibliographic information into many new areas. One of the most common areas is the generation of browse pages for the EAD finding aids. These browse pages are updated with the production updates every month. Marla Banks has the sole responsibility of maintaining and generating this workflow which uses EAD and MARC (found in the METS object) to generate these browse pages. Another common part of the EAD 2002 Schema strength is the required validation of date syntax using ISO8601 standard. The current EAD 2002 Schema allows only the Gregorian calendar and date normalization must be valid ISO8601 format. Marla generates dates browse lists based on these values in the EAD finding aids enabling finding aid materials to be browsed by date.

The strength of date information in EAD 2002 is also used in generating metadata for Project one brief pages. One of the major projects EAD 2002 workflow has created is the Project One automated digital object linking from finding aids to Project one pages (offering grouping images of the digital items in the finding aid using a page turner). The EAD 2002 captures date information in the UNITDATE element which requires a Gregorian calendar and also offers ISO8601 date normalization. ISO8601 is a standard that enables the parsing of date information in the finding aid in any UNITDATE element. The real strength of EAD 2002 UNITDATE is that the EAD 2002 Schema validates the document when it is created. Thus at the lowest level of the workflow the choice to use dates can be checked. Project One uses XML Document tag set to populate the content and provide metadata for searching and display of any digitized object. EAD 2002 Schema has enabled us to create a workflow that can generate these xml documents from the finding aid. Furthermore, this workflow is also able to populate the finding aid with links to the digital item (relieving the coders from the tedious task of doing them by hand) and generating handles that can be registered for each link (from the finding aid to the Project one pages, and from the Project One pages to the item in the Container list via the eXist web application). Moving to EAD 3 would demolish this workflow.

One of the supposed features of EAD 3 is the ability to use international calendars. The EAD 3 Schema UNITDATE element has been stripped of all required validation and the ISO8601 standard has been completely removed from the element. These values are now simply free text with no validation of the document at any level. It is difficult to describe how catastrophic this will be for our data and our scripts and code. It will be necessary to write consider check code to process these elements and necessitate a standalone validation for the document unavailable to the coders creating the finding aid. It is simply inconceivable that EAD 3 Schema would strip the UNITDATE element of these requirements and standards and not even provide some kind of controlled vocabulary (of calendar names).

EAD 3 presents another issue with allowing international dates. Current Library of Congress finding aid creation discourages using other calendars when giving dates (except as a note). Finding aid materials and their contents are converted to Gregorian dates only when expressed in the finding aid container list and elsewhere. Moreover, MARC Cataloging practice follows the same guidelines by converting all dates to Gregorian calendar, but allowing a free text field to express the given date (i.e., 245$f, 245$g, 260$c, 264$c). Although the MARC record offers fields that allow for dates, the most critical field allows only 2 options, no date (19uu, uuuu), or a Gregorian date. In addition to this, the Voyager system, reads only the 008 position (it is possible to have 2 dates representing a span in the 008) to determine date sorting and searching for the record. Finally, Oracle tables date capabilities require Gregorian dates (dates that have time zones). Therefore, EAD 3 UNITDATE element will be a dramatic and detrimental loss to the current level of data we possess for dates in our finding aid information and encourages a practice contrary to overall practices at the Library of Congress for capturing date information. There is simply no system at the Library of Congress that uses international dates.

EAD 3 also requires verbose tagging of CONTROLACCESS information. The current EAD 2002 workflow uses the MARC bibliographic data for populating these elements (using a toolkit to keep it updated, reduce the coding time, labor, and possible human error). The controlaccess elements are populated with access point fields from the MARC record (1XX, 2XX, 6XX, 7XX, etc) through XSLT transformation stylesheets (I created). These fields often must be filtered and arranged to present the information in the EAD finding aid that will best serve the user (in accord with Voyager display practices). This information is displayed as text to the user in the current finding aids. In the eXist finding aid application the MARC record is also used to generate index terms as live searches to the Voyager catalog. EAD 3 CONTROLACCESS element has changed the EAD tagging for this information requiring new elements for every possible subfield that is retained from the MARC record. Thus EAD 3 assumes that users will want to capture this MARC field information to reconstruct or harvest this portion of the XML finding aid for other purposes. But this assumption is again, contrary to how finding aids created in the Library of Congress. The Library of Congress special collections materials are presented in more than one representation. Finding aids are only one of those representations. Voyager catalog and other systems capture this type of high level cataloging information for possible search and processing that is completely apart from the finding aid. EAD 3 effort to capture these subfields in new elements in the CONTROLACCESS elements would greatly add to our workload and require verbose coding of these elements. Moreover, CONTROLACCESS elements can be used in other areas of the EAD finding aid. Thus, the necessity for coders to attempt to add these elements in those places as well. EAD 3 insistence on adding these elements to the CONTROLACCESS elements is completely contrary to the processing and use of special collection materials at the Library of Congress and would also burden our workflow with unnecessary and unwanted verbose coding.

Contributors & Additional Information

EAD 3 & UNITDATE @normal Attribute Change:
What's the difference between @standarddatetime, @standarddate, @notbefore, @notafter, @normal for dates?
EAD3 refines dates and has added the new attributes for information about complex cases. @normal remains available within <date> as a way to present a normalized form of the date or date range but does not address cases where only partial information about a date is known. @standarddate, @notbefore, and @notafter are available within <todate>, <fromdate>, and <datesingle> to record either a specific date or dates which can help in placing it. A photograph taken during a childhood summer, for example, might be @notbefore 1970-05-01 and @notafter 1970-09-07. A letter might be @notafter 1871. Ideally, data recorded in any of these elements would conform to ISO 8601, e.g. YYYY/YYYY-MM/YYYY-MM-DD, although it is not constrained by the schema and allows for non-Gregorian calendar use.
http://www2.archivists.org/groups/encoded-archival-description-ead-roundtable/frequently-asked-questions-about-ead-and-ead3#.VwKACV5cCVo

EAD Group Technical Documents:
http://www.loc.gov/staff/rr/ead/

EAD Toolkits:
http://findingaids.loc.gov/search/rr_fa_html_page_query.html
http://findingaids.loc.gov/search/lccn_marcslim2ead_query.html
http://findingaids.loc.gov/search/rss_search_query.html

EAD Processed Finding Aids Monthly Stats:
http://rs5.loc.gov/xmlcommon/eadwork/mfer_files/all_ead_report.html

EAD eXist Finding Aids Web Application:
http://findingaids.loc.gov

EAD eXist Finding Aid Browse Pages:
http://findingaids.loc.gov/browse/collections/a
http://findingaids.loc.gov/browse/dates/main
http://findingaids.loc.gov/browse/locations/main
http://findingaids.loc.gov/browse/names/a
http://findingaids.loc.gov/browse/titles/a
http://findingaids.loc.gov/browse/subjects/a

EAD Tutorial:
http://www.loc.gov/staff/rr/perform/mfer_pkmn/pkmncntr/kanto/viridiancity/ead_slices_tutorial/parr.st.10011002.html

Music Division Finding Aids Processing Manual:
http://www.loc.gov/staff/rr/perform/procmanual.pdf

EAD 2002 UNITDATE Element specs:
https://www.loc.gov/ead/tglib/elements/unitdate.html

Staff Contributers:
Marla Banks (EAD Group : EAD Browse Pages Developer)
Digital Media Project Coordinator
OCOO/CIO/WEB

Colleen R. Cahill (Cataloging : Non Gregorian Dates)
Digital Conversion Coordinator
Geography & Map Division

Bennett Heggestad (Cataloging : Non Gregorian Dates)
Sr Cataloging Special
Manuscript Division

DeAnna Evans (MARC Standards Documentation & Standards)
Digital Projects Coordinator
LS/ABA/NDMSO

Ardith Bausenbach (Voyager : MARC Standards)
Information Technology Specialist
ILS Office

David Bucknum (Voyager : MARC Standards)
Digital Projects Coordinator
ILS Office

Tracy Chen (Head Oracle DBA Supervisor)
Supervisory Info Tech Specialist
OCOO/CIO/DBA

Raymond White
Music Specialist / Librarian
Acquisitions & Processing Section, Music Division

Kate Rivers
Music Librarian
Acquisitions & Processing Section, Music Division