

**Nine-Tenths Planning, One-Tenth Tagging: The Physics History
EAD Consortium
by Katherine A. Hayes**

Introduction

My first attempt at using Encoded Archival Description occurred in early 1998, about a year after taking the 2-day RLG (now SAA) course. I was asked to apply it to the finding aid for the Samuel Goudsmit papers, which was chosen as a sample for a grant application being submitted to the NEH. I'd been excited about EAD ever since hearing Daniel Pitti and Elizabeth Dow describe and demonstrate it at SAA meetings several years ago. Before I'd even finished this first attempt, however, frustration hit me. I used WordPerfect 7's SGML module, and Panorama Pro for display. Both were poorly documented, and had plenty of bugs.

Our grant application was approved the following spring. When I learned we were going to undertake the transformation into EAD of 4,118 pages of 76 legacy finding aids, I tried not to think of the stress ahead. I didn't realize that planning, discussing and managing this project would take a large portion of our time and energy.

First I'll give you some background on The structure and goals of the project. I'll talk about some of the issues and problems we confronted. I'll try not to digress into too much technical jargon. Although the project isn't quite finished, I'll include some preliminary results, and then thoughts we've had about where to go from here.

Background

In the summer of 1998 we applied to NEH's Division of Preservation and Access for funding to develop "a union database of history of science finding aids." The funding covered about 42% of the total project budget. We planned to cooperate with eight major academic

archives to create a core database of legacy finding aids describing collections in the physical sciences. The grant emphasized that we'd test the feasibility of a small repository tackling such a large project. By the end of the year we'd know if we could maintain and continue this work after the project period ended. We wanted our database of finding aids to include significant, heavily used collections, most of which came from other repositories. We asked for their cooperation and help in preparing the finding aids and in proofreading them.

We started by taking a close look at actual finding aids. We wanted completed, well-formed descriptions that weren't already encoded elsewhere. They also had to represent important collections in the realm of the physical sciences. The holding repositories agreed to participate by supplying us with up-to-date copy at the beginning of the project period, in the form of electronic or paper documents. They were also asked to proofread their finding aids at a non-public URL, and then approve the electronic publication. Consortium members would be credited in each finding aid and on the web site.

Organizing the project

Obviously any project requires planning, especially if you are using a new technology. Some of the questions we asked ourselves in the beginning:

- What software tools to use;
- What kind of technical expertise is needed;
- How should we treat different styles of finding aids; how much (if any) editing to do;
- Which tags to use in our template, and how to "standardize" as much as possible;
- How to automate tagging as much as possible, especially the container list;

- How to input paper finding aids;
- How to present finding aids on the web;
- How will they be accessed on-line? How should the interface be composed;
- How will finding aids be indexed for searching;
- How will we proof or check finding aids; how will the holder do it?

But first we had to make staffing decisions. We initially asked for funding to pay for a full-time archivist to cover our work so we could concentrate on the project. However, when we had the luck to hire a recently graduated archivist (Clay Redding) who was technically skilled with EAD experience, the plan changed. We also weren't sure how much internal tech support we could depend on, because AIP's web support staff is scattered over several departments with their own priorities. As a result, we hired Clay and allocated the temporary position entirely to the grant. Existing staff would continue doing regular archival work, as well as working on the project. As it turned out, the technical side of the project required more than 1 FTE's hours.

As we talked about how to approach the project, we realized that all the different steps would get confusing. We knew that eventually we'd want to analyze the process as well as the results. So, to track the progress of each finding aid we decided to create a Microsoft Access database. Finding aid styles ranged from simple inventories to large printed documents describing multiple related collections. Some required scanning, some only needed partial text input. We tried to anticipate all possibilities by creating 54 fields for data, with numerous queries, forms and reports. As we hoped, this database proved useful in compiling our progress statistics, reporting to NEH and consortium members, and writing papers for this session.

[template overhead]

Now I'll briefly get a bit technical and mention some of the EAD tags in more detail. We created an EAD document template so that display, indexing and searching would be more uniform. After studying many sources, we arrived at our template. Because container lists varied the most, we concentrated on uniformity at the front of the finding aid, before the <dsc> or description of subordinate components. We chose not to use tabular display in the container list, although that would have made the display easier. We eliminated the <frontmatter> tags, since much of its content was duplicated in <eadheader>. The collection-level <did> was the most tightly controlled part of our template.

Another important step in the process included comparing each collection's MARC record with the finding aid. We planned to link them, so it was important that they "matched". We also decided to add the 600-level fields from the MARC record to the finding aid within the <controlaccess> tag.

We made extensive use of the published Tag Library and Application Guidelines, and web sites with examples of templates and conversion guidelines. The EAD Help Pages provided information on other repositories' methods of applying EAD and contact information. As a result, I was not shy in emailing and telephoning these and other individuals with demonstrated expertise and willingness to help.

Five types of problems We encountered difficulties that fell into five categories. They were: finding aid format; software; text conversion; participant cooperation; technical assistance.

The biggest problem we had was with the *widely varied formats of legacy finding aids*. Initially we decided to avoid rearranging information to fit our template. With some finding aids this decision worked; others required some tinkering. Since we decided that the collection-level

descriptive identification <did> must be complete, sometimes that information was rearranged, duplicated or added. The printed and bound finding aids were hardest to adapt to EAD. One very large publication actually described four collections which we decided to separate. In each one we duplicated contextual information, adding a note describing the original document.

Idiosyncratic container lists required the most time to tag, and were most difficult to display.

Inconsistent placement of dates or confusing hierarchies required us to examine each document and decide on how to treat it. Would it be easier to rearrange the data elements or leave them all in a <unittitle> tag? The scripts we used for the container lists also had their pitfalls, requiring testing on each new type of container list.

Conversion of print documents presented some difficulties. For a few very large printed documents we decided to see how an outside vendor performed the rekeying and tagging. The finding aids that were only missing parts of their data were rekeyed in-house. We also scanned some ourselves. Internal scanning resulted in the most work checking the text. Characters such as the number "1", capital "I", and lower-case "l" weren't distinguished. The in-house rekeyed text was the cleanest and took the least time in tagging, but there weren't many examples. The outsourced rekeyed and tagged text required considerable refinement, due to the non-standard documents.

Software presented the next hurdle. Which to use, how much it cost, would it work -- we discussed, tested, researched. The most comprehensive, user-friendly tools were also the most expensive. Our original proposal said we'd use WordPerfect and Panorama. However, my early experiences with both, the disappearance of the Panorama Free viewer, and Clay's expertise made us change courses. We ended up using freeware and shareware for most of the work. As

we learned how to use newer web technologies, our display methods evolved. First we planned to use a PERL script to convert the SGML to HTML. However, it couldn't process long container lists. In the end we used XSL transformations, which converts the EAD/XML document into HTML using Cascading Style Sheets. We're sticking with displaying HTML because not all web browsers recognize XML; Internet Explorer uses a proprietary, non-standard formulations of it. We made it a priority to be sure our documents conformed to the open standards agreed upon by the WorldWideWeb Consortium for XML.

Any project involving *multiple institutions* produces different expectations on the part of the participants. Ours was no exception. We asked project participants to submit their finding aids early in the project, and once finished, to check them for errors. Exceptions were expected and accommodated. In reality, a number of finding aids weren't received until well into the project year. One institution took a second look and decided to rework some of its finding aids. One collection wasn't completely processed until the third quarter. We selected new documents and institutions in order to meet our projected goals. Although responses to the online proofs are slow, we'll go public with finding aids 30 days after they're finished.

Lastly, we also had to adjust our expectations of *internal institutional assistance*. The Niels Bohr Library is a small unit within the American Institution of Physics. AIP has its own site indexing software, web servers, and staff to run it. Initially we hoped to have AIP's web programmers help with our PERL scripts. Their time limitations, and the complexity of EAD's SGML DTD forced us to reevaluate this approach. We ended up doing the conversions ourselves. However, the web staff did perform the indexing based on our instructions, which took very little time.

We planned to *proofread* the EAD files in house. But when a staff member took a new job out of state we brought in outside help with money unspent on scanning and rekeying of text. However, we still had to create tools and guidelines for someone who was unfamiliar with finding aids or archives. The proofreaders had difficulty telling the difference between a text error and a style sheet error. The person doing the corrections frequently had to look up the original text, and catch errors the proofreader missed. All this took longer than anticipated.

Conclusions

The end of the project year approaches, and our project has been successful. We're still receiving finding aids that were part of the project. However, we've converted more document pages than we planned. We've learned what it takes to sustain EAD conversions. The pitfalls and problems are obvious, the benefits more elusive. Some initial findings are:

- Trying to maintain a timetable for a collaborative project is difficult. Asking multiple repositories to coordinate their time and resources is nearly impossible.
- EAD exposes the weaknesses in poorly formed finding aids; the opportunity to re-evaluate them may result in rewriting, or even reprocessing. COROLLARY: Poorly formed or weak finding aids are difficult and costly to tag.
- One must look critically at standardizing descriptions before starting to apply EAD.
- EAD requires technical expertise in web applications and scripting beyond the basic tagging courses. Depending on the desired results, it may require some level of programming skill.
- In lieu of in-house technical personnel, money for software packages or contractual programming will help ensure successful conversions and searching.

- Retrospective tagging is complex, difficult, unpredictable, and time-consuming. However, EAD is well-suited to the creation of new finding aids.
- It's important to use available resources, like the Application Guidelines and Tag Library, the EAD Help Pages on the web, and those who will share their experiences, advice, and tools.

Future

What's next? We intend to finish what we started, although it may take a month or two. We'll continue to add EAD finding aids to our web site, but may be more selective in the future.

How do researchers view the usefulness of EAD finding aids? We'll attach a user survey to our web site. Others at this conference have already completed user studies. How will their findings affect us?

We are looking at ways to share others' encoded finding aids, either by adding them to our web site, linking from the MARC record, or using distributed servers. We're now thinking of using EAD as an XML 'wrapper' or administrative tool to more efficiently present legacy information. We can save time and money by enclosing existing descriptions, especially long or complex container lists, in XML and reformatting only the collection-level information. EAD might gain wider acceptance if it's viewed as an intermediary device instead of a destination; a door instead of a wall. EAD demonstrates the need for standardizing finding aids. Its compliance with XML and emerging open standards for sharing information gives it the flexibility to persist as a tool for archival description.