PROSPECTING AN ARCHIVISTS' DIGITAL TOOLKIT

INTRODUCTION

Archives and manuscript resources have been an indispensable part of the modern research endeavor. Many groundbreaking works in numerous disciplines would not have been executed without access to the kinds of documentary evidence present in institutional archival collections. Initial development and application of online technologies to archival resources in the 1990s has been a key means for promoting archival resources to a greater number of researchers across research and national domains. The Encoded Archival Description Document Type Definition has provided the means for an increasing number of repositories to publicize their holdings to remote researchers and to participate in the development of union databases of finding aids. These databases have matured quickly to include not only finding aids but also digital surrogates of collection materials that can be accessed with or without the use of a finding aid. Presenting digital surrogates via the internet permits archivists and curators to share their historical content with other kinds of audiences, for example, K-12 classrooms, in a manner more efficient for all concerned and without threatening or degrading the condition of the original resources.

While the initial convergence of archival information and digital technology was and continues to be very beneficial and exciting, it also has revealed or created problems that could be easily ignored in the earlier days of printed finding aids and MARC collection records but that now represent significant obstacles to the aggregation and usability of archival finding aids and resource surrogates. For researchers to use union databases of finding aids and digital surrogates effectively and efficiently, standards for content and structure of resource description must be adhered to by all participating repositories. Without employing such standards, union databases will only be able to serve gross chunks of information. This will become intolerable and useless to researchers as the magnitude of union databases increases and the chunks of information increases from three or five to fifty or more finding aids to browse through. However, application of content and structure standards requires substantial training and modification of work patterns. As many repositories will attest, EAD encoding of finding aids has resulted in adding another work routine in an already labor intensive processing regimen.

The development of a suite of digital tools to support archival processing work and access would help to solve this problem to a substantial degree, although not completely. A suite of digital tools or toolkit could be designed to force adaptation and adherence to extant content standards. It could be constructed so that structure standards are applied automatically in the production of outputs such as EAD encoded finding aids and standardized digital objects (e.g., METS), thereby reducing the need and cost of training. And it can be built to completely or nearly completely automate some routines, thereby streamlining a repository's processing work. But most important a toolkit designed according to the requirements suggested above and described more fully below will lead to more compatible data streams into union

databases and to more efficient and productive use of the those union databases. Such a toolkit will promote and support good research.

BACKGROUND

Sponsored by the Digital Library Federation (DLF) and the California Digital Library (CDL), twenty-one archivists and information technologists met in La Jolla, Calif., on February 4-5, 2002. The purpose of the meeting, known as the Archivists' Workbench meeting, was to discuss the concept of a workbench or suite of digital tools that would facilitate collection and management of information about archival materials at the various points along the life cycle of those collections. Ideally, a workbench would facilitate integration of the disparate filing systems and databases now used in most archival repositories for collecting and managing their archival information, and it would enable more efficient production of various outputs, ranging from encoded finding aids for use by end users to internal administrative reports.

Chief among the meeting's successfully met objectives was validation of a broad need for a digital toolkit that would:

- Create efficiencies in data capture and reuse at various points in repository workflows;
- Reduce barriers to participation in consortial and institutional access systems by making digital encoding for online access a system byproduct rather than a complex additional segment of staff work;
- Reduce educational requirements and training tasks by automating complex encoding procedures and other kinds of work routines;
- Increase application of data content and structure standards, assuring greater interoperability of end-user access products such as encoded finding aids and digital objects; and
- Integrate in one system, serving one or more archival repositories, archival data typically dispersed across several databases and filing systems, digital and analog.

Participants in the February meeting also discussed strengths and weaknesses of a variety of technological solutions that might serve as a possible platform on which to build this suite of tools. In addition, participants considered incomplete or unsuccessful efforts by the archival community during the late 1980s and early 1990s to construct a comprehensive data management utility, as well as the lessons derived from those efforts.

In light of lessons learned from previous unsuccessful attempts to build archival information systems, meeting participants concluded it was extremely important to focus narrowly initial design of an archivists' workbench. Earlier attempts at the creation of such tools had failed in part because they aimed for comprehensiveness of process and participation at the outset. Participants in the Archivists' Workbench meeting decided it would be best focus initial construction and application of the toolkit to a homogeneous group of repositories, smallish archives and special collection units in which one professional is typically responsible for most, if not all, of the archival work. This group was targeted because meeting participants believe such repositories are lacking in staffing resources to standardize their archival processes and contribute their descriptions and surrogates to consortial databases and because publication of the archival materials administered by these repositories would greatly benefit the

research community. In addition, such repositories represent a middle ground between the "lone processor" historical society and the multi-staffed manuscripts and archives unit that exist at a few of the nation's research libraries. Workflows would be easier to discern in those environments, and it would be easier to build upon those results, presuming their success, to enlarge subsequent designs to include a broader range of repositories and more complicated workflows.

Participants in the February meeting also cautioned that this current effort to construct a suite of digital tools for archivists not become paralyzed at the outset due to too grand a vision. They advised that a few key archival functions be targeted. That advice has been considered thoroughly in the aftermath of the February meeting and during the composition of this grant request. The planning process, for which funding is being requested with this proposal, will be devoted in large part to identifying those archival functions that are typical and related and, hence, could and should be accommodated in a toolkit. The objective is not to be comprehensive in the initial design but, rather, to make sure we allow collection of related data when it can be collected relatively easily and enable thorough use of all data collected. Another objective is to build the toolkit with an eye toward facilitating future modifications and extensions. In short, an accommodating design, and not a comprehensive design, is the target of the planning sessions. The particulars of that design will be the product of the planning sessions.

The meeting concluded with the commitment of twelve participants, known as the Archivist Workbench Core Team, to begin defining the functional requirements and system attributes of a workbench by elaborating and specifying the high-level requirements agreed to during the meeting and to join together in a planning process, the objective of which is to define a paper prototype of the archivists' workbench and secure funding for building and testing a working prototype.

DESIGN CONSIDERATIONS FOR AN ARCHIVISTS' WORKBENCH

First among the high level requirements validated at the meeting is that the tool set needs to be informed by the life cycle of an archival collection or item as it progresses through a repository, from first contacts with a creator or donor of the archival materials through completion of the arrangement and description to use of the resource by the research community. However, while it is true that all collections or documents reflect the same basic life cycle, how that life cycle is articulated in one repository may differ in some ways from its articulation in another repository. Work may be sequenced one way in one repository and another way in another repository. One repository may cluster its data differently than does another repository. And one repository may choose not to collect information than another repository believes indispensable.

Second, every archival function typically has two basic aspects. One aspect is the physical labor required to perform the function, such as transferring a set of boxes to the custody of the repository. The other aspect is the documentation or representation of the task and its results. Archival representation is the sum of the recording of the archival work of acquiring, processing, and servicing of archival materials. Historically, data generated from these events has been stored in a variety of locations, some digital (e.g., spreadsheets, databases, word processor files) and some analog (e.g., paper collection files, rolodexes, printed finding aids). As a consequence, the richness of this information and its myriad relationships has rarely been utilized to its fullest potential by archivists and curators.

Third, as demonstrated during the February meeting, there are significant differences across repositories regarding the sequence or workflow of the archival functions generating the representations, not to

mention differences in how repositories represent each function (i.e., character and number of data elements). Meeting attendees agreed that an archivists' workbench would need to be flexible and adaptable to different work environments and able to accommodate different workflows. With minimal customizing, the suite of tools should be deployable on a single desktop in a one-person repository, or on a network serving a larger repository or even a consortium of repositories such as the Five Colleges or participants in CDL's Online Archive of California.

Meeting participants also agreed it was important for the toolkit to accommodate processes and workflows as established by individual repositories, since variance in institutional missions, staffing patterns, funding, and space are important determinants for how a repository represents and sequences its archival work. Accommodating a range of representational practices and workflows is complicated by the probability that not all archival repositories define their archival functions with the same delimiters. This state of affairs necessitates building flexibility into the toolkit that permits implementers to tailor it to their own needs but without compromising archival standards for content and structure that are imperative for developing broadly useful consortial access systems to archival resources. Obviously, it is inevitable that successful design and implementation of an archivists' workbench will require repositories to analyze their local practice and evaluate whether or not changes to those practice would be beneficial; however, the toolkit will enjoy even greater success if it can accommodate a wide range of those local practices and minimize the need for conformity to the toolkit.

The strong consensus reached in the February Archivists' Workbench meeting was that a modular design would best accommodate different work environments and workflows; hence, a blueprint for a suite of tools or toolkit would be the desired outcome of the planning phase of this project.

Modules determined by archival functions or predictable archival representation events allow for sequencing the modules in a manner that best conforms to the actual workflow employed in a given repository. In simple terms, a modular toolkit would consist of input templates and associated program code, storage data tables, and output formats and associated program code. The configuration of input screens would be determined by repository workflows, and they would funnel data to the storage data tables. These storage tables would not necessarily reflect boundaries or relationships suggested by the input templates. When the same data is required in the representation of different archival functions, it would be collected at the first available opportunity in the workflow, stored in a single location in the storage tables and reused for representation of subsequent functions. Data would be entered and stored according to community content standards. For example, controlled access terms would be entered and stored in accord with the principles of the LC Name Authority File, the LC Subject Heading list and other established thesauri. Data structure and transmission standards would be applied on export of information in one of the defined output routines. Output products would minimally include encoded and printed finding aids, standardized digital objects (MOA2 or METS), and cross collection browse lists created by archivists in response to end user queries, but they could also include provisional MARC and DC cataloging records for the collection and selected sub-parts and a wide and diverse set of administrative reports such as shelf lists, or periodic quantitative statements on major functions such as acquisition, digitization, or cataloging.

Effective delimitation of the modules, accompanied by sufficient documentation, should make the suite of tools capable of being implemented differently by different repositories, or of being modified by a single repository through time to reflect changes in the workflow pattern due to changing staff levels or repository goals. In addition, if modules are defined at high enough levels of granularity, it will be possible for modules to be combined in such a way that best reflects how archival functions are defined and represented in a specific repository. Finally, this design approach will enable repositories to use only those modules pertinent to their current workflow. Assuming, for example, that the toolkit includes a

digital object production module, a repository not creating digital objects could elect not to use it at all or use it at a later date when the repository begins to create and upload digital objects.

Participants in the February Archivists Workbench meeting clearly confirmed that the most pressing need at present is a tool to facilitate the output of encoded finding aids to enable online access to archival resources through repository websites and union databases. Nonetheless, participants also agreed that while efficient production of finding aids and other access products should be the primary rationale for building a toolkit, it should not be sole objective for an archivists' workbench. Consideration should also be given to how the archival information might be re-used for other purposes already extant in archival repositories and how it could be adapted to future needs. The toolkit we envision incorporates finding aid production but looks well beyond it to include a greater range of functionality that could result in significant efficiencies for archival workers across the range of archival work and not just for finding aid encoding. For example, we envision a toolkit that, with some adaptation, could facilitate ingestion of electronic records and their associated metadata, as well as other kinds of born digital materials.

A service and maintenance model is the final critical feature for an archivists' toolkit. Meeting participants concurred it would be folly to invest considerable resources in constructing a suite of digital tools and not address how the toolkit will be maintained and modified over time to keep current with technological developments and changes in archival work. A good service model would satisfy several basic requirements:

- Provide training for repositories in the use of the toolkit;
- Provide ample documentation of all component parts of the toolkit;
- Provide assistance to toolkit users with implementing and customizing the input templates and output formats;
- Provide structure and procedure for updating the toolkit in a timely and appropriate manner to keep pace with technological evolutions; and
- Provide a mechanism for tracking all registered users so they can be easily notified of new modifications and features.

FUNDING REQUEST

The Five Colleges, a Consortium in Western Massachusetts made up of Amherst College, Hampshire College, Mount Holyoke College, Smith College and the University of Massachusetts at Amherst, and the California Digital Library request funding of $40,000.00 to support 5 two-day planning meetings over the course of a year for developing the functional requirements, system attributes, paper prototype, and business / service model for a digital toolkit that would embody the objectives agreed to in the February Archivists' Workbench meeting in La Jolla. The meetings will lead to the development of a paper prototype for the toolkit and a grant request for construction and trial implementation of a working prototype.

Team members

A core group of 12 persons will participate in each of the projected 5 two-day meetings. The exact composition of meeting attendees will be adjusted where necessary to bolster the content and fulfill the objectives of the particular meetings.

For the planning phase, a core team of 12 persons will be composed of Five Colleges and University of California personnel and other participants from the original Archivists' Workbench meeting who have volunteered their contributions to this project. The Five College Archivists Group (Daria D'Arienzo, Amherst College; Susan Dayall, Hampshire College; Peter Carini, Mount Holyoke College; Nanci Young, Smith College and a staff member from the University of Massachusetts) led by Peter Carini and Kelcy Shepard will represent the Five Colleges. Robin Chandler, Bill Landis, and Brad Westbrook will represent the University of California. Other members will be Mary Lacy (LC), Merrilee Proffitt (RLG), Chris Prom (Univ. of Illinois), Clayton Redding (American Institute of Physics), David Ruddy (Cornell Univ.), Elizabeth Shaw (Univ. of Pittsburgh), and Elizabeth Yakel (Univ. of Michigan). Archivists and curators from the Five Colleges will also participate in the meetings, contributing substantial information to the first few meetings. It is also expected that other domain experts may be needed for specific aspects of the planning phase; for example detailing storage and platform options. It is not expected that a facilitator will be required during this planning phase of the project.

The core team will be broken in to sub-teams, which will be assigned tasks for the entire planning process or particular meetings. Sub-teams are identified in the description of the meetings below.

Projected Meetings:

Data Modeling (2 Meetings)

On the basis of examining work flows and case scenarios for several archival repositories conducted prior to this meeting, participants will identify a range of archival functions and the data elements used to represent them. Attention will then turn to defining the input templates. This will require specifying which data elements need to be governed by community standards and best practice guidelines.

While primary emphasis will be placed on making sure the templates enable adequate representation of each function, consideration will also be given to customizability and usability of the input forms. Usability, and methods for testing it, will be high priorities throughout the entire project.

Sub-team: Peter Carini, Chris Prom, Kelcy Shepherd, and Beth Yakel will assume responsibility for workflow descriptions and data specifications from a number of archival repositories. They will analyze the information they obtain and present a list of data elements used by surveyed repositories and a descriptive analysis of the variance among workflows. This data will be used in the meeting to determine the number and range of data elements required for the toolkit and basic kinds of workflows that need to be encompassed. This understanding can then lead to productive design of input templates and specification of input or data entry rules.

The sub-team will draw substantially but not exclusively on input from Five Colleges archivists and curators.

Output Products (1 meeting)

The chief concern of this meeting will be defining a variety of output products that will be available in the prototype archivists' toolkit and assessing the products of the previous two meeting to insure that adequate

data has been captured and stored to support the output of these specific products. Attention will also be devoted to the need for support of some degree of customization in these output routines, for example, layouts of printed finding aids and administrative reports.

Sub-team Bill Landis, Merrilee Proffitt, David Ruddy, and Brad Westbrook will identify the outputs enabled by the data elements legitimized at the conclusion of the first meeting. They will present versions of these outputs, with recommended formatting, to the second meeting for modification.

Storage Architecture (1 meeting)

Once the data elements are identified, the input templates developed, and the desired outputs specified, attention will be given to the architecture for storing the data to enable variable outputs. Efforts will be made to identify repeating data elements that need be stored only once and other data elements useful for linking data subsets. Prior to the meeting participants will investigate various technical options for the storage architecture for the toolkit and during the meeting will discuss the strengths and weaknesses of each, deriving specifications that will be used during the development of the working prototype.

Sub-team Clayton Redding and Liz Shaw will present models for how the data is to be mapped from the input templates to storage "containers". This analysis may require the presence of another domain expert.

Platform and Service Considerations (1 meeting)

Participants will come to the last meeting in this sequence of meetings prepared to discuss the advantages and disadvantages of various software environments and platforms that might be used for the toolkit. Arguments will be weighed for it being a relational database or an object oriented database, for it being an SQL or XML database, and for the software being proprietary or open source.

Sub-team Clayton Redding and Liz Shaw will present options for software environment and platform for the toolkit. Again, this component of the planning phase may require the contribution(s) of additional domain experts.

During the second day of the meeting, participants will discuss service requirements and models for the archivists' toolkit. In addition, an inventory of documentation needs for the toolkit will composed, and a strategy for evaluating the utility of the toolkit will be developed. Finally, an attempt will be made to identify institutions interested in sole or joint governance of the toolkit.

Sub-team Peter Carini, Robin Chandler, Mary Lacy, and Merrilee Proffitt will present varying models for governing and sustaining utility of the toolkit. As part of their presentation, sub-team members will attempt to gain some sense of institutional support for each service model. Meeting attendees will evaluate the different models and rank them according to projected success.

Overall Process, Budget, and Outcomes

It is expected the 5 two day meetings will be conducted over a 12 month period beginning shortly after funding is granted. Typically, meetings will be held at either a Five Colleges site or a University of California site (Oakland or San Diego), but meetings may be held at other sites if it is cost effective and convenient to do so.

Meetings will be separated by adequate time to allow for resolution of the issues raised and objectives targeted in the previous meeting and preparation for the upcoming meeting.

Each meeting will be substantially documented. Sub-team members Robin Chandler, Kelcy Shepherd, Brad Westbrook, and Beth Yakel will be responsible for collecting and creating documentation and for synthesizing into a subsequent grant proposal to support construction of a working prototype. The documentation sub-team will record meeting contents and deliver them to meeting participants well in advance of the next scheduled meeting.

Expectations are that each meeting will cost approximately $8,000.00. Each meeting includes a cushion of $400.00 (a total of $1,600.00 to $2,000.00 for the 5 meetings) to invite additional domain experts to particular meetings. It is expected that additional domain experts will only be required in two of the meetings at most. Meetings will be held in either the Five Colleges Area, San Diego, or Oakland, thereby alleviating the need for two airfares for each meeting. Meeting space and technology will be donated by the Five Colleges, the CDL, or UCSD, depending on where the meeting is held.

$4,000.00 Ten airfares at average cost of $400.00 per airfare

$1,800.00 Six rooms for two days at $150.00 per day
(assumption is team members will share rooms)

$1,800.00 Meal per diems of $75.00 for each per team member
(12) for two days

$400.00 Extra guest support
_____

$8,000.00 Total cost per meeting

The projected total cost for four meetings is $32.000.00, or $40,000.00 for five meetings if a fifth meeting proves necessary.

The planning meetings will be extensively documented and result in several discernible sets of data:

- Archival data elements, rules for their entry in templates, and template mock ups
- Data storage model, depicting relationships among data elements
- Output products and the rules for formatting them
- An informed decision for the best platform and software environment
- One or more detailed service models (e.g., centralized vs. distributed) for how the toolkit is to be maintained and sustained over time and an assessment of what kinds of education and training mechanisms will be useful

At the conclusion of the five meetings, team members will use online and telephone interaction to refine these data sets and then knit them together to form a paper prototype of the toolkit. This prototype will serve as the basis for a subsequent and more substantial funding request to support development and trial implementation of a working prototype and a more detailed service model. After the successful

recruitment of programmers, a working prototype will be designed, accompanied with an effective user interface, and then deployed for trial implementation.

It is expected that implementations will be tested at the Five Colleges, selected repositories participating in the OAC, and other repositories represented by the team members. During the implementation trial, the results obtained during the planning phase will be iteratively tested and, where necessary, modified to meet real life practices. Programmers and toolkit implementers will work closely and quickly to optimize the tool's performance. These trials will be accompanied by on-going evaluation and iterative design modification. Evaluation techniques will include usability analyses, on-site observations, surveys, and interviews with archivists at participating institutions.

_____

Note: A single three day meeting, to take place in western Mass., was funded by DLF. (6/21/02)