REEA - Building an Encoded Archival Description (EAD) generator for the HDMS

Aim

The aim of this project is to build an EAD Finding aid generator for the HDMS, and to report on application of EAD in an Australian context.

Introducing EAD

'Cutting edge' is not a term one usually associates with archivists, but the development of the Encoded Archival Description standard is an important initiative that certainly gives the archival profession a head start as the next generation XML Web starts to take shape. One has only to look at the W3C website (http://www.w3c.org/) to see a range of other professions developing 'EAD-type' frameworks for their information.

For the background to the development of EAD and information on other EAD projects see:

- Pitti, Daniel, 'Encoded Archival Description An Introduction and Overview', DLIB Magazine, November 1999, vol 5 no 11, http://www.dlib.org/dlib/november99/11pitti.html
- The Official EAD web site at http://lcweb.loc.gov/ead/
- National Library of Australia Encoded Archival Description projects at http://www.nla.gov.au/initiatives/ead/
- *The American Archivist* EAD editions, Encoded Archival Description Part 1-Context and Theory, September 1997, vol 60 no 3; Encoded Archival Description Part 2-Case Studies, Fall 1997, vol 60 no 4.
- EAD in Action: Applications of the Encoded Archival Description, *Archives and Museum Informatics* Special Edition, vol 12, nos 3-4, 1998.
- Join the EAD listserv for the latest discussions see 'How to Subscribe' at http://lcweb.loc.gov/ead/eadlist.html

For information on XML developments see:

- W3C The World Wide Web Consortium XML pages at http://www.w3.org/XML/
- *The XML Cover Pages* at http://www.oasis-open.org/cover/
- Microsoft, http://www.microsoft.com for information about XML support in Microsoft products and some basic XML introductory information.

The files making up the EAD Document Type Definition (DTD) 1.0 are available for download from http://lcweb.loc.gov/ead/eadv1ann.html#whattodo. Also available at http://lcweb.loc.gov/ead/tglib/tlhome.html is the EAD DTD Version 1.0 Tag Library. The SAA publishes a print version of the tag library and also *Encoded Archival Description Application Guidelines* which include EAD to ISAD(G) and Dublin Core metadata mappings. Unfortunately there are no plans at the moment to make these application guidelines available on the Web.[1]

⇨ Should an Australian distributor of the publications be encouraged? Should we lobby the SAA for the application guidelines to be made available on the Web?

A small modification of the SGML EAD DTD is needed to make it an XML DTD. The technical details are included in comments within the ead.dtd file itself.

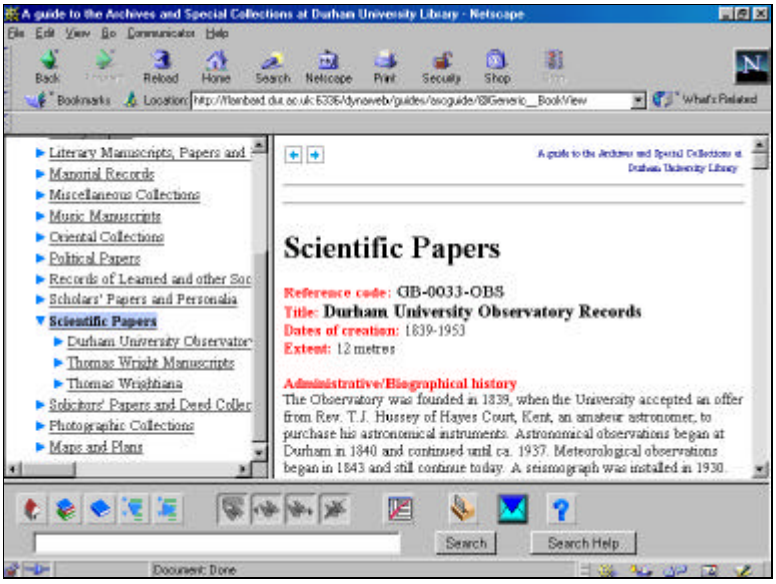
⇨ It would be helpful to have this information spelled out on the download page and in less technical terms.
⇨ Alternatively are there any plans to provide an official XML version for download?

[1] EAD list serv, Crosswalk EAD, DC and MARC Topic, 6-7 June 2000

EAD and SGML

The majority of EAD projects as detailed on the official website are SGML based. Most involve the conversion of existing finding aids through manipulation of digital versions in text processing software. Macros are devised to transfer the existing text to EAD, with subsequent manual verification and correction. The macros are specific to the content and structure of the legacy finding aids being processed. Other projects have involved rekeying, in particular where digital versions of the existing finding aids are not available, and some have involved the creation of EAD finding aids from scratch.

The SGML documents are delivered over the Web using SGML database software like DynaText/DynaWeb. The software manages a database of the SGML documents and supplies HTML versions created 'on the fly' to browsers according to style sheets.



Example of DynaWeb interface for Durham University Library EAD finding aids

As you can see from these examples, quite sophisticated display and searching is possible, although it does require technical expertise to manage the server, the database and create the stylesheets. The purchase and implementation costs of proprietary SGML/XML database web serving software are also significant.

EAD and XML

XML has the potential to be a more cost effective alternative. XML is a subset of SGML. It has less of the complexity of SGML, whilst still retaining most of the functionality. Hence XML software - parsers, authoring tools, processing tools etc - are reportedly much easier for the software developers to develop.

However XML is still immature. The core XML Syntax was only finalized in February 1998, XSL in November 1999, with XLINK, XPATH, etc. still being worked through by W3C. It means that software developers can only work with the latest draft of the standard and then be prepared to adjust as the standard is worked through to acceptance. As a critical example, XSL support in Microsoft's Internet Explorer 5 (IE5) is based on the working draft, http://www.w3.org/TR/WD-xsl, whereas later products, e.g. James Clark's vt.exe, support the finalized version, http://www.w3.org/1999/XSL/Transform.

The other side of the immaturity is in the degree and direction of sophistication. As we have found in the development of our database systems, you need a critical mass of real world applications to both

determine and prioritise functionality. You also cannot build the bells and whistles, until the core structures and functions have been through vigorous testing and use. Keep in mind this immaturity and remember what is was like to use earlier versions of the software tools you now take for granted!

With browser developers moving towards XML support, XML documents can be viewed like HTML. The browser dynamically transforms the XML to HTML for display in the browser viewer according to a stylesheet, i.e. XSL or CSS. The figures below show the rendering of an XML document in IE5, with and without a XSL stylesheet. Viewing the source displays the XML content.

XML document in IE5 with stylesheet

XML document without stylesheet

Alternatively XML documents can be transformed into static HTML pages, according to a stylesheet, using a product like James Clark's vt.exe. The HTML document produced can then be uploaded to a web server and viewed as usual through any browser. Note that the HTML document is not an EAD document, the structural tags have been transformed to HTML presentational tags.

The EAD project at Cornell University (see http://cidc.library.cornell.edu/xml/) illustrates both these methods, delivering a XML document for browsers with XML support and a HTML document for browsers that don't support XML.

Heritage Documentation Management System (HDMS)

The *Heritage Documentation Management System* provides a suite of tools for an archivist to process and manage any collection or grouping of records. It has evolved from Austehc's[2], processing of archival collections, referencing available standards and guidelines e.g. CRS, ISAD(G), ISAAR(CPF) etc., to determine structure and functionality. It is a relational database system built on a Microsoft Access platform.

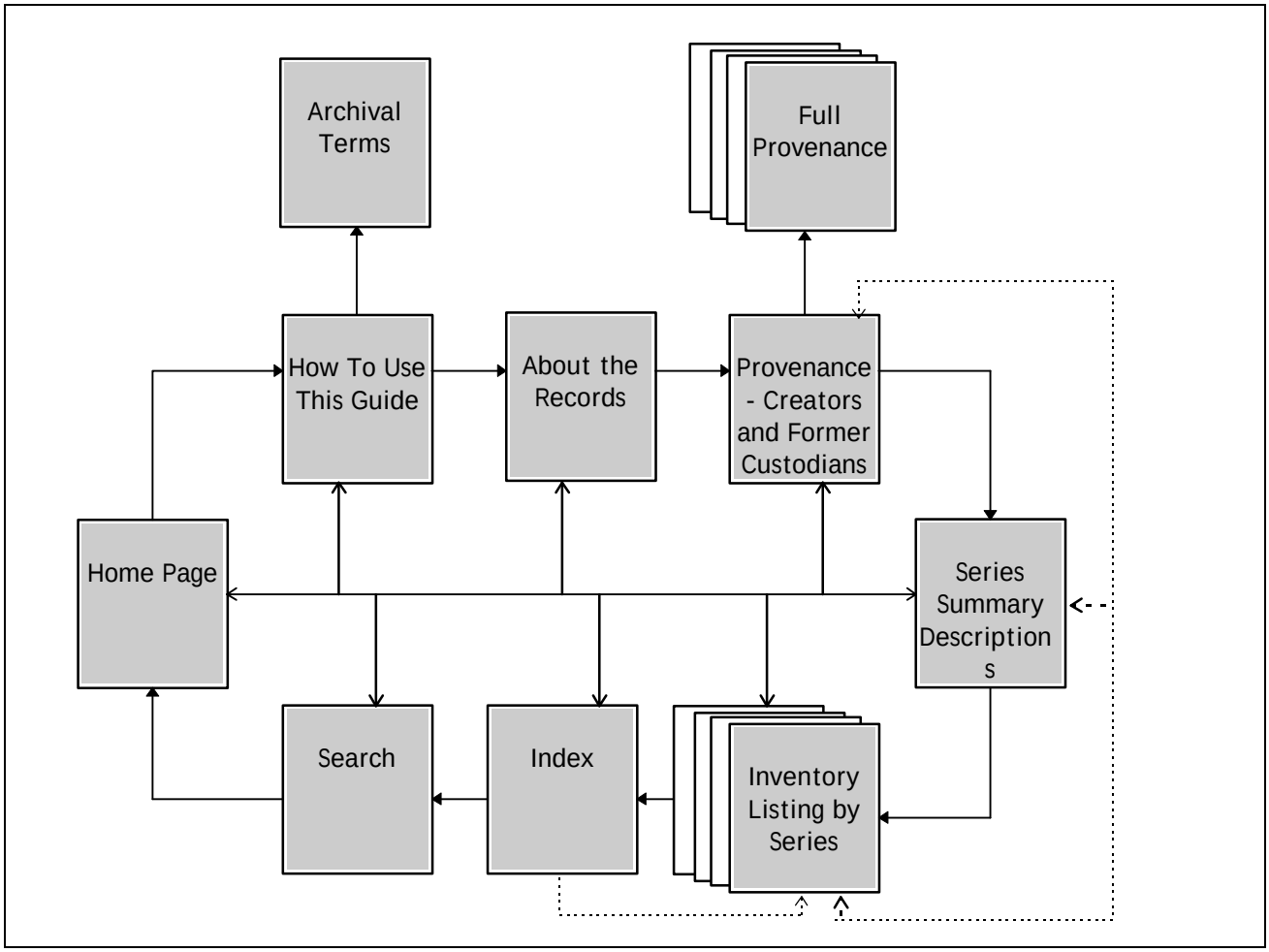
Printed finding aids, and more recently an HTML finding aid, can be produced from the archival documentation and administrative information held in the HDMS. The finding aids are focused on providing researchers with information about a 'collection' of records and their context. Not all the information held in the HDMS is output to the finding aid. The idea is that a researcher uses a finding aid to identify the records they wish to access, and then through quoting of the appropriate reference number, i.e. Inventory or other ID, the archivist managing the collection can use the HDMS database itself to locate the records, provide them to the researcher and record their usage.

[2] Note Austehc was formerly Australian Science Archives Project (ASAP) which has been involved with the processing of records of Australian science and technology since 1985. Information about the development of the HDMS can be found at http://www.austehc.unimelb.edu.au/HDMS/

In developing the HDMS in general, and the finding aid outputs in particular, Austehc has confronted the same sorts of issues that go into applying EAD, i.e. What data should be captured for each unit of description? What other elements make up a 'good' finding aid? How should these be assembled?

HDMS HTML Finding Aid

The HTML finding aid generator produces a multi-page HTML finding aid, which includes links between series, provenance, inventory, related series, and related provenance. An index and photo gallery pages may also be added. A wizard takes you through a series of steps to specify finding aid elements, like title, publisher, web address, etc., and html elements, like names and location of logos and other images. The pages are then generated in a specified directory from which they can be uploaded to a web server.



Examples:-

- *Philip Garth Law Guide to Records* at http://www.austehc.unimelb.edu.au/guides/lawp/LAWP.htm
- *Ian William Wark Guide to Records* at http://www.austehc.unimelb.edu.au/guides/wark/WARK.htm

HDMS EAD

In an ideal world this project would have looked at replacing the HTML output with EAD output, and writing the style sheets to transform or view the EAD output into the desired HTML. However, the technology in the real world is just not at a stage to support this. The technical concerns are

- File size - Philip Garth Law Guide to Records as an EAD document is 897KB, compared with file sizes from 5 KB to 160KB for the pages that make up the HTML guide.
- XML support only in IE5 - access to the finding aids would be impossible for users with earlier versions or alternative browsers hence you would need to provide HTML transformed pages as well.
- XSL - support of different versions has been identified above. In addition XSL at the moment is very difficult for a non programmer to use.
- EAD - would need stylesheet 'work arounds' to deal with provenance structure, multi-provenance, related provenance links and related series links.

In this context, the aim of the EAD generator for the HDMS is to provide a basic SGML or XML EAD finding aid. The SGML finding aid produced may be uploaded into an appropriate SGML viewer, and the XML finding aid can be viewed, via an XSL stylesheet in an XML enabled browser, i.e. IE5 . Thus, not all fields in the HDMS are output to EAD, nor are all the possible mappings of HDMS fields to EAD used. The aim is to strike an appropriate balance, and produce a structured finding aid, that is not 'overloaded' with data, nor is 'too large' to be delivered over the web.

The document *HDMS Application Guidelines* details the mapping of HDMS fields to EAD elements.

Finding Aid Generator

Developing the EAD finding aid has lead to some refinement of the HTML finding aid, as well as verifying its structure and content. The HTML Finding Aid Wizard has been rebuilt as a more general finding aid wizard which can lead to the production of one or more of an HTML, XML EAD, and SGML EAD finding aid.

The document *HDMS Finding Aids* shows how to use the new HDMS Finding Aid wizard. It will be incorporated in the online help and *HDMS The Essentials* manual as quickly as possible.

Issues and Future Developments

1. Provenance and <bioghist>

From the perspective of the HDMS, <bioghist> gives a very primitive representation of provenance entities. Not having the structure to 'title' and 'date' the provenance entity is of particular concern. We have used generic title fields in order to title each provenance entity and also made assumptions in the XSL style sheet to display the born/died, established/ceased date in a ‘unitdate’ way.

The ability to express provenance entities in a structured way and also to express relationships amongst provenance entities is critical for Australian practice.

Þ *Recommendation: To progress the ISAAR(CPF) DTD initiative*

2. Provenance and <chronlist>

The <chronlist> element within the <bioghist> element does have a lot to offer the HDMS. Austehc's style for biographical/historical notes in finding aids has been to record 'CV' style information rather than narratives. Our development of the OHRM (Online Heritage Resource Manager) has led to the creation of an 'Events' table to allow this CV style of data to captured in fields, i.e. dates, location, description, source etc. A further development aim could be to roll this event functionality into the HDMS and then output as a <chronlist> element in an EAD finding aid.

Þ *Recommendation: Develop an event module for the HDMS*

3. Related Series

The HDMS allows for relationships amongst series to be recorded in a RelatedSeries table. Information captured includes the nature of the relationship between the series, a description of the relationship and the dates of the relationship. The relationship may be one of the 'standard' relationships of Australian

practice, e.g previous, subsequent, controlling, controlled, related, or any other type, i.e. you can define your own relationships.

From this information network and hierarchical representations of series may be built. EAD allows for hierarchy, through recursive <c> elements in the <dsc>, but can it support a network model to the same extent?

The <ref> element is defined as 'An internal linking element that provides for movement from one place in a finding aid to another place in the same finding aid.' It may be used to ' provide a dynamic link from one Component <c> to another related Component <c> in the same way that See and See also references direct readers of paper-based finding aids'[3] It can also support a range of tags within it, to give some structure to the data it encapsulates. There is also a <relatedmaterial> tag within <add>, which supports <ref> and is comparable to ISAD(G) 3.5.3 Related Units of Description[4] .

The key question is whether the related series structuring of the CRS, i.e. relationship, related series, description of relationship, dates of relationship, can be supported by the prescriptive use of existing tags, like <ref> and <relatedmaterial> or whether development of EAD is required?

- *Recommendation: Discuss with EAD Working group*

4. Related Provenance

As with series, the HDMS allows for relationships amongst provenance entities to be recorded in a RelatedProvenance table. The same type of information, i.e. relationship, description, dates, as for related series is captured. As with series the question is whether this information can be expressed reasonably within EAD at the moment, given that the <bioghist> tag only supports the <ref> element through a <p> element.

- *Recommendation: To progress as part of the ISAAR(CPF) DTD initiative*

5. HDMS EAD Refinements

The mapping of HDMS fields to EAD elements has been based on Austehc's extensive use of the HDMS across a range of different projects. However, we do acknowledge that for various reasons it is impractical for other HDMS users to slavishly follow our data entry protocols. The question then is how robust our mapping is given the potential variety of application? It will be answered as we roll out the HDMS upgrade across our data, our existing users and onto new users.

- *Recommendation: Upgrade existing users of the HDMS to the EAD version, i.e. version 7.5, as quickly as possible and establish a listserv for users to discuss this and other HDMS issues.*
- *Modify the EAD generator and application guidelines as per feedback maintaining the balance between general applicability versus customisation.*

6. EAD Data Interchange

Mention has been made in the literature of the potential for EAD to be used as a standard for the interchange of archival descriptive data. This project has focused on building a particular EAD export facility for the HDMS, i.e. to create a 'one document' finding aid for delivery over the Web. Further work should look at how 'all' of the data in the HDMS can be exported as EAD for translation into another archival description management system, and then how EAD could be imported into the data structures of the HDMS.

[3] 'EAD Elements: <ref> Reference' *EAD Tag Library Version 1.0* at http://lcweb.loc.gov/ead/tglib/tlin117.html

[4] ISAD(G) 2nd edition 3.3.3 rule is given as 'record information about units of description in the same repository or elsewhere that are related by provenance or other association(s)' which contrasts to the <relatedmaterial> tag guideline of ' An Adjunct Descriptive Data <add> subelement for information about materials that are not physically or logically included in the material described in the finding aid but that may be of use to a reader because of an association to the described materials. Materials designated by this element are not related to the described material by provenance, accumulation, or use.'

This project gives the preliminary mappings between EAD and the HDMS structure on which this further work can be based. But at this stage there needs to be testing and evaluation of the existing mappings before moving on to the next stage.

7. Evaluation Studies

The archival community is in need of user and other evaluation studies of Web based archival resources. Some studies have been undertaken[5], however it is also necessary to look at what the evaluation criteria should be for Web based archival resources - i.e. as well as general Web usability how important is it for Web based archival resources to be contextualised, citable, and linked into existing archival infrastructure?

- *Recommendation: Analyse the criteria used in existing evaluation studies and encourage further studies to both evaluate and test evaluation frameworks.*

Conclusion

EAD is an initiative to be highly commended and its rollout in Australia to be encouraged. Australian perspectives must be fed back to the EAD working group to facilitate its evolution as an 'emerging encoding and structural standard for archival description'[6].

Any EAD implementation involves significant investment in time, resources and skills development. At whatever level it is taken on at and by whatever method, it involves the establishment of new systems and acquisition of new skills. It also requires long term commitment to the development of those systems and skills as EAD and SGML/XML technology evolves.

So an EAD implementation at the moment is not for the faint hearted. Most of the software tools are still relatively immature and, indeed in the case of XML structural searching tools, little more than gleams in an application developers eye. Also lacking are user studies that evaluate existing electronic finding aids and associated Web systems. These studies are crucial in order to assess what has been done to date and too prioritise future functional development.

What we have done with the development of an EAD generator for the HDMS is to make the production of an EAD finding aid a byproduct from the processing of records. As my colleague Tim Sherratt stated in an internal ASAP email in 1997

> 'The important thing is not just to agree on standards but to maintain the data in systems that are open to establishing new connections. This is one of the great benefits of the ADS [HDMS], which in some ways is the epitome of the "open and scalable" strategy I talk about in Pathways[7].'

With the HDMS being made available under license at no charge to non-profit organisations or for public good purposes, we can facilitate EAD rollout. We can build a databank of EAD finding aids, hopefully from a broad range of perspectives and experience, as part of records processing. It can then be used to refine the HDMS EAD generator itself, provide feedback to the EAD Working group on the Australian experience and help to determine functional priorities for the development of the technological tools.

[5] EAD list serv, 'EAD User Studies' emails of 27th June 2000, see the EAD listserv archive at http://www.loc.gov/cgi-bin/lwgate/EAD/archives/ead.log0006/date/

[6] Pitti, Daniel, 'Introduction to Encoded Archival Description', Powerpoint presentation, Canberra, March 2000 available at http://www.nla.gov.au/initiatives/ead/.

[7] Sherratt, Tim, 'Pathways to Memory - Accessing Archives on the WWW, paper presented at AusWeb96. http://www.scu.edu.au/sponsored/ausweb/ausweb96/cultural/sherratt/, unfortunately this URL goes to a missing page notice. I am in contact with the publishers to see whether we can mount a copy on the ASAP web server.