Archivists Workbench: White Paper

Robin Chandler, Online Archive of California

Bill Landis, University of California, Irvine

Bradley Westbrook, University of California, San Diego

1 November 2001

Background

Ten to fifteen years ago the process of archival description was fairly simple. Typically, archivists created inventories or finding aids for archival collections using a word processor or, in some cases, a typewriter. Administrative information-such as deeds of gift, accession records and action logs-was kept as printed forms in collection control files. Some repositories with sufficient staff expertise and access to an online bibliographic utility created collection-level MARC records for their archival holdings.

Beginning around 1990 the complexity of the descriptive process increased dramatically as archivists began to experiment with the Internet as a tool for publicizing their collections to the research community. Archivists first utilized Gopher and WAIS technology to deliver ASCII versions of collection finding aids but quickly migrated to HTML-encoded finding aids once that encoding scheme was broadly introduced in 1993. HTML served to improve the presentation of finding aids; however, its limitations for facilitating searching and navigating online finding aids was quickly apparent to archivists. Furthermore, it did not help to promote consistent identification of encoded data elements within and across repositories. Dissatisfaction with these drawbacks led to the development of an SGML DTD specifically for encoding archival collection descriptions and facilitating their publication online. This DTD, known as Encoded Archival Description (EAD), allows archivists to represent the hierarchical structure inherent in archival collections in encoding and utilize it for searching and navigating through a finding aid or groups of finding aids. EAD also makes possible the kind of data encoding standardization that more predictable, less idiosyncratic access systems require. The success of EAD quickly led to the construction of union databases of EAD-encoded finding aids, of which the Online Archive of California (OAC) was the first. Similar statewide efforts are underway in New Mexico, Texas, Virginia, and North Carolina, along with several international projects.

An analysis of OAC efforts thus far reveals two areas clearly in need of additional work if the OAC is to mature satisfactorily as a user-responsive database of finding aids and associated digital objects representing archival holdings in California repositories. First, while many significant archival repositories in California are currently participating, many more are not. Moreover, some repositories are not able to participate very actively. Among the factors that help to explain this are the difficulty of integrating encoding with description and the cost and

complexity of maintaining separate description and encoding processes. The majority of archivists create a finding aid using a word processing application, followed by a secondary encoding process utilizing one of several available methods: manual encoding, use of scripts and macros, or commercial encoding tools such as Author/Editor or XMetaL. A recent posting to the Archives listserv by the Director of the Five College On-Line Finding Aids Access Project is indicative of the challenges:

We are starting a three-year EAD project that involves five institutions, and will encompass both conversion of legacy finding aids and the creation of new ones. Four of the five institutions are currently using Word or WordPerfect to create finding aids. Of these, two also have some collection-level description in a database format (InMagic and Minaret) but neither of these use the database for complete finding aids. One institution generates complete finding aids for all of its collections from a database (Minaret). We would like to utilize the same encoding method for new finding aids at all institutions. (Archives listserv, 25 Jan. 2001)

Second, the encoding of data is highly inconsistent, thereby impeding the functionality of the union database. For instance, searches of scope and content notes suffer a certain amount of imprecision since many encoded finding aids in OAC lack a scope and content note, while other finding aids have the scope and content note coded with a tag other than the scope and content tag. Achieving optimal performance in a union database requires a high degree of encoding and, to a lesser extent, content consistency. Such consistency enables the construction of navigational interfaces and search indexes to support more sophisticated and precise use of the data by both archivists and researchers. Lacking encoding consistency, a union database of SGML-encoded finding aids has not much more functionality than one created utilizing ASCII text or HTML encoding. In short, integrating multi-institutional descriptive and encoding processes and normalizing archival data are essential for developing the OAC efficiently and effectively, but each requires a time-consuming and expensive effort that most individual repositories simply cannot undertake.

In the fall of 2000, the OAC Metadata Standards Committee formulated best practice guidelines to reduce encoding inconsistencies in newly encoded OAC finding aids. To be effective, however, these guidelines must be incorporated into a work process that integrates description and encoding. In recent years, several efforts have been made to reduce the learning curve for incorporating EAD encoding into an individual repository's workflow. The Society of American Archivists' EAD Application Guidelines and Michael Fox's EAD Cookbook are two notable examples. Neither of these tools, though, helps effect the integration of description and encoding into a single process.

Proposal

As a solution to this continuing problem, we believe a national level project to build tools for addressing these issues is needed. We envision an initial planning meeting to strategize about these issues and the tools their solutions require. Individuals attending the meeting would include domain experts in archival description, information technology and administration. We anticipate this initial meeting will lead to a series of planning meetings based on identified high level tasks. The foremost purpose of these meetings would be to begin development of a suite of Open Source tools to increase the efficiency of managing archival collections and producing

EAD-encoded finding aids by integrating description and encoding and creating metadata for digital objects associated with finding aids. The meetings would also identify funding sources to support construction and testing of a prototype. When implemented, an archivists' workbench would result simultaneously in more consistently encoded data in the OAC, in more sophisticated searching and navigation of finding aids and attached digital facsimiles, and more streamlined processes for administering archival collections. A suite of digital tools to support archives would be indispensable in building multifunctional virtual collections that would satisfy the interest and needs of disparate audiences.

This suite of tools--an archivists' workbench--would satisfy requirements for several archival tasks or processes:

- It would support input and editing of information elements derived from archivists' management of archival collections, including appraisal, accessioning, and processing.
- It would support input and editing of all necessary forms of metadata regarding original or surrogate digital objects associated with a collection.
- It would facilitate manipulation and use of that data by archivists in management of collections, both online and in printed reports.
- It would support searching, extracting, displaying, and publishing the data for a variety of research needs in both online encoded and print formats.
- It would promote quality assurance of the data. • It would enable exporting of data in multiple encoding standards, and it would be adaptable to emerging encoding standards.

The suite of tools would likely consist of the following components:

- Databases of archival administrative and descriptive data.
- Web-based data inputting and editing templates / forms.
- Specialized scripts for querying data in various ways.
- Specialized output style sheets for a number of encoded (e.g., EAD, MOAII, TEI, HTML) and print (e.g., printed finding aids and other printed research and access tools) formats.

Constructing and implementing an archivists' workbench would alleviate the two fundamental problems mentioned above: it would increase standardization of descriptive data elements and would allow data encoding to be done behind the scenes, as it were, according to pre-established encoding protocols. The use of input / editing templates would increase data consistency to a certain degree, while still allowing repositories a reasonable amount of latitude in degree of detail used in a given description. The tools in an archivists' workbench would streamline the descriptive process, as archivists would be able to begin describing a collection at the point of accession, amplifying and completing the description as the collection becomes

fully processed. In the end, after collection descriptions are completed using the tool suite, the workbench could easily facilitate the development of more sophisticated access mechanisms that would benefit specialized researchers.

Standardization and consistency of encoding and description will facilitate more sophisticated uses of encoded data within the larger world of the California Digital Library and nationally. Within the OAC testbed, encoding done with this suite of tools would enable the creation of topical views, based on controlled access terms, of OAC resources for which curators and other specialists might provide a contextual overview, as well as permit end users to extract, merge, and otherwise manipulate information resources from the OAC in a way that is more meaningful to their individual needs. A very specific objective of this endeavor will be to enable "out of context" searching of finding aid data and attached digital objects, which will greatly supplement the "in context" searching now supported. "In context" searching returns a set of finding aids that contain matches to the search query. Each finding aid then must be searched individually for the match(es). "Out of context" searching will return only the part of the finding aid or the digital objects that match the query. However, these results should be presented in such a way that the researcher can easily identify the collection and its repository to which the description pertains and, also, that the researcher can easily jump to the part of the finding aid from which the description or object is taken. Enabling "out of context searching" is fundamental to establishing true, multipurpose virtual collections. The archivists' workbench would make managing collection data and digital objects much easier than it is currently.

PROJECT WORK SEQUENCE

Phase 1: Basic design.

A project team will convene a series of retreats with a dozen or so identified experts in archival description and / or information technology, the goal of which will be to determine and elaborate the data and metadata requirements for an archivists' workbench, to develop clear output pathways for extracting descriptive data in a number of predefined structures, and to discuss the advantages and drawbacks of the various technological options available to realize this suite of tools. Ultimately, the project team and the experts invited to these retreats will produce the specifications for a prototype archivists' workbench. The specifications, in turn, will be offered to the archival community for comment and additional refinement.

Phase 2: Prototype development.

The project team will develop a prototype archivists' workbench based on the specifications defined in Phase 1. The prototype, at various phases in its development, will be tested by a representative group of participants, including the OAC as a testbed, and their feedback will be incorporated into ongoing refinement of the prototype. At the same time, members of the project team will begin work on formulating documentation strategy needed for training project participants in the use of the archivists' workbench.

Phase 3: Funding procurement

Members of the project team will develop funding request(s) to support construction and

implementation of the tool suite.

Phase 4: Archivists' workbench implementation.

The project team will develop a documentary infrastructure to support implementation of the archivists' workbench among all project participants, including the OAC testbed, and training of staff from these repositories in use of the workbench. This work will also include identification of costs associated with training and continued documentation. Phase 4 will also include the definition of a procedure within the CDL for future development of the archivists' workbench to insure that it remains synchronized with pertinent technological developments. As happened with EAD, the workbench will be available to other repositories and consortia outside of CDL-OAC. It is anticipated the archivists' workbench will be tested and implemented first at the partnering institutions. Potentially, these include UCI and UCSD, and prospectively Cornell University, Library of Congress, Minnesota Historical Society, University of Pittsburgh, and Yale University.

PHASE ONE DETAIL

It is expected the initial meeting will cost between $5,000.00 to $7,000.00 dollars, depending on how many participants invited are from the mid-west or east. Funding will be used to cover travel and per diem expenses. Phase One High Level Tasks for Discussion (not prioritized)

1: Administration

Identifying organizational structures and resources required to develop and implement the project.

2: Data Modeling (Description):

Identify the descriptive elements required to be managed by the tool set, as well as the preferred type of tools for managing them.

3: Data Modeling (Administration):

Identify the administrative elements required to be managed by the tool set, as well as the preferred type of tools for managing them. Administrative data elements are key to managing collections, but they are often ancillary to the descriptive data that constitutes the larger portion of a finding aid. Moreover, they don't necessarily belong in a finding aid, as their value is largely administrative and not research. Attention must be given to integrate descriptive and administrative elements in the same tools, but it is recognized that the differences between descriptive and administrative data, as well as the idiosyncratic ways archival collections are managed by different repositories, may require the construction of separate but related tool sets.

4: System Integration:

Discuss and define the system parameters based on data modeling considerations.

5: Prototype Development:

Develop a prototype archivists' workbench, a paper based schematic of the tool set, showing its functions and relationships between component parts.

6: Response to / Critique of Prototype:

Selected archivists, administrators, and information technologists critique and refine the prototype and define the process for constructing the prototype.

7: Education / Promotion / Community Participation

Define method for promulgating the workbench and assisting its implementation in the general archival community.

8: Maintenance of Archivist's Workbench

Discuss issues and develop strategies for evolving workbench to take advantage of technological developments.