Harvard University EAD revision submissions (final version, Feb. 25, 2011)

Brief description of change	Rationale
system identifiers • **optional** • **repeatable**	The purpose of this revision is to provide a place in EAD to record identifiers applied to the EAD record in various systems. This is roughly analogous to the MARC 035 field: control number of a system other than the one whose control number is considered primary is a given context. In EAD, the EADID is the primary identifier, but as the finding aid moves among systems over time, it may acquire additional identifiers that are useful to track. Uses include record migrations or aggregator databases. Attributes allow for either use of a resolvable identifier or any other type, such as an explicit expression of a system name. Examples: <sysid type ="other" othertype="MH-ATprod"> 12345</sysid> *for an Archivists' Toolkit resource record number from which the EAD was derived* <sysid type="urn">http://nrs.harvard.edu/urn-4:12345</sysid> *for the system number of the EAD in an EAD aggregating database*

Brief description of change	Rationale
external metadata reference • **optional** • **repeatable**	The purpose of this revision is to provide a place in EAD to identify/link to related metadata. Such information would include but not necessarily be limited to information about the described collection, about any described component of the collection, or about concepts, names or vocabulary terms expressed anywhere in the EAD. Examples might include: • a collection-level MARC record that describes the same materials described in the EAD document • an analytic description in an image catalog for a photograph that is also described as a component of a collection in EAD • an EAC record for a name Uses include interoperative indexing and display for end-users, automated updates to EAD text of controlled vocabulary terms, and linkages between EAD and the related metadata records. This should be valid in <ead>, <c> and various elements that can hold controlled terms: <name>, <corpname>, <famname>, <persname>, <geogname>, <subject>, <formgenre> It should be possible to record the type of metadata and the type of location identifier supplied, controlled or uncontrolled for both the type and the location. In EAD, this could be accomplished using a single set of attributes applicable at many points in the schema, or a mix of mechanisms if that would better fit the existing EAD element structure. For example, PREMIS and METS both use a mechanism with paired controlled attributes and uncontrolled attributes, i.e. -- LOCTYPE, with controlled values such as ARK, URL, URN and OTHER; and OTHERLOCTYPE to specify another kind of location identifier, such as a specific local system. -- MDTYPE, with values such as MARC, DDI, FGDC and OTHER; and OTHERMDTYPE to specify metadata not from the enumerated list. MODS permits linking to related authority metadata by using @authorityURI and @valueURI attributes to identify the source of the vocabulary and the applicable record within the source. Possible examples: • <mdref mdtype="marc21" loctype="URN"> http://hollis.harvard.edu/?itemid=|library/m/aleph|009504960</mdref> • <mdref mdtype="other" othermdtype="OLIVIA" loctype="otherloctype" otherloctype="OLIVIAwork">416461</mdref> • <mdref mdtype="marc21" loctype="other" otherloctype="viaf"> http://errol.oclc.org/laf/n 86140996.html</mdref> • <mdref mdtype="marc21" loctype="other" otherloctype="lcnaf"> no2005088392</mdref> • <mdref mdtype="eac" loctype="url"> https://wiki.nla.gov.au/download/attachments/32189/oliphant-BrSp.xml?version=1&modificationDate=1201813806000</mdref>

Brief description of change	Rationale
digital archival object identifier • optional • repeatable	The purpose of this revision is to provide a place in EAD for the system identifier of a digital object. This data would support management of digital objects by facilitating data sharing between systems describing digital objects and systems storing digital objects. Unlike the current digital archival object links (dao href attribute), this would not be a reference to a delivery version of a digital archival object, but would instead be an identifier of an archival object in its host system. The identifier should be repeatable to allow for tracking a local repository copy and a copy in a shared repository, e.g. HATHI. Uses include cascading updates made in EAD to the digital object's host system metadata. For example, a title change in EAD could be sent to update a METS file in the digital object's host system. The proper location for this new element would be as a new sub-element or set of attributes on the <dao> and <daogrp>. A possible name for this element is the "Digital Archival Object Identifier." Given that a digital object may be described in an EAD instance, but not yet available for research, if this element is made valid for the <dao> and/or <daogrp>, then a <dao> lacking a link, but with a <doaid>, must be valid. In fact, implementing this enhancement may be especially valuable to contemporary archivists who face the influx of hybrid paper and electronic collections. Archivists can currently label and box analog collections and control their identifiers in EAD regardless of their restriction status. Controlling the whereabouts of virtual materials requires a new type of identifier. Identifiers might conform to established schemes such as ARK, URN, or PURL, or may be local. A "type" attribute would be needed to specify. One type would require Another type would be a resolvable identifier that leads to a name resolution system that interprets an URN and returns a system identifier and a system number. Examples: <daoid type="other" othertype="Cornell University eCommons">12345</daoid> <daoid type="urn">http://nrs.harvard.edu/urn-4:hul.arch:12345</daoid>

Brief description of change	Rationale
increase encoding granularity in <controlaccess>	The purpose of this revision is to enhance the compatibility and functionality of controlled access terms. Note: This suggestion has an impact on two kinds of <controlaccess> sub-elements: topics and names. Names are addressed in a separate revision proposal due to their validity in contexts other than <controlaccess>. The conversation of <controlaccess> granularity may not be possible without also considering the granularity of encoding of names. In pre-coordinated subject strings where part of a subject string is may be a topic, genre, time period, or a name, the various pieces are undifferentiated, leaving archivists pondering which to prefer. For example, an archivist may wish to convey that records of a business concern cobblers in turn-of-the-20th-century Carlisle, Pennsylvania, or that a personal papers creator wrote extensively about Thomas Jefferson's religious views. Further, many archivists wish to convey these notions in a form compatible with LCSH, such as "Shoemakers--Pennsylvania--Carlisle--1900-1910"or "Jefferson, Thomas, 1743-1826 -- Religion." In the present EAD encoding, an archivist must choose whether to prefer the topical, geographic, or name aspects of such concepts, and has no recourse at all with the temporal characteristics. In MODS, the name, topic, temporal, and hierarchical geographic sub-elements of these subject headings would have their own sub-elements. See: http://www.loc.gov/standards/mods/mods-outline.html#subject To assure backwards compatibility, EAD could permit the inclusion of sub-elements in <subject> within <controlaccess>, while not requiring these sub-elements.

Brief description of change	Rationale
increase encoding granularity in <name>, <corpname>, and <persname> elements	The purpose of this revision is to enhance the compatibility and functionality of name elements. Note: The granularity of encoding names has a companion question in the granularity of encoding of <controlaccess> terms. In EAD, name parts are not differentiated, thus the textual parts of an authorized form of a name and date or other portions are encoded in a single string that cannot be parsed by computers. This makes EAD name data incompatible with other standard expressions of names, in which the parts of an authorized name are carefully parsed. Both MODS and EAC present models of encoding names in which the various parts can be parsed. See: http://www.loc.gov/standards/mods/mods-outline.html#name and http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#d1e5077 To assure backwards compatibility, EAD could permit the inclusion of sub-elements in <name> within <controlaccess>, while not requiring these sub-elements. An additional question may be whether it would be necessary to deprecate <persname>, <corpname>, and/or <geogname> and exactly how to do so.

Brief description of change	Rationale
file size Addition of a place to record the file size of EAD instance • optional • non-repeatable	The purpose of this information is to improve the researcher experience when interacting with large finding aids. Researches can experience slow load times when they attempt to view large finding aids. To allay researcher frustration, it can be useful to provide a warning about potentially slow downloads or to indicate file size information before researcher elects to view a finding aid (i.e. in search results). Calculating files sizes in real time for all finding aids in a large result set can slow display of search results beyond reader tolerance. Encoding file size information within EAD allows for its display and thus it can alleviate user frustration. If other institutions are using a mechanism to provide this information to researchers, it makes sense to encode the data in a standardized way across the archival community. Harvard University has implemented this through a local change to the EAD schema in order to store the file size of EAD instances for display in search results. We have created a local version of the EAD schema which has an additional attribute, fileSize, which is valid for the ead element. For example, <ead fileSize="9910">. This attribute and its value are machine-generated when an EAD file is ingested by OASIS, our EAD catalog. Our research revealed that the styled html we derive from our EAD is roughly the same size. We suggest that this should be an optional, non-repeatable piece of information. We recognize that if the subcommittee is willing to include this information in EAD, an attribute on the EAD element is not the only possible location, a new element or attribute may be better suited to this purpose. One option may be to allow extent within filedesc, and include an attribute to indicate the unit of measurement. For example, "<filedesc><extent unit="megabytes">9.6</extent> ... </filedesc>". Indeed, since the EAD filedesc is patterned after the TEI fileDesc element, which includes an <extent> sub-element, the subcommittee may want to revisit this part of the TEI standard: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-extent.html

Brief description of change	Rationale
digital archival content flag Addition of a place to flag for the presence of <dao>s in the EAD instance • optional • non-repeatable	The purpose of this flag is to support functionality in discovery systems for finding aids that link to archival digital content. Researches often want to find only those EAD finding aids that contain links to online content. Calculating whether or not such content is present based on polling EAD instances for the presence of a <dao> in real time for all finding aids in a large result set can slow display of search results beyond reader tolerance. Flagging for the presence or absence of digital archival objects allows for faster retrieval of the desired result set. This flag can also be used to "brighten" an interface option (a button or tab on which a user clicks) for an alternative view of the finding aid in which the digital objects are prominently featured. Other repositories and implementations may have uses for such a flag. If other institutions are using a similar mechanism, it makes sense to encode the data in a standardized way across the archival community. Harvard University has implemented this through a local change to the EAD schema by creating an attribute, digitialLinks, on the <ead> element. This attribute and its value are machine-generated when an EAD file is ingested by OASIS, our EAD catalog. We suggest that this should be an optional, non-repeatable piece of information. We recognize that if the subcommittee is willing to include this information in EAD, an attribute on the EAD element is not the only possible location, a new element or attribute may be better suited to this purpose.

Brief description of change	Rationale
begin year and **end year** Additional attributes for <unitdate>s • optional • non-repeatable	The purpose of these attributes is to facilitate indexing of date information in discovery systems in order to support date range searching. For some systems, parsing the "normal" attributes of <unitdate>s for date range searches can slow the delivery of search results beyond reader tolerance. In other systems-- those in which EAD instances are indexed along with MARC data-- the availability of simple, separately parsed 4-digit dates that are compatible with MARC 008 date1 and date2 is of great value. Other repositories and implementations may have uses for 4-digit dates. If other institutions are using or in need of similar data, it makes sense to encode the data in a standardized way across the archival community. Harvard University has implemented this through a local change to the EAD schema, via the attributes beginYear and endYear on the <unitdate> element. This attribute and its value are machine-generated when an EAD file is ingested by OASIS, our EAD catalog. We suggest that these should be an optional, non-repeatable attributes on <unitdate>. We recognize that if the subcommittee is willing to include this information in EAD, it may determine that different attribute names are desirable.

Brief description of change	Rationale
permit simplified encoding of indexes	The purpose of this revision is to permit repositories to reduce the amount of time and effort required to mark up indexes by eliminating some of the currently required sub-elements. It may be that <list> was originally conceived as a mechanism to produce such simplified encoding within <index>, but an <index> that includes only <list> and <item> sub-elements will not validate against the schema. This may be merely a bug in the schema, but it may also be intentional. The following is an example of what seems to be the least complex valid combination of sub-elements of <index>: <indexentry><persname>Mary Smith</persname> <ref target="indexEntry1">Folder 5</ref> </indexentry>. That encoding, however, also requires insertion of the value of the target attribute in the ID attribute value of the targeted element. Such targeting is extremely time-consuming. What is desired is encoding within <index> as simple as one of the examples below or some other combination of tags that the Technical Subcommittee may determine. Possible examples: • <indexentry>Mary Smith, Folder 5</indexentry> • <list><item>Mary Smith, Folder 5 </item> </list> • <list><defitem><label>Mary Smith</label> <item>Folder 5</item></defitem>

Brief description of change	Rationale
expand valid options for locating <dao>s	At least one archival repository at Harvard is doing extensive on-demand digitization of individual items in folders. The primary access point for these digital copies is the EAD-encoded finding aid, many of which were created in the early days of EAD before digitization was common. The finding aids in question contain <list>s of <item>s, rather than nested <c>s below the folder level. It would be advantageous if <dao>s were valid within <item>s. If other institutions are facing the same issue, it makes sense for the EAD community to discuss options.