

Around and About

Five College Archives & Manuscript Collections: Building a Dynamic, Searchable Web Site of EAD Finding Aids

Kelcy Shepherd, Project Director, Five College Finding Aids Access Project

*F*ive College Archives & Manuscript Collections is a recently developed Web site of Encoded Archival Description (EAD) finding aids from the archives and special collections of Amherst, Hampshire, Mount Holyoke, and Smith Colleges, and the University of Massachusetts Amherst. The site, at <<http://asteria.fivecolleges.edu/index.html>>, is a product of the Five College Finding Aids Access Project, funded by the Andrew W. Mellon Foundation. The project's goals are to encode approximately 1,100 finding aids using the EAD standard; make these finding aids available on the Internet in a single, searchable union catalog; create a MARC record for each uncataloged collection that will link to the online finding aid by using the 856 field or update existing catalog records to include the 856 field; and establish a methodology that will be sustainable after the project has ended.

This discussion will focus on the second goal: the development of a Web site that dynamically delivers EAD finding aids and provides the ability to do both full-text and field-specific searching of those finding aids.

Project Background

When the Five College project began in 2001, the EAD standard was well documented, particularly the aspects related to the encoding of finding aids. Numerous articles had been published in the archival literature outlining best practices and procedures for converting finding aids to EAD. The

EAD Cookbook was available, providing tools, guidelines, and instructions for encoding. Most training opportunities also focused particularly on marking up finding aids. Understanding the EAD standard, establishing a template for encoding the Five College's finding aids, and creating tools to efficiently convert our finding aids into EAD was not difficult.

But project participants agreed that the ability to encode finding aids was not in itself sufficient for a meaningful implementation of EAD. The project's goals specifically stated that each institution's finding aids be delivered on the Internet in a unified searchable form, and we wanted to develop a Web site that made use of XML's ability to generate content dynamically and support powerful, field-specific searching. Before we could even attempt such a goal, we needed to overcome a number of barriers, not the least of which was learning what exactly this accomplishment would require in terms of tools and expertise. Early EAD adopters had primarily implemented Standard Generalized Markup Language (SGML), though it was clear that eXtensible Markup Language (XML) was eclipsing SGML, and would therefore be the sensible solution for new projects such as ours. Information about delivering and developing searchable access to XML documents on the Internet was difficult to find. The EAD Application Guidelines provided an overview of software solutions, but was already out of date.

Many of the tools mentioned were developed for SGML. Others were no longer being supported. Little else had been written about the Web development aspects of EAD, and there were no training opportunities in the archival community at the time.

Within the Five Colleges, little work with SGML or XML had been done. A small project to mark up two books using the Text Encoding Initiative (TEI) standard had been completed in May 2000, but the librarian who had led the project and the programmer who had provided technical support were no longer working at the library. Systems staff were hoping to adopt the Oracle relational database management system as a standard platform for digital library projects, and Oracle was beginning to tout its capabilities for handling XML. Project staff were uncertain, however, of its suitability since Oracle's XML features were at that time limited. One of Oracle's XML solutions was to translate the XML document into a relational database model, a model that does not easily accommodate the often complex hierarchical relationships of data in a finding aid. Oracle's other method for handling XML documents was to store the full text of the documents, using its built-in search engine to index them. Using Oracle simply as an XML search engine seemed an overly complex solution. In addition, the number of other digital library projects being supported by a small systems department indicated that project staff would need to take a lead in exploring XML.

Learning About XML and Web Development

It was clear that the project's staff needed to discover more about implementing this relatively new technology. Research provided us with a high level of understanding. We learned that solutions for searching XML documents include XML search engines, XML-enabled databases, and native XML databases. An XML search engine is capable of creating indexes that retain information about the context of the indexed term, so that users can search within specific elements. XML-enabled databases offer tools for transferring data in and out of an XML format, but either break up the XML document into a relational data structure, or simply store the document as text without recognizing the inherent data structure of the XML. Native XML databases are specially designed to store and manipulate XML data. They maintain the XML document as the fundamental storage unit and utilize XML-related standards to access data. At the time of the project's inception there were few true native XML databases available, but this area has since grown.

There was also a wealth of information available about the various software packages available in each of these categories, but developing a more detailed understanding of how any of these software applications could be implemented was more difficult. Before developing the Five College's finding aids Web site and choosing software, project staff needed to learn what skills and expertise would be required to implement these solutions, and what kind of time would be involved.

A workshop sponsored by the Association of Research Libraries, *Web Development with XML*, provided the first opportunity for the project director to gain actual hands-on experience with these kinds of tools. The week-long course covered the XML basics, XSLT stylesheets used to transform XML to HTML, and components for developing a dynamic, searchable Web site of XML documents. Though ultimately the tools used in the course were not the ones project staff selected for the Five College project, the workshop provided a better understanding of the pieces necessary to deliver and develop searching features for XML documents on the Web. It also demonstrated that the development of an XML-based Web site was within the project's grasp. Site visits and discussions with other EAD implementers helped to expand our understanding. Through these conversations we gathered information about specific products being used, learned more about the amount of time and expertise other implementers devoted to their efforts, and began to explore different models for sustaining EAD.

Developing an XML Solution for the Five Colleges

Implementation of the Five College Archives & Manuscript Collections Web site relied primarily on the project director, a part-time project assistant, and some additional support from the UNIX systems administrator at the University of Massachusetts Amherst library. The project director, an archivist, directs and supervises all aspects of the Web development and holds direct responsibility for HTML content and development of XSLT stylesheets for display of the finding aids. The project assis-

tant, a graduate student in computer science with Perl and Java programming skills, installs, configures, and troubleshoots software used in the site and develops additional stylesheets that support the search interface. The systems administrator assists with installation and troubleshooting and maintains the server that houses the Web site.

The project director and assistant researched a few specific products, including the DLXS system offered by the University of Michigan and the Tamino native XML database. Both are in use by other EAD implementers. The amount of time and expertise needed for each of these systems was more than we were willing to commit, given that the finding aids project was then the only effort that would benefit from the software and skills. We decided to focus on developing an initial system as rapidly as possible, using the system to learn more about XML Web development, and then evaluating the system and its use over time. If the Five College libraries develop additional XML content, it may be possible to migrate to a more robust system that would be shared by a number of digital library projects, but in the meantime a more straightforward and easily executed approach is desirable.

Given these parameters, we chose to implement a system consisting of the Cocoon XML publishing suite and the Lucene search engine, both no-cost, open source products of the Apache Software Foundation. With these tools we were able to acquire, install, and develop the software rapidly. And, because the underlying source code is available to users, the project assistant was able to modify Lucene's Java code to customize index-

ing for our finding aids. In addition, the tools are based on XML and its related standards, so the stylesheets and other tools developed for the Web site can be migrated to a different XML-compliant system in the future if necessary.

Considerations for Other EAD Implementers

The challenges the Five Colleges faced in developing the expertise necessary to build a Web site of dynamic, searchable EAD finding aids are not unique. For smaller institutions, these challenges will likely continue to be insurmountable barriers for some time. There are signs, such as the recent development of advanced training courses, that the archival community is beginning to address some of the issues we faced. The adoption of XML by libraries may also provide opportunities for some archivists to benefit from the technical expertise of systems staff or other colleagues. Other repositories may develop collaborative projects, as we did, in order to gain the necessary support.

The Cocoon/Lucene implementation chosen by the Five Colleges will not be the model for every EAD project. Evaluation of the many options for delivering XML on the Web requires an assessment of the individual situation. Important administrative questions include:

- What technical expertise is available? Do you have a programmer or database administrator to assist you?
- How much money for software is available?
- How much time do you have to develop and customize? Are you looking for an out-of-the-box solution or are you willing to build something from different components?
- Are there similar efforts underway at your institution on which to build?
- What searching and display features are necessary, or desired?
- Are you interested in delivering only finding aids, or will you need a system that handles other forms of XML as well, for example metadata for digitized images?

These administrative questions are very important, but there will also be technical considerations such as the operating platform, performance needs, scalability, etc. If possible, work closely with those providing technical support to evaluate these kinds of concerns. Talk to others implementing EAD to get their impressions of the software they're using and learn more about implementation time and expertise. If training funds are available, attend workshops that cover XSLT stylesheets and other aspects of Web development for XML, for example, the *Publishing EAD Finding Aids* course now offered at the University of Virginia's Rare Books School, or the Society of American Archivists *Stylesheets for EAD* workshop.

Future of EAD at the Five Colleges

With the Finding Aids Access Project in its last year, participants face new challenges as we plan for the sustainability of the Web site. The decision to rely on project staff to complete the bulk of Web development work was an effective one in terms of meeting the project's goals on schedule. Additional planning and training will now be required to ensure that the Five Colleges have the expertise necessary to maintain and continue development of the Web site once the project has ended. Ideally, the skills necessary to support the archives' XML implementation would also be used to develop additional XML projects in other areas of the Five College libraries.

The project may provide new opportunities as well. The Five Colleges are considering implementation of a federated searching system, an interface that would allow users to enter a single search that returned results from the library catalog, commercial databases and full-text resources, and local electronic resources. Because the finding aids are available online with a searchable interface, they are ready to be included in a federated searching system. This would allow us to integrate the finding aids with the full range of print and electronic resources available in the Five College libraries, improving access to the Five College's primary sources beyond what we had envisioned at the project's inception. ☛

Please visit us online at
www.newenglandarchivists.org