Communications

Using a Native XML Database for Encoded Archival Description Search and Retrieval

Alan Cornish

The Northwest Digital Archives (NWDA) is a National Endowment for the Humanities–funded effort by fifteen institutions in the Pacific Northwest to create a finding-aids repository. Approximately 2,300 finding aids that follow the Encoded Archival Description (EAD) standard are being contributed to a union catalog by academic and archival institutions in Idaho, Montana, Oregon, and Washington. This paper provides some information on the EAD standard and on search and retrieval issues for EAD XML documents. It describes native XML technology and the issues that were considered in the selection of a native XML database, Ixiasoft's TextML, to support the NWDA project.

Pitti, one of the founders of the EAD standard, noted the primary motivation behind the creation of EAD: "To provide a tool to help mitigate the fact that the geographic distribution of collections severely limits the ability of researchers, educators, and others to locate and use primary sources."[1] Pitti expanded on this need for EAD in a 1999 *D-Lib* article:

> The logical components of archival description and their relations to one another need to be accurately identified in a machine-readable form to support sophisticated indexing, navigation, and display that provide thorough and accurate access to, and description and control of, archival materials.[2]

In a more recent publication, Pitti and Duff noted a key advantage offered by EAD that relates to the focus of this article, the development of an EAD union catalog:

> EAD makes it possible to provide union access to detailed archival descriptions and resources in repositories distributed throughout the world. . . . Libraries and archives will be able to easily share information about complementary records and collections, and to "virtually" integrate collections related by provenance, but dispersed geographically or administratively.[3]

In a 2001 *American Archivist* article, Roth examined EAD history and deployment methods used up to the 2001 time period. Importantly, two of the most prominent delivery systems described by Roth—DynaText (a server-side solution) and Panorama (a client-side solution)—were, by 2003, obsolete products for EAD delivery. This is indicative of the rapid pace of change in EAD deployment, in part due to the migration from SGML to XML technologies. Roth described survey results obtained on EAD deployment that underscore the recognized need at that time for a "cost-effective server-side XML delivery system." The lack of such a solution motivated institutions to choose HTML as a delivery method for EAD finding aids.[4]

Articles like Roth's that describe specific EAD search-and-retrieval implementation options are in short supply. One such option, the University of Michigan DLXS XPAT software, is employed for the search and retrieval of EAD and other metadata in the University of Illinois at Urbana-Champaign (UIUC) Cultural Heritage Repository.[5] Another option, harvesting EAD records into machine-readable cataloging (MARC) to establish search and retrieval access in an integrated library system, was described by Fleck and Seadle in a 2002 Coalition for Networked Information Task Force briefing. Using an XML Harvester product created by Innovative Interfaces, MARC records are generated based upon MARC encoding analogs included in the EAD markup and loaded into an Innovative Interfaces INNOPAC system.[6] This product has been used to create access to EAD finding aids in the catalog for Michigan State University's Vincent Voice Library.

In a 2001 article, Gilliland-Swetland recommended several desirable features for an EAD search-and-retrieval system. She emphasized the challenge of EAD search and retrieval by noting the nature of finding aids themselves:

> Archivists have historically been materials-centric rather than user-centric in their descriptive practices, resulting in the finding aid assuming a form quite unlike the concise bibliographic description with name and subject access most users are accustomed to using in other information systems such as library catalogs, abstracts, and indexes.[7]

Without describing specific software tools, Gilliland-Swetland argued for a user-centric approach to the search and retrieval of finding aids by examining the needs of specific user communities such as genealogists, K–12 teachers, and historians.[8]

Several initiatives similar to the NWDA effort are described in the professional literature. The Online Archive of California (OAC), which was founded in the mid-1990s, is a consortium of California special-collections repositories. A number of key consortium functions are centralized, including "monitoring to ensure consistency of EAD encoding across all OAC finding aids" according to agreed-upon best practices, a critical need in the creation of a union catalog.[9] Brown and Schottlaender also describe the integration of the OAC into the California Digital Library, which enables linkages between EAD

Alan Cornish (cornish@wsu.edu) is Systems Librarian, Washington State University Libraries, Pullman.

finding aids and digitized copies of original materials.[10]

Finally, one important development area is the possibility of integrating EAD documents into Open Archives Initiative (OAI) services in order to enhance resource discovery. A 2002 paper written by Prom and Habing, both of whom work with the UIUC Cultural Heritage Repository, explored the possibility of mapping EAD to OAI, the latter of which is based upon the fifteen-element Dublin Core Metadata Set (unqualified). While noting, "we do not propose that the full capabilities of EAD finding aids could be subsumed by OAI," Prom and Habing suggested that it is possible to map the top-level and component portions of EAD into OAI, resulting in multiple OAI records from a single EAD finding aid. In this scenario, a single OAI record is created from the collection-level information and multiple records from component-level information in an EAD document.[11]

Evaluation of EAD Search and Retrieval Products

In order to identify a software solution for supporting a union catalog of EAD finding aids, the consortium conducted a product evaluation. The strengths and weaknesses of the native XML technology employed by the consortium can be best understood by looking at alternative XML products and product categories. Table 1 shows the products considered during an evaluation period that consisted of both product research and actual trials. In approaching the evaluation, the consortium and its union-catalog host institution, the Washington State University Libraries, had several specific needs in mind. First, the licensing and support costs for the product needed to fit within the consortium's budget. Second, the search-and-retrieval software had to support several basic functions: Keyword searching across all union-catalog finding aids; specific field searching based upon elements or attributes in the EAD document; an ability to customize the look and feel of the interface and search-results screens; and the ability to display search term(s) in the context of the finding aid.

As noted in the table, three of the evaluated products are native XML databases. Cyrenne provides a definition of native XML as a database with these features:

- The XML document is stored intact: "the XML document is preserved as a separate, unique entity in its entirety."
- "Schema independence," that is, "any well-formed XML document can be stored and queried."
- The query language is XML-based: "native XML database vendors typically use a query language designed specifically for XML" as opposed to SQL.[12]

Of the three native XML products, only the licensing costs of Ixiasoft's TextML and the open-source XIndice software fell within the available project funding. Both packages were extensively tested, with TextML proving superior at handling the large (sometimes in the MB-size range) and structurally complex EAD documents created by consortium members.

One key strength of TextML that met an NWDA consortium-need involved field searching. In TextML, it is possible to map a search field to one or more XPath statements, enabling the creation of search fields based upon the precise use of an element or attribute in EAD documents. The importance of this capability is shown with the EAD <unittitle> element, which can appear at the collection level and at the subordinate component level in a document. With TextML, using its limited XPath support, it is possible to reference a specific, contextual use of <unittitle>.

In addition to the native XML solutions, several other product types were considered. An XML query engine, Verity Ultraseek, was tested and produced good results when used for the search and retrieval of consortium documents.[13] Ultraseek can be used to search discrete XML files, supports the creation of custom interfaces for the search-and-retrieval system, and has strong documentation. Probably the most obvious limitation in this XML query-engine product concerned the creation of search fields. To contrast Ultraseek with a native XML solution: Ultraseek 5.0 (used during the product trial) lacked XPath support. Instead, it required a unique element-attribute combination for the creation of a database search field. Returning to the <unittitle> example, contextual uses of <unittitle> could not be indexed without recoding consortium documents to create a unique element-attribute combination on which to index.

An XML-enabled database, DLXS XPAT, has been successfully used in several EAD projects, including OAC. One disadvantage of this product is that it requires a UNIX operating system for the server. Additionally, XPAT, as a supporting toolset for digital-library collection building, provides functionality that duplicates other media tools at the host institution (specifically, OCLC/DiMeMa CONTENTdm).

The use of a Relational Database Management System (RDBMS) to establish search and retrieval for EAD XML documents was considered as well. The advantage to this approach is that it would enable the use of coding techniques built up through other Web-based media delivery projects at the host institution. The most obvious negative issue is the need to map XML elements or attributes to tables and fields in an RDBMS, which, as Cyrenne notes, "is often expensive and will most likely result in the loss of some data such as processing instructions, and comments as well as the notion of element and attribute ordering."[14] The

Table 1. NWDA project—evaluated search and retrieval products

Product	Vendor	Product category	License
MySQL/PHP	N/A	Relational database management system	Open source
Tamino XML Server	Software AG	Native XML database	Commercial
Textml	Ixiasoft	Native XML database	Commercial
Ultraseek	Verity	XML query engine	Commercial
Xindice	N/A	Native XML database	Open source
XML Harvester	Innovative Interfaces	Integrated library system	Commercial
XPAT	DLXS	XML enabled database	Commercial

use of native XML avoids the task of exploding XML data into the table and field structures of an RDBMS.

Finally, another approach considered was the use of an integrated library system product. This was a realistic option for NWDA because consortium member institutions had decided to include MARC encoding catalogs for selected elements in union-catalog finding aids. Innovative Interfaces produces an XML Harvester that can be used to generate MARC records from EAD finding aids that include MARC encoding analogs. For this project, a local (or self-contained) catalog could have been created and populated with MARC records containing metadata for the EAD documents, including a URL for online access. This approach offers important strengths and weaknesses. On the positive side, it is a relatively easy method for enabling search-and-retrieval access to EAD finding aids. In contrast to the interface coding requirements for TextML, the XML Harvester provided an almost turnkey approach to XML search and retrieval. On the negative side, two factors stood out during the evaluation. First, it would be difficult to fully customize search-and-retrieval interfaces as needed for the project. Second, using the XML Harvester, there is no ability to display search terms in the context of the finding aid. Search and retrieval is based upon the metadata extracted from the finding aid using the MARC analogs. In Michigan State's Voice Library implementation of this solution, the finding aid is an external resource with no highlighting of search terms.

Strengths and Weaknesses of the TextML Approach

Each project has its own specific needs; thus, there is no correct approach to establishing search and retrieval for EAD XML documents. In taking the needs and resources of the NWDA consortium into account, Ixiasoft's TextML, a native XML product, provided the best fit and was licensed for use. The use of TextML enables the creation of customized interfaces for an XML database (or Document Base, using the TextML terminology) and provides support for keyword and field searching of consortium documents. The qualified XPath support in TextML enables search fields to be built upon precise element or attribute combinations within EAD documents.

The existence of a major finding-aids Internet site employing TextML was a factor in the project's selection of the software. The Access to Archives (A2A) site, accessible from URL www.a2a.pro.gov.uk/, provides an excellent model for a publicly searchable finding-aid site. The A2A site supports keyword searching and searching by archival facility; provides multiple views of search results (a summary records screen, search terms in context, and the full record); highlights search term(s) in the displayed finding aid; and supports the presentation of large finding-aid documents. While A2A uses General International Standard Archival Description, or ISAD(G), as opposed to EAD for its description standard, the similarities between the two standards makes the A2A site a valuable example for development.[15]

One weakness of TextML is the implementation model supported by Ixiasoft, which assumes significant local development of the application or Web interface. The relationship between software capabilities and local development was considered with each of the products listed in table 1. As noted, the Innovative Interfaces solution was the most straightforward approach, assuming the existence of the MARC analogs in EAD markup, but provided the least flexibility in terms of customization and establishing a true linkage between the search system and the actual document. In contrast, while Ixiasoft makes available a base set of active server pages using visual basic script (ASP/VBScript) code for TextML application development and provides very good training and support services, the responsibility for

that development rests with the local site. For the NWDA consortium, this development, using the code base, has been manageable. The current state of interface development for the NWDA project can be reviewed at http://nwda.wsulibs.wsu.edu/project_info/.

Conclusion

In selecting an EAD search-and-retrieval system, one important question for the consortium was, Which software solution had the best prospects for migration in the future? Because of the inherent strengths of native XML technology in comparison to the other product categories listed in table 1, a native XML database appeared to be the best approach, and TextML provided the best combination of licensing costs, software capabilities, and support.

It is important to note that the distinctions between native XML databases and databases that support XML through extensions (XML-enabled databases) may become more difficult to discern over time, in part due to the existing expertise and investments in RDBMS technologies.[16] Nevertheless, capabilities central to native XML, such as the use of an XML-based query language, are integral to the success of such hybrid systems.

References and Notes

1. Daniel Pitti, "Encoded Archival Description: The Development of an Encoding Standard for Archival Finding Aids," *The American Archivist* 60, no. 3 (Summer 1997): 269.

2. Daniel Pitti, "Encoded Archival Description: An Introduction and Overview," *D-Lib Magazine* 5, no. 11 (Nov. 1999). Accessed Nov. 2, 2004, www.dlib.org/dlib/november99/11pitti.html.

3. Daniel V. Pitti and Wendy M. Duff (eds.), "Introduction," in *Encoded Archival Description on the Internet* (Binghamton, N.Y.: Haworth, 2001), 3.

4. James M. Roth, "Serving Up EAD: An Exploratory Study on the Deployment and Utilization of Encoded Archival Description Finding Aids," *The American Archivist* 64, no. 2 (Fall/Winter 2001): 226.

5. Sarah L. Shreeves et al., "Harvesting Cultural Heritage Metadata Using the OAI Protocol," *Library Hi Tech* 21, no. 2 (2003): 161.

6. Nancy Fleck and Michael Seadle, "EAD Harvesting for the National Gallery of the Spoken Word" (paper presented at the Coalition for Networked Information fall 2002 Task Force meeting, San Antonio, Tex., Dec. 2002). Accessed Nov. 2, 2004, www.cni.org/tfms/2002b.fall/handouts/H-EAD-FleckSeadle.doc.

7. Anne J. Gilliland-Swetland, "Popularizing the Finding Aid: Exploiting EAD to Enhance Online Discovery and Retrieval," in *Encoded Archival Description on the Internet* (Binghamton, N.Y.: Haworth, 2001), 207.

8. Ibid, 210–14.

9. Charlotte B. Brown and Brian E. C. Schottlaender, "The Online Archive of California: A Consortial Approach to Encoded Archival Description," in *Encoded Archival Description on the Internet* (Binghamton, N.Y.: Haworth, 2001), 99.

10. Ibid, 103–5. OAC available at: www.oac.cdlib.org/. Accessed Nov. 2, 2004.

11. Christopher J. Prom and Thomas Habing, "Using the Open Archives Initiative Protocols with EAD," in Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries (Portland, Ore., July 2002). Accessed Nov. 2, 2004, http://dli.grainger.uiuc.edu/publications/jcdl2002/p14prom.pdf.

12. Marc Cyrenne, "Going Native: When Should You Use a Native XML Database?" *AIM E-DOC Magazine* 16, no. 6 (Nov./Dec. 2002), 16. Accessed Nov. 2, 2004, www.edocmagazine.com/article_new.asp?ID=25421.

13. Product category decisions based upon definitions and classifications available from: Ronald Bourret, "XML Database Products." Accessed Nov. 2, 2004, www.rpbourret.com/xml/XMLDatabaseProds.htm.

14. Cyrenne, "Going Native," 18.

15. Bill Stockting, "EAD in A2A," Microsoft PowerPoint presentation. Accessed Nov. 2, 2004, www.agad.archiwa.gov.pl/ead/stocking.ppt.

16. Uwe Hohenstein, "Supporting XML in Oracle9i," in Akmal B. Chaudhri, Awais Rashid, and Roberto Zicari (eds.), *XML Data Management: Native XML and XML-Enabled Database Systems* (Boston: Addison-Wesley, 2003), 123–4.

Using GIS to Measure In-Library Book-Use Behavior

Jingfeng Xia

This article is an attempt to develop Geographic Information Systems (GIS) technology into an analytical tool for examining the relationships between the height of the bookshelves and the behavior of library readers in utilizing books within a library. The tool would contain a database to store book-use information and some GIS maps to represent bookshelves. Upon analyzing the data stored in the database, different frequencies of book use across bookshelf layers are displayed on the maps. The tool would provide a wonderful means of visualization through which analysts can quickly realize the spatial distribution of books used in a library. This article reveals that readers tend to pull books out of the bookshelf layers that are easily reachable by human eyes and hands, and thus opens some issues for librarians to reconsider the management of library collections.

Several years ago, when working as a library assistant reshelving books in a university library, the author noted that the majority of books used inside the library were from the mid-range layers of bookshelves. That is, by proportion, few books pulled out by library readers were from the top or bottom layers. Books on the layers that were easily reachable by readers were frequently utilized. Such a book-use distribution pattern made the job of reshelving books easy, but created some inquiries: how could book locations influence the choices of readers in selecting books? If this was not an isolated observation, it must have exposed an interesting