Acknowledgements:

The Making of America has from its inception been a highly collaborative effort, and this attempt to summarize some of its history and methods is no less so. I am particularly indebted to conversations with and documents prepared by Wendy Lougee, Assistant Director for Digital Library Initiatives at the University of Michigan and John Price-Wilkin, Head, Digital Library Production Service, University of Michigan, as well as current and former staff of the Digital Library Production Service, many of whom will hear their voices echoing here.

Building A Digital Library: The Stories of the Making of America

The University of Michigan Making of America (MoA) is a digital library of primary sources in American social history from the antebellum period through reconstruction. It was born out of a major collaborative endeavor between the University of Michigan and Cornell University, initially funded by the Andrew W. Mellon Foundation. This effort aimed to preserve and make accessible through digital technology a significant body of primary sources in United States history. MoA also seeks to develop protocols for the selection, conversion, storage, retrieval, and use of digitized materials on a large, distributed scale.

The U of M MoA collection is a collaborative effort within the University Library, involving staff from Collection Development, Preservation, Technical Services, and the Digital Library Initiative. Primary responsibility for the production of the MoA system lies with the Digital Library Production Service (DLPS). The U of M collection contains approximately 1,600 books and 50,000 journal articles with 19th century imprints, a total of over 630,000 pages. The selection of materials for inclusion focused on monographs in the subject areas of education, psychology, American history, sociology, science and technology, and religion and periodicals of literary and general interest. These texts were chosen through a process in which subject-specialist librarians worked with faculty in a variety of disciplines to identify materials that will be most readily applicable to research and teaching needs.

The Making of America has enjoyed enormous success both within the scholarly community and with the general public. User reception of the searchable pages available at the site has been overwhelmingly positive: materials previously unused and in storage are now being searched as many as 120,000 times a month, and users are displaying more than 75,000 pages each month. Further, other institutions have begun to adopt the U of M deployment strategy in their preservation efforts. MoA serves as a model for conversion that accommodates both automatically processed and carefully prepared (proofread and fully encoded) materials, where journals can coexist with monographs, and where preservation and access are equally well supported. The MoA system is not only extremely well received by users but is also being embraced as a model by other institutions, such as members of the Digital Library Federation (DLF).

MoA is extensively used, both by users the developers anticipated and by those we did not. MoA has been used in history classrooms at the University of Michigan and at other institutions. Faculty and graduate students from all over the country use the collection for their research. Scholars at the *Oxford English Dictionary* use MoA to search for earliest uses of words. In addition, MoA has seen significant use from more surprising audiences, such as genealogists, hobbyists, and literary societies.

The breadth of uses and users of MoA has been one of the more surprising and rewarding aspects of the project. As was originally anticipated, MoA is an electronic research repository serving historical scholars, and as such is part of the academic library context out of which it grew. But because of its appeal to the general public, it also serves another mission. This broad, valuable, freely available collection serves as a public digital library, providing useful and interesting materials to patrons of all backgrounds and levels of expertise.

At this point, reading about a digital collection in a paper volume, perhaps far from your Internet connection, you may well be asking, "but what does it let me *do*?" It lets you do many different things: You can find materials on premature burial, the *code duello*, phrenology, and "fancy fairs" with simple phrase searching. You can compare the collocation of the words "virtue" and "vice" near "poverty" in articles in 19th Century journals. You can put together a bibliography of abolitionist tracts or look at illustrations from reports on explorations by the US Army Corps of Engineers. In short, MoA gives you the ability to search thousands of pages of full text very quickly and in a number of ways. The sample screen images that accompany this article give examples of some of the possibilities for uses of MoA, as well as the flavor of the materials in the system.

The Stories of the Making of America

The story of the development of the Making of America collection is not a single linear narrative. Instead it is a collection of stories that have a common focus but very different points of view. There is a genealogical story that traces the attempts to build a stable coalition of institutions dedicated to the project, of negotiation and collaboration and the productive tension between the aims of preservation and access. There is an architectural history of the building of a digital collection that will be able to grow into a true digital library, not simply remain a special collection. And within that there is a technical report detailing the methods for creating a viable and sustainable online system with a high degree of functionality and the possibility of scaling to accommodate future growth. Although this discussion can not do justice to the complexity of all those stories (and there are no doubt others to be told as well), it hopes to suggest some of the key points of all three.

A Project Genealogy

The Making of America was initially conceived as a multi-institutional partnership with emphasis on the digital conversion of monographs. In an early and ambitious iteration the project partners proposed the staged conversion and deployment of 100,000 volumes from a 100 year period

(1850-1950). From early exploratory discussions to the initial work on conversion and deployment, the project went through a three year period of definition, refinement, and negotiation among possible partner institutions, as well as active fund seeking. In the end, the Making of America emerged as a project with firm commitments from the University of Michigan and Cornell University and with funding from the Mellon Foundation.

Cornell and Michigan shared a common set of goals in their development of the MoA collection. The collection was built to facilitate large-scale preservation of 19th Century American materials. Within that aim, the institutions wished to base the collection on a clearly articulated and intellectually viable *selection* process, to create an efficient *conversion* process, to build a functional *access* system, and to create a system that could undergo *evaluation* through actual use. Although the institutions began from this common point, they brought to the project different histories and experiences with digital projects. These experiences led to differing emphases within the goals and differing outcomes in the development of the MoA systems.

In a number of significant early efforts, Cornell demonstrated the value of being able to scan and print preservation-quality images. In efforts involving materials such as mathematical monographs, Cornell was able to create print replacement copies while storing valuable digital surrogates. In light of this successful earlier experience, in the MoA project Cornell continued to focus its energy on the specifications and quality control of the conversion process and page image production.

The U of M brought to MoA a history of creating full text access to digital materials. In the TULIP project and in UMLibText, the U of M Library had already created a number of successful methods for searching and displaying text in digital form (Lougee, 1998). Work done at the University of Michigan for the JSTOR project had also demonstrated the value of marrying preservation standards and access strategies of OCR and indexing (Guthrie and Lougee, 1997). Added to this was the expertise of the U of M Humanities Text Initiative in dealing with monographic materials. Building upon this combination of experiences, the U of M was able to commit itself to creating a highly functional access system that allowed fast searching, retrieval of relevant pages, and access to both bibliographic and full text information, as well as more conventional page turning and browsing mechanisms.

These divergent paths have led to quite different outcomes in the systems. Cornell has produced a significant number of replacement volumes for the items it disbound for the MoA collection. At present it provides online access to a sample of the journal volumes it has converted. These volumes are navigated by browsing the tables of contents of the original volumes and using a "go to page" mechanism to locate a desired article. The U of M Library has embraced a different model; while not putting a premium on reprints (U of M anticipates making reprints in the future only on demand), it has put online the entire body of its converted materials and built a system that allows a variety of ways of accessing and navigating the materials.

In the end, Cornell and U of M decided to at least temporarily eschew the difficulties of creating a fully integrated system. The two institutions shared a procurement process and have communicated about their plans and methods for online implementation. They have remained

committed to facilitating searching across the two collections. Cornell has recently made a decision to adopt the U of M access model and has contracted with DLPS for the OCR of their page images and for implementation of an access system. When this process is complete, the partner institutions can move back to a vision of an integrated collection. After an extended period of independent exploration and development, the two efforts hope to come back together in mutually enriching ways.

An Architectural History: Building A Special Collection into A Digital Library

Like most full text collections available on the World Wide Web, MoA is thematically focussed. The building of such online collections has led to a number of high quality and carefully constructed scholarly resources. Where MoA differs from these collections however is in its size and its ability to grow even larger. The mass of the content and the scalability of the system are the key features that will allow MoA to grow out of being a special project into a true digital library. As this growth happens we hope to better understand how we can apply our existing knowledge about libraries to online systems and about the continuities and discontinuities in user behavior as users move from physical to electronic stacks.

MoA hopes to promote a productive and dynamic synergy between a relatively new and developing system architecture for *digital libraries* and an established intellectual architecture for *libraries*. The system architecture of MoA has three guiding principles. First, we are intent on developing a system that can accommodate further treatment of the materials (such as full encoding) and better access processes, as they become available. Second, there is a strong commitment to a system that can scale as we include more materials. We hope to add thousands of volumes to the system in the next few years and the access and delivery mechanisms must be able to grow with the content. Finally, we are committed to a high degree of usability. We facilitate behavior that is consistent with the existing use of paper journals and monographs, aiming to make the digital copies an acceptable surrogate for the print counterpart.

Our understanding of how people use these texts – and our vision of the whole MoA system – comes, to a high degree, out of the professional and intellectual experience of librarianship. MoA was built by library staff: the bulk of the programming, the system design, the management of OCR processes, and the interface development were all carried out by professional librarians working with library staff in the more traditional areas of collection development, preservation, and cataloging. We have aimed to bring together technical expertise with our training in and understanding of information organization, collection development, and research needs. Moreover, we have drawn upon existing library standards for comprehensive and varied collections and commitment to broad public service. By doing so, we hope to make MoA a model digital library that reflects the best principles of the traditional library.

Not only do we aim to make MoA a digital library in *conception*, but we also look for it to work as a library in *use*. By seeding MoA with substantial content, by pursuing opportunities for growth, and by facilitating a number of access strategies, we seek to simulate the experience of research in the library stacks. Users can conduct finely targeted searches – they can also browse and play.

They can search on subject headings. They can work closely with individual volumes, or they can get a sense of the coverage within a period on a particular subject. There is enough material that anyone interested in the period can return repeatedly for different purposes and find useful texts. To a large degree we have tried to support such variety of use by not over-anticipating types of use. We aimed to build as large a collection as was feasible given our resources; we thought about research behavior and how to facilitate it, without being overly deterministic. Actual use of MoA has shown us that users are more creative and ingenious than we could ever have imagined. Like a physical academic library, we have had to work on maintaining a sense of balance: between dedication to defined primary users (e.g. "research faculty") with serving the general public (e.g. "citizens of the state of Michigan") and between anticipating user needs and avoiding limiting the possibilities for use of the material.

Technical Notes: Methods for Creating a Scalable and Sustainable Online System:

The above section details some of the theory that went into the design of MoA. The success of the implementation also depends, of course, on a set of practices that make that theory realizable. What follows are brief descriptions of some of the methods that make MoA work:

- **Selection**

At Michigan, collection specialist librarians gave the Systems Office a set of criteria (date range, US imprint, materials that had been removed to remote storage) and asked the office to generate lists of titles that met those criteria. The subject specialists then went through the lists, marking likely looking titles. Hourly staff moved the selected volumes from remote storage to the specialist's office for review and final selection. Special collections had the option of reviewing all volumes before the texts were sent on into the conversion process.

The thematic focus of the initial phase--antebellum period through reconstruction, 1850-1877--was chosen for several reasons. First, scholarly and general interest in this period of American history remains high, thus increasing the potential of the collection to support the research and teaching needs of the partner institutions. Second, much of the literature of this period is deteriorating rapidly and to preserve it the materials must be reformatted – the materials were already high priority preservation candidates. Third, the body of literature is of a manageable size, so a cohesive collection in digital form could be assembled quickly. Finally, the publications from this period are not covered by copyright protection and thus can be made freely available to the public.

As MoA continues to grow, the developers hope to move toward a more automatic method for identifying materials for conversion. By establishing and using broad selection criteria (for example, *all* volumes within a certain date and imprint range that are currently housed in remote storage) rather than making volume by volume decisions, we hope to considerably reduce the time and effort of selection.

- **Conversion**

The materials in the MOA collection are scanned from the original paper source. The volumes were disbound – the texts were often so brittle and damaged that this had to take place locally rather than risking destruction in shipping. (After scanning all pages were returned and are now stored in preservation boxes in the remote storage facility, pending a final decision about their disposition). A service vendor did the scanning itself. The images are captured at 600 dpi in TIFF image format and compressed using CCITT Group 4. Minimal document structuring occurs at the point of conversion, primarily linking image numbers to pagination and tagging self-referencing portions of the text, such as indices and tables of contents. In the case of serials, low-level indexing was added post-conversion by the partner institutions; Cornell and Michigan staff collaborated to determine low-level indexing guidelines for this complex group of serial titles (Shaw and Blumson, 1997).

The MoA images also went through a quality control process when they came back from the service vendor. Making this process efficient proved to be an important lesson of the project. Originally, the images were loaded into our system and processed, including the OCR, as soon as they arrived from the vendor. The Preservation Department then checked each image for acceptable clarity and degree of skew. If images were rejected by Preservation they were returned to the vendor, and we waited for them to come back through our process again. This proved to cause extensive duplication of effort, to consume a great deal of time and to not allow communication with the vendor in a timely enough fashion to correct systemic problems. In the future any substantial additions to MoA will pass through quality control *first*. Preservation will review a statistically determined sample of the images and either reject or accept batches of images. When the images have passed through quality control, they will then only have to be loaded and processed once.

- **Access**

Once the page images were returned to the U of M, DLPS conducted fully automated OCR to generate the searchable text for the system. Additional processing cleaned up some basic problems with the OCR (non-ASCII characters) and converted bibliographic metadata into a TEI conformant header.

A commitment to flexibility and extensibility was central to the development of the OCR process. When improved OCR technology is developed, as has happened twice since the inception of the first project, the materials can be treated again at very little cost. Further, high priority items can be moved through a process where the OCR is fully corrected, and the text fully encoded. This is an expensive process but adds functionality (e.g., chapter navigation or display of encoded text in addition to images), and can be applied selectively as resources and demands allow.

Since the completion of the most recent OCR process, the Digital Library Production Service at the U of M has conducted an analysis of the accuracy of the OCR and the confidence ratings assigned by the OCR software. The goal of this project was to develop a method for distinguishing accurate OCR files from OCR files with an unacceptable number of errors, without having to examine each file. This ability will enable the Digital Library Production Service to put

online those OCR files with a high probability of accuracy and to estimate the amount of "clean-up" required to correct pages with an unacceptable number of errors (see Bicknese, 1998).

Another important aspect of the MoA system is the method of page image delivery. As indicated above, the MoA images are stored in TIFF format. TIFF, however, is not a format that is widely understood by World Wide Web browsers. Because of this, page images that are presented to the user are converted to GIF format, which is universally understood. This is facilitated by the use of Tif2gif, a specialized utility which converts TIFF images to GIF images quickly, but with a limited set of scaling options. Tif2gif, written by Doug Orr, was originally developed at the University of Michigan and is used in a variety of our digital collections(Shaw and Blumson, 1997). As John Price-Wilkin has pointed out, there are two important rationales for make the material available through dynamic rather than pre-computed and stored transformations: "First, we assume that the patterns of use in our collections mirror those of traditional libraries, where many of the materials go unused for significant periods of time, and where many resources are used only once in an extended period of time. Consequently, creating derivatives for potential use will result in most derivatives being unused and both computational and human resources being wasted." Second, this method also ensures flexibility and the possibility of forward migration in the system. It makes the quick and widespread delivery of the page images possible in the present and allows for the possibility of better delivery in the future. As better delivery strategies are developed, we can return to the original high quality images and adapt their presentation to these better methods (Price-Wilkin, 1997).

Future Directions for the Making of America

The development phase of Making of America ended in 1997 (periods of assessment and refinement are part of the ongoing work of MoA) and MoA has been in full production and use for over a year. MoA is, however, very much a work in progress, and the developers are committed to adding both to the content and functionality of the system. Our immediate goals include:

- Further integration with the Making of America materials at Cornell University
- Gradual migration from raw OCR to fully corrected and encoded text, based on the availability of resources and specific demands. The Humanities Text Initiative, a part of DLPS at the UM, will undertake the process of proofing OCR and refining markup based on user demand. The HTI, as part of its American Verse Project is currently in the process of encoding over 200 volumes of poetry from the MoA collection.
- The UM Library will be incorporating digital conversion into its Preservation Department's "Brittle Books" program. New materials will be added to the MoA site as they are converted.
- We are working with other institutions and funding agencies to make more significant additions to the MoA site.

One substantial measure of our success in building MoA will be whether we have indeed designed a system that has room to grow and that can accommodate and facilitate these goals.

MoA is a complex system. Designing and building a complex system is a complex process. As these brief stories may illustrate, it takes significant efforts to find funding, to building coalitions between and within libraries, to define standards, to undertake the conversion, and to specify and implement the system. MoA also requires an ongoing commitment to user support, to periodic assessment, and to maintenance and upgrading. Those are a lot of trees to keep track of. But we still need to keep looking at the forest as well. It is valuable for both the system developers and those who would learn from our experience to periodically remind ourselves of what we are trying to do. At the U of M Digital Library Production Service in general and in MoA in particular we are trying to build sustainable and usable systems for putting library content online – systems that can grow with growth in our expertise and technological capability and can grow with users as they become more comfortable and experienced with using online texts. In the Making of America, we are designing a system to help preserve an intellectual history that is in danger of being lost and to make accessible that history in powerful ways.

Bibliography

Bicknese, Douglas A. Measuring the Accuracy of the OCR in the Making of America (1998). URL http://quod.lib.umich.edu/m/moagrp/moaocr.html

Guthrie, Kevin M. and Lougee, Wendy P. (1997). The JSTOR Solution: Accessing and Preserving the Past. *Library Journal*, 122(2) (February 1, 1997), 42-44.

Lougee, Wendy. The University of Michigan Digital Library Program: A Retrospective on Collaboration within the Academy (1998). *Library HiTech*, 16(1), 51-59.

Price-Wilkin, John . Just-in-time Conversion, Just-in-case Collections: Effectively leveraging rich document formats for the WWW (1997). *D-Lib Magazine*, May 1997. URL http://www.dlib.org/dlib/may97/michigan/05pricewilkin.html

Price-Wilkin, John and Bonn, Maria (1998). Making of America IV: The American Voice, 1850-1876. Proposal Submitted to the Andrew W. Mellon Foundation, September 18, 1998,

Shaw, Elizabeth J and Blumson, Sarr. Making of America; Online Searching and Page Presentation at the University of Michigan. *D-Lib Magazine*, July/August1997. URL http://www.dlib.org/dlib/july97/america/07shaw.html