How Do We Get This Online?

A Step-by-Step Guide to Converting Archival Catalogs into
Customized XML for Online Presentation

Introduction

In January 2009, the Massachusetts Historical Society (MHS) was awarded a grant by the National Historical Publications and Records Commission (NHPRC) to digitize a fifty-year-old paper catalog into an online database. The database would be used concurrently by the public wishing to access information about Adams family documents held by the MHS and elsewhere and by the documentary editing project, The Adams Papers (housed by the Society), to better facilitate and improve their workflow in publishing the papers of John Adams and his family. One of the goals of the project was for it to serve as a model to other institutions with similar item-level catalogs. The following guidelines detail how this project was successfully completed and how others can replicate these basic methods at their home institutions.

The Adams Papers Control File Digitization: What We Did

Over a fifty-year period, the Adams Papers documentary edition created an item-level, color-coded catalog over fifty years. It tracks the physical archive held by the MHS as well as every known Adams document held in other repositories and any Adams document that has been auctioned off in private circles over the past hundred years. There are about 110,000 slips in duplicate, organized both alphabetically by correspondent and chronologically. In 2001, the entire chronological run of the catalog was microfilmed for preservation purposes. There are 42 reels of microfilm.

The project duplicated the microfilm and sent it to an offsite data entry vendor, DataStream Content Solutions. The vendor converted all microfilm images into high resolution PDF files. Those were read by human keyers and optical character recognition (OCR) technology to create 42 Extensible Markup Language (XML) files of encoded text. Open-source XML was chosen for the database because of its flexibility in marking-up text and ability to organize disparate data. The vendor, based on detailed instructions created by the project manager, marked up every slip with a unique identification number and each piece of information was encoded with a one to two-letter tag that was later expanded into semantically logical elements (i.e. date, author, etc.). Providing clear, specific instructions to the vendor was essential for the return of quality data with a minimum of errors. We were also realistic about the tasks assigned to the vendor, i.e. several of the more problematic fields were lumped into a general “notes” category and later parsed out by encoders.

Upon receipt of the XML files, the project staff completed three phases of work. The first was a character-by-character proofreading of the text itself. In the second phase, we created a set of Extensible Stylesheet Language Transformations (XSLT) to manipulate the data. These XSLT stylesheets provided a set of instructions to automatically encode the more numerical and basic information (dates, codes, page numbers, etc.). We then checked the encoding record by record. The third phase tackled the more intellectual content encoding such as names, titles, locations, and citations. These were first encoded using Extensible Stylesheet Language (XSL), which allowed us to transform and render the XML documents in a specific format, and then checked for accuracy by the encoders. Concurrent with the last phase, we created five supporting databases that would be used to apply controlled authorities to the individual records.

Finally, the XML phase of the work was completed and all 42 files were moved into a MySQL database. This backend database manages the workflow of updating and improving the records. Searching is handled by an open-source web search engine, SOLR, which is robust and extremely fast in delivering results.

The entire project was completed primarily by a project manager, a half-time proofreader, two half-time encoders, and a web developer. All administrative and budgetary concerns were handled by the project director. All of these individuals are MHS employees working on other projects as well. There were several other key staff members of the MHS that contributed between five and ten percent of their time at various stages of the project.

The Adams Papers Control File Digitization: What We Learned

Step 1: Funding

Adequate funding is critical for any project that seeks to bring analog material into an electronic format. But the dilemma familiar to many non-profits is that you cannot secure funding without a clear plan in place and you cannot establish a clear plan without adequate funding. Thus, the case must be made to the home institution that key employees must be given an opportunity to assess the project parameters and plan ahead to save time and money down the road.

The Adams Papers Control File Digitization Project was funded by several agencies that already contribute significantly to the documentary editing project. The NHPRC awarded the bulk of the money in the form of a strategies and tools grant. The Packard Humanities Institute supported the initial project launch prior to the NHPRC grant, which had been delayed due to federal funding cycles. It allowed the project staff to receive key training and covered half of the $33,000 cost of the data entry vendor. The remaining funding for the project was committed by the home institution, the MHS, which provided 55% cost share, including the long-term commitment to host the final product on its website.

Depending on the home institution, some projects may have grant writing resources through their development arm. However, for smaller independent projects that do not, funding could be secured for different stages of the project that build on each other. For example, you could approach a donor to cover the relatively small costs of planning the project and writing a larger grant proposal, while the larger grant could then cover staff time and vendor costs.

Step 2: Planning

The importance of planning ahead cannot be overstated in digitization projects. There will always be unforeseen flaws in your source material and hurdles in finding the right software and online delivery systems, but with proper planning, these issues can be tackled in a more efficient and thorough manner. It is also critical to document your plans and policy decisions, especially as they change. And they will change.

For this project, the staff created a blog to track our encoding guidelines and update staff on policy changes. It was also a good way for our funders to track our progress between the formal reports submitted at six-month intervals. The blog can be viewed at http://adamspaperscatalog.blogspot.com/. Another planning format that could be used is the wiki model. Just as easy to set up, the wiki model allows for easy searching of encoding rules that you've established and resolutions to problems that arise.

Another important reason to document the evolving plans is that staff will change. The project was a three-year endeavor (with the NHPRC grant covering two of those years) and experienced several staff changes. Having clearly documented plans at each stage of the project kept us on pace despite staff upheavals.

In addition to documenting your plans, it is important to remember that *planning ahead is an activity that should occur throughout the entire project*. A granting agency will rightly ask for the long-term work plan. This document is written at the outset and should be constructed loosely to allow for adjustments. The most efficient way to reach your end goal of an online catalog is to tackle each phase of work in its own time and build on it. Anticipating each new phase of work, the project staff calculated how long the task would take in hours. At the end of each phase, we would audit the work just completed and use that knowledge to project our next phase. In this way, we have been able to more accurately predict our schedule as we've progressed. Many of these types of projects won't stick to the original schedule, but knowing exactly how far off schedule you are will be extremely useful if you need to apply for an extension to complete the work.

Planning: Identify the Source Files

Identifying the source files does not simply mean knowing that you'd like to convert the old Smith Papers catalog into on online database, it is understanding what exactly you will be working with. Is it typed information or handwritten, or a combination of both? Are they traditional card stock catalog cards or thin slips of carbon paper like the Adams Papers file? Do they exist only as cards or have they been reproduced in some manner? Have they been photocopied or microfilmed? The Adams Papers had microfilmed the entire "slip-file" in 2001, so that was our starting point. Is the file regularly consulted and updated by editors (as in the Adams Papers) or a static inventory that will not change (see the Saltonstall Collection, below)?

If your source material only exists in one place, you will need a duplicate in some form. Ideally, the original source material should not be the working copy, especially if it is regularly consulted by researchers. Also, depending on if you will be using an offsite vendor to assist in the initial data conversion, you will want to consider the costs associated with mailing large amounts of material. Reams upon reams of paper add up in shipping costs, but a box of 42 microfilm reels is quite affordable. Many data entry vendors are equipped to use microfilm as a source file.

Planning: Identify a Data Model

Due to the unique nature of archival materials, archival catalogs never follow an exact pattern. They are unlike book catalogs and yet they usually have some inherent order that can be found upon closer inspection. In many ways, Extensible Markup Language (XML) is the perfect mark-up language for archives because it can accommodate the organic nature of archival materials. If your material just doesn't quite fit into a structured database, XML is the perfect solution. And if you eventually find that your XML data does fit into a relational (structured) database, it can be converted. Additionally, non-proprietary XML has the added advantage of being compliable and sustainable, and is becoming the "industry" standard for humanities publishing and organization.

The key to identifying a data model is consistency. Find the consistent patterns and fit them into a hierarchy of data. Is every card dated? How do the dates look? Are they constructed in a consistent manner? Do all the cards have at least one correspondent? Are titles or abbreviations employed? Are any identifying features noted? Page numbers? Is the type of document, a draft or recipient's copy? Citation information? Are there any internal codes in use? Are they applied uniformly? Once you've identified a pattern, the other major question to answer in archival catalogs is whether you have a one-to-one match to the documents. Many archives catalogs are simply item-level records of one discrete collection. But there are some catalogs that represent both physical and intellectual collections of documents. The MHS has both kinds.

The Historical Society houses the Saltonstall Family Collection and recently converted its item-level control file into an online database using the Adams Papers model. It presents one record for each physical document housed in the MHS. However, the Adams Papers control file presents a much more much complicated challenge. It tracks the physical documents in the Adams Family Papers on colored slips of paper. But it also tracks if those same documents have been printed elsewhere on different slips of paper. So it is not a simple one-to-one match. For this reason, we had to create a database of the control file, not of the documents themselves. This is a more complicated database, but it is a lesson in efficiency: adjusting to and working with the existing data was much more efficient than reconstructing a catalog of over 100,000 records that was created over 50 years. Ultimately, the technology will aid us in presenting a database that is easy to use and understand much more than re-inventing the wheel would have done.

Once you have constructed a model of how each record is supposed to look, it is time to define the structure and content of the XML documents with a schema. This schema will provide the semantics that will be used during the data entry phase, either by in-house staff or an outside vendor. The Adams Papers project chose the RelaxNG schema type because it was simple and easy to learn and must be written in XML. It did not require any other programming knowledge. This basic schema was then used by the data entry vendor and eventually expanded to create the full schema for the database (See addenda).

Planning: Create Encoding Guide

The Encoding Guide will be an extremely useful and constantly changing tool for all the people working on the conversion. During data entry (particularly if you have an outside vendor), the Encoding Guide will serve as part of the contract for services and is your opportunity to explain exactly how you want every piece of information treated. In completing this phase of work, it is important to have simple, clear instructions to produce reliable data from the vendor.

An example from the Adams Papers project's vendor Encoding Guide states:

> Every record has an author name (or initials) on the second line under the date. All text that appears *before* the first instance of the word "to" should be considered an author and coded under one author <a> tag. If there is no clearly defined "author to recipient" statement, the second line under the date should be tagged as a title, <ti>, see example #5, below. To conform to the RelaxNG schema, there should always be either an AUTHOR and/or a TITLE present in each record.

Since almost all Data Entry Vendors calculate their prices based on the number of keystrokes, the initial schema employed one- or two-letter tags to reduce the number of keystrokes. As quoted above, <author> is shortened to <a>. This can be very easily expanded at a later stage using an XSL stylesheet.

Planning: Identify the Easiest Path to Data Entry

Converting text on paper into electronic text is the first step, but to avoid headaches later on, it is important to clearly articulate the path of conversion. If you will be handling the data entry in-house, you should identify the keyers and their workspace. Will they be typing everything by hand? Will they be cleaning up text produced with OCR software? If possible, you should complete a sample set to estimate the time it will take to complete this step.

It is worth noting again that it is helpful to track the time it takes to complete each phase of work throughout the duration of the project. Knowing how long it is taking staff, whether it be proofreading 500 cards in an hour or encoding 50 cards in two, will assist you in adjusting workflows and priorities as needed. Remember, no project will play out exactly as you've described it in the proposal, but it is extremely useful to anticipate the length for completion once you've begun a phase of work.

If you will be calling on the services of a Data Entry Vendor, it is particularly important to map out the path of conversion. Conducting an RFP from a pool of vendors is an excellent way to identify the best vendor for your needs. We asked the potential vendors to do a test sample as part of the selection process and requested recommendations from previous clients. It is also important to be clear about what they will need. Generally, the vendor will want the source material in some visual format that they will project on a screen for their keyers. In the case of the Adams Papers project, we sent our vendor 42 reels of microfilm that they converted into PDF files. They used a combination of OCR technology and hand keyers to type up the cards and complete the initial markup in XML. Within the first two months of the project, we had all of our source text in flat XML files with basic encoding completed. Although hiring an outside vendor was a relatively large amount of money (22% of the NHPRC grant), it was well worth it given the time it would have taken in-house staff to complete.

Much of this decision depends on the size of the catalog. The Adams Papers control file is 110,000 records and would have taken enormous staff time to complete in-house, both for sheer numbers and setting up the infrastructure to work off of the microfilm. The Saltonstall Collection numbers only about 3,000. This much smaller collection wasn’t worth outsourcing, so the Saltonstall project leader compressed several steps to complete data entry with the encoding.

Step 3: Execution

At this outset of executing the work, we highly recommend creating checklists for each stage of work. The checklists should serve as a quick companion to the larger Encoding Guide and should include the most detailed steps needed to be taken for completion of each phase of work. As you will see below, we “chunked” out the material into 42 files of work, both in paper and XML, and for each of these files, we attached a checklist to make sure every phase of work was completed the same way (See addenda). This included exactly how electronic files should be named, where they should be stored, and which pieces of information must be checked for each stage of work. This aided work in several ways, keeping people on track when wading through over 100,000 records, as well as keeping new staff in line with established practices. Consistent application of the Encoding Guide and project policies is critical to the ultimate efficacy of the database.

Execution: Staffing

Finding the appropriate staff for a project like this is critical. The first step is understanding the workflow and what kind of experience and skill set is required for each phase. Some projects would do well to hire more technically-minded people to tackle encoding at this level, however a case can be made that some catalogs require contextual knowledge of the material. The Adams slip file required both. Because the control file is a complex and idiosyncratic filing system that only makes sense once you start using it, we felt it was important to hire several individuals familiar with the file and how it was originally constructed.

In total, the work of proofreading, encoding, and populating the supporting databases was completed by the project manager, one proofreader/encoder, two separate encoders (the second replaced the first), and an archivist/encoder with XML experience. The in-house web developer was consulted throughout and has done the bulk of the work in pulling the XML into the various web presentations and construction of the supporting databases. We also utilized an XSLT consultant throughout the project for a total of five working days.

Project oversight and guidance was conducted by the project director and a team of various MHS employees with either technical knowledge or specific Adams-related knowledge. The team met on a quarterly basis to keep the project on track and set priorities as the work progressed.

Execution: Proofreading

While no data is perfect, it is important to establish a level of trust in the source files. For the Adams Papers, although it had been built over generations of editors and undoubtedly included errors, we had to simply decide to trust the original records and reproduce them exactly. It was entirely outside the parameters and priorities of the grant to correct errors on the source files at this time. The best solution to that would be to build a system that allowed for easy correction later on, as we discovered them. Thus, we decided to proofread the entire file against the vendor’s XML files, paper against paper.

Using the 42 XML files received from the vendor, we created a very basic XSL transformation that pulled all the information in each XML record and arranged it to mirror the physical slip and printed it out through a web browser. This layout would allow for faster proofreading. We then printed out all 42 files and proofread each record against the physical slip. We used this opportunity to document the color (an incredibly useful piece of information that could not be captured from black-and-white microfilm) and conduct an inventory to include records that may have been skipped.

Proofreading was the one phase that took twice as long as the initial estimate, but it proved extremely useful to verify and correct the data because we used XSL stylesheets to complete automated markup. If the text is clean and trustworthy, then the stylesheet logic will apply correctly and will save time later on.

Execution: Encoding

Encoding is a task that most will have to learn as they go. It is rare to find someone who knows both the material and the intricacies of XML markup, so one aspect of the knowledge will need to be gained on the job. Structuring the work to allow the 'easy' material to be tackled first will allow for a more seamless familiarity as the work continues.

Encoding was separated out into two levels. The first level focused on data that was easily determined, either through automation or by human encoders. It did not require contextual knowledge of the material. We determined, as noted above, that the entire project staff would have to learn the intricacies of encoding on the job, so we tackled the simpler, more straightforward data first. This included dates, codes, page numbers, colors, places, and copy types. At this stage, encoders moved from working with paper printouts to editing the electronic files themselves, inputting changes directly in the XML editor Oxygen. The encoding work consisted primarily of scrolling through the records one by one, adding or fixing attributes and tags. Detailed checklists for the encoders ensured that they didn't overlook anything. To manage the workflow and maintain version control—a significant issue with a project this size—the team used the version control software TortoiseSVN.

For both levels of encoding, we hired an XSL consultant to write a stylesheet that would automatically populate as much markup as possible. To do this, we created detailed logical analyses of the data that the consultant could then turn into an XSL transformation. For example, codes are always found in the top right corner of the slips. The vendor marked them as <c>. They generally consist of a library cataloging code and a number assigned in-house. These two pieces of information are separated by a colon. We instructed the consultant to change the <c> to <code> and to find two pieces of information in the text of that element and use it to populate the attributes, or controlled information. Thus, he wrote an XSL that took <c>MHi:1234</c> and converted it to <code type="accession" institution="MHi" number= "1234">MHi:1234</code>. This information was then verified by the encoder. It saved a great deal of time to automate this and allowed us to focus our encoder's time on more questionable data in level 2.

The second level of encoding focused on the variable data that would require more interpretation on the part of the encoder. As with the first level, we automated as much as possible, but due to the greater variations in the text it needed much more correction from the encoders. Encoding level 2 included the mark-up and improvement of data for authors, recipients, titles and bibliographic citations. The encoders also added special category attributes to about 9,000 slips, making it possible for editors to limit searches by type.

Concurrent to the encoding work, the encoders populated the supporting databases to manage supplemental information. These supporting databases represent people and organizations (over 14,000), places and locations (over 3,000), accessions (over 27,000) and institutions holding those accessions (over 500). The remaining supporting databases contained abbreviations for bibliographic citations found on the slips, i.e. short-titles. The encoders added new entries whenever they appeared and assigned attributes to those entries. With three encoders working concurrently, these databases were essential for sharing information and maintaining consistency across files.

The data was enriched with links to additional information not available on the original slips. We were able to focus on the project objective of adding cross-references to at least 30% of the records in the new catalog to the printed, online and microfilm editions, or websites of the appropriate repository. This included:

- providing specific reel-level locations to the 608-reel microfilm edition of the Adams Family Papers at the MHS for 74,000 records;
- adding direct links to online finding aids, when available, and collection records in the library's online catalog for almost 4,500 documents
- identifying over 4,000 records directly linked to the Adams Papers Digital Edition;
- linking approximately 1,500 records to the Adams Electronic Archive;
- directly linking over 23,000 records of documents not owned by the MHS to the repository's homepage.

Thanks to this added content, all of the information related to a single document can be pulled together and displayed when that record is retrieved.

The ability to revise data and cancel slips is essential for the Adams editors, so it was important that the database be designed as a dynamic catalog with a robust editing interface. XML files were moved into a MySQL database to manage the additions and updates to records. Once the basic encoding work was completed the work of cleaning up and cross-checking cancelled and re-dated records began. This required tracking every slip with a "z" attribute (added during the initial data entry because the slips were literally crossed out) and finding the newly corrected slip and linking it with a corresponding "y" attribute. There were about 460 corrections made. Inconsistencies in the original data also produced discrepancies in the attributes for individual names and dates. To reconcile these errors so that search results would sort correctly, ongoing checks and revisions took place outside of the scope of the grant. The attribute database of authors and recipients was built during the second phase of encoding and required significant revision after the files were loaded into the online delivery system.

Step 4: Completion

The web developer loaded the 42 XML files into a SOLR database for beta testing. SOLR, an open-source search platform by Apache Lucerne, handles the robust searching capabilities necessary for such a large database. This transition away from XML was an important milestone in the project, and the timing had to be right; from this point on, records would be edited individually, using drop-down boxes in the editing interface, and global changes would be more difficult. The end-user interface was designed with two sides, one internal for use by the Adams Papers editors and key library staff and one external for the wider public.

After the decision was made to move our work into MySQL to manage the workflow and supporting databases and SOLR for searching and browsing, the webmaster devoted the final months of the grant to building the internal web-interface. In addition to the interface design, the webmaster spent much time constructing a reliable back-up system for the 100,000 plus flat XML files that stores the data. SOLR, the search engine, indexes the XML and by virtue of it being on the web server is backed up at two offsite locations once a week. The workflow (versioning/editing of records) is backed up in MySQL twice daily. Once the internal (and more complicated) end-user interface was completed the webmaster focused on constructing the external interface as a streamlined and simpler version of the internal one. When the data went "live," the encoders then used the end-user interface (powered through MySQL) to add the final piece of data enrichment, document type subject headings.

With the data viewable in a browser, the slip file team could hold meeting with MHS staff and solicit feedback about the two interfaces. Adams Papers editors met with the team to discuss the more complex internal interface—its features and bugs, available fields, searching capabilities, etc. Building and refining this dynamic interface, with its complicated editing functionality, was the most difficult part of the project, and it would not have been possible without a full understanding of the Adams Papers editorial process. External beta testers, including Adams scholars and documentary editors at other founding families projects, provided valuable input about the public interface and suggested tools to make it user-friendly for library patrons. For the public interface, the assistance of the MHS library staff was essential in designing the detailed search tips and help screens.

Addenda:
Encoding Level 1 Checklist
Final Schema