Archivists' Toolkit: Overview

Outline

Archivists' Toolkit: Overview of Software Specification

Ov-1: Problem Statement

Twenty 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 practices 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 introduced in 1993. HTML served to improve the online presentation of finding aids; however, its limitations for facilitating searching and navigating online finding aids became quickly apparent to archivists. Furthermore, HTML did not help to promote consistent application 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 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 have gained footholds in New Mexico, Texas, Virginia, and North Carolina, not to mention several international projects.

No doubt, this development is highly laudatory and a great benefit to the general research community. It has sparked the development of best practice guidelines and a wide range of tools for producing finding aids in an automated environment. About those tools, several general characteristics are noteworthy. First they are often under-utilized. A database, for example, may be used only to encode finding aids but not to support other archival tasks. Second, the tools often are not integrated. As in the previous print environment, each archival function or task typically has its own tool. There is a database for accessions. Another for donors. A word document for tracking locations. And so on. As a consequence, some data expressions, most obviously a resource title or identifier, needs to be rekeyed several times and stored in different places.

****

Updating the data often requires updating it at each unique storage point. Finally, the tools are highly localized and not designed in a manner that promotes standardization and interoperability beyond a repository's immediate institutional boundaries and needs.

The deployment of numerous, single-purpose tools increases the cost of archival processing in several ways. The work flow is inefficient because the same data has to be re-entered at various times during the description process. Training costs are increased, as staff has to be knowledgeable about using each distinct tool. The breadth of skills and training required for proficiency with all the tools utilized in the descriptive process can prohibit assigning some descriptive work to lower staff levels. The tools collectively have to be managed and kept up to date.

There are hidden costs as well. For example, one consequence of encoding tools that do not promote or enforce standardization is inconsistent data and, thus, union databases that cannot be searched or navigated at fine levels of granularity. The promise, or at least one of the promises, of the encoding is not realized. Additionally, increased costs for descriptive work absorb valuable resources from an archive's operational budget (rarely large), resources that could be used for other archival functions such as collection development, fundraising, and reference service.

The development of a generally deployable digital application, such as the Archivists Toolkit (AT), to support archival processing work could serve to lower processing costs dramatically, to promote standardization of archival information, and to foster development of more robust union databases of archival information. An application could be designed to "push" adoption and adherence to extant content standards. It could be constructed so that encoding standards are applied automatically in the production of outputs such as EAD encoded finding aids and METS digital objects, thereby reducing significantly the need and cost of training. And it could be built to automate completely, or nearly completely, some routines for managing archival information, thereby streamlining a repository's processing work. But most importantly, a toolkit designed according to the objectives suggested here 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. In short, such an application would foster and support good research.

The ultimate objective of the AT, as described in this software specification, is to reduce the costs of archival processing by facilitating more efficient work flows and quicker throughput of archival information. The AT will do so by integrating key archival functions into an integrated application environment. This will make it possible for data about archival materials to be more easily repurposed and output in different formats to support different needs.

****

In addition, the AT, due to its adherence to archival content standards, will contribute to the standardization of archival information, to the extent the AT is implemented by various archival repositories.

To promote its acceptance, usability, and development, the AT must be based in the standards essential to the creation and communication of archival information. The application must be open source, and it must be modular, allowing repositories to use only the functional areas they need to support their local work. The interfaces and outputs of the application must be customizable, allowing repositories to configure the application to their basic work flow and staffing structure rather than, what is often more typical, adjusting work practices to fit the design of an application.

Reducing archival processing costs and increasing data standardization will benefit researchers by allowing archival materials to be described more quickly and by promoting standardized access tools for archives such as EAD finding aids and METS records.

Ov-2: Purpose of the Software Specification

This specification is intended primarily for the AT Software Design Team, but it will also be shared in varying forms with the AT partner repositories and with members of the general archives community in order to elicit their comments on the specifications formulated for the application.

The software specification stipulates the features or functions the application is to support, drawing on tasks identified in collaboration with the project's partner repositories. The specification describes the high level features of the application: open source, modular, customizable to work setting. The specification also describes each functional area in detail, indicating the task sequences supported by the functional area and then the data inputs, computer processes, and data outputs required to satisfy the tasks. Provisional screen prototypes and entity relationship diagrams are provided as part of the description of each functional area.

Ov-3: Scope of the Archivists' Toolkit

The Archivists' Toolkit does not attempt to accommodate all archival information. Rather, the application will address only the following essential archival functions: accessioning; location tracking; source registration (names of donors); description of items, collections, and surrogates; and application of authoritative names and subject descriptors. The application will allow for ingesting legacy data in the form of MARCXML and EAD v. 1 or 2002 finding aids.

****

Satisfying these key functions will support typical archival tasks such as recording accession transactions; multi-level description of archival resources, including authoritative forms of names and subject terms; and shelving and retrieving archival materials. It will also support production of access outputs, such as finding aids (EAD encoded and printed), catalog records (MARC, DC, OAI), METS records for digital surrogates, and a range of administrative reports, including collection profiles and production statistics.

The application will be offered as an open source application, with a Web-based interface, designed for deployment as either a stand alone or network application on the Linux, Mac, and Windows platforms. The application's modular design will allow for local customization of input templates and output formats.

The application does not include functions for managing or using any of its outputs, especially encoded outputs such as EAD finding aids, MARCXML catalog records, or METS encoded digital objects. Storage, searching, and management of such outputs is handled by external systems. The AT is simply a production tool that will produce objects that will work in those external systems.

Ov-4: Acknowledgements

This effort to design and build an application for managing archival information would not be possible without the assistance and support of a many people and organizations. The project team is deeply indebted to the generous support of The Andrew W. Mellon Foundation and Donald Waters, who provided acute readings of early drafts of the proposal.

The AT project team would also like to express its gratitude to members of the Advisory Board and to the AT partner repositories, all of whom contributed information toward the completion of the specification.

Advisory Board

Robin Chandler, California Digital Library
Michael Fox, Minnesota Historical Society
Merrilee Proffitt, Research Libraries Group
Richard V. Szary, Yale University
Guenter Waibel, Research Libraries Group
Beth Yakel, University of Michigan

****

Partner Repositories

Five Colleges, Inc.

Amherst College Archives and Special Collections (Daria D'Arienzo, Peter Nelson)
Hampshire College Archives (Susan Dayall)
Mount Holyoke College Archives and Special Collections (Jennifer Gunter King)
Smith College Archives (Nanci Young)
Sophia Smith Collection, Smith College (Sherrill Redmond, Margaret Jessup)
Special Collections and University Archives, University of Massachusetts Amherst, (Robert Cox, Danielle Kovacs)

Participating New England Area Archives

The Edmund S. Muskie Archives and Special Collections Library, Bates College (Katherine Stefko)

New York University

Fales Library, NYU (Ann Butler)
NYU University Archives (Nancy Cricco)
Tamiment Library, NYU (Mike Nash)

Participating New York City Archives

The American Museum of Natural History (Barbara Mathé, Kristen Mable)
The Brooklyn Museum of Art (Deb Wythe)
Carnegie Hall (Kathleen Sabogal)
The Center for Jewish History (Bob Sink)
Manhattan College (Amy Surak)

University of California, San Diego

Mandeville Special Collections Library (Steven Coy)
Scripps Institution of Oceanography Archives (Deborah Day)

Daniel Greenstein and David Seeman, respectively former and current Directors of the Digital Library Federation, secured funding for two early exploratory meetings held at UCSD and at the Five Colleges in 2002. Those meetings were important for establishing the basic requirements and features of an archival information management application. The following persons all attended those meetings and have in some measure influenced the work represented by this specification.

First meeting, February 4-5, 2002

Peter Carini, Mount Holyoke College
Robin Chandler, Online Archive of California

Morgan Cundiff, Library of Congress
Michael Fox, Minnesota Historical Society
Bernie Hurley, University of California, Berkeley
Mary Lacy, Library of Congress
Bill Landis, University of California, Irvine
Bertram Ludaescher, San Diego Supercomputer Center
Stephen Miller, University of Georgia
Regan Moore, San Diego Supercomputer Center
John Ober, California Digital Library
Merrilee Proffitt, Research Libraries Group
Chris Prom, University of Illinois
Clayton Redding, American Institute of Physics
David Ruddy, Cornell University
Elizabeth Shaw, University of Pittsburgh
Kelcy Shepherd, University of Massachusetts, Amherst
Mackenzie Smith, Massachusetts Institute of Technology
Brian Tingle, California Digital Library
Brad Westbrook, University of California, San Diego
Stephen Yearl, Yale University
Beth Yakel, University of Michigan

Second meeting, November 4-6, 2002

Peter Carini, Mount Holyoke College
Robin Chandler, Online Archive of California
Mary Lacy, Library of Congress
Chris Prom, University of Illinois
David Ruddy, Cornell University
Kelcy Shepherd, University of Massachusetts, Amherst
Brad Westbrook, University of California, San Diego
Beth Yakel, University of Michigan

And, of course, we wish to thank our immediate library administrators. Lorna Peterson and the Librarians' Council of the Five Colleges; David Ackerman, Carol Mandel, and Jerome McDonough of the New York University Libraries; and Luc Declerck and Brian Schottlaender of the UCSD Libraries. They have allowed us to take this opportunity and have provided the project team with very useful advice throughout the first year of the project.

****