Assessing The Costs of Conversion

MAKING OF AMERICA IV: THE AMERICAN VOICE 1850-1876

A HANDBOOK CREATED FOR
THE ANDREW W. MELLON FOUNDATION

THE UNIVERSITY OF MICHIGAN
DIGITAL LIBRARY SERVICES
JULY 2001

TABLE OF CONTENTS

Introduction

From February 1999 to February 2001, the University of Michigan University Library engaged in a large-scale digitization project entitled "The Making of America IV: The American Voice, 1850-1877," commonly known as MoA4. MoA4 extends and tests the methods used in the original Making of America project (referred to as MoA1[1]), as well as increasing the content of the Making of America by almost 500%. The project was undertaken to answer the question: *what are the costs and methods of using digital technologies for preserving and deploying monographic materials?*

The project was funded by The Andrew W. Mellon Foundation. While the Foundation welcomed the creation of this digital content, its primary interest in supporting the project was in the accompanying documentation of costs and methods. This documentation is collected in this handbook; it is intended to support the Mellon Foundation and the library community in evaluating and developing similar projects. While the costs in the handbook are not fixed, because they will vary according to local conditions and changes in both the labor and service markets, they can serve as reasonable benchmarks against which to judge future project budgets. This handbook can be of assistance in identifying unusually high or unrealistically low projected costs, in assessing methodologies and in assuring attention to the central component parts of the digitization process.

This manual includes an extensive detailing of the component costs of conversion for preservation and access, along with documentation of the methods used by the University of Michigan. The documentation of the process focuses on the methodology of converting monographs to the simplest digital form, scanning pages as preservation-quality image files, and automatically generating OCR and simple SGML. The online system that delivers these digital materials is predicated on this method of conversion, and as such the costs detailed are those that provide the highest level of functionality for a relatively minimal investment.

This handbook does not detail, at the component level, the costs of online implementation. It makes some assumptions about a digital library infrastructure being in place to deploy these converted books – assumptions that are outlined in the course of this discussion. The handbook does describe the activities necessary for putting the converted materials online, but at the present time there is too much variety in local practice at different institutions to successfully benchmark the costs of system building and deployment from the ground up.

Assessing the Costs of Conversion is designed to allow assessment and understanding of methods in a component-by-component approach. An institution might find any part of the process (e.g., quality control) to be more cost-effectively performed by using other methods or services. As well as establishing a framework for such decisions, by identifying parts of the process and methods for performing each task, this report may aid in the establishment of a marketplace, and marketplace standards, for services surrounding the conversion process.

This handbook begins with a description of the Making of America project, including a brief history, an overview of its principles of conversion of materials to digital form, and a description of the existing infrastructure at the University of Michigan University Library that underpins the Making of America. This last is intended to both give readers a context for understanding the

[1] The Making of America 2 and Making of America 3 are non-affiliated projects carried out at other institutions. They are at various stages in their development.

project and to indicate where some economies of scale have already been achieved, economies that affect the costs presented here.

Following this introductory material, the handbook details the conversion methods and process used in MoA4, translates the steps in that process into cost categories and reports on the costs associated with each step. Each section contains an explanation of the manner in which costs are assessed and described. These analyses also look at some ways in which variation in local practice among institutions might affect cost and at the possibility of outsourcing parts of the conversion process. The cost section reports on the differences between the actual costs and what can be achieved under optimal conditions and some suggestions for creating those conditions.

Finally, the handbook closes with a more detailed description of the context in which these materials are put online and made widely accessible. While this section does not take on a component-by-component investigation of building an online system, it does describe some of the elements necessary for providing full access to digitized material. This infrastructure was already in place at the University of Michigan Library. Similar endeavors starting from scratch will need to plan for the development or acquisition of staff and systems to deliver digital library content.

This handbook also contains several appendices: the spreadsheets that detail the costs calculations used in this handbook, a list of resources for planning and budgeting for digital projects, a discussion of calculating costs of OCR, and position descriptions for staff employed by the project.

The University of Michigan Making of America

The University of Michigan Making of America (MoA) is a digital library of books and journals focusing on 19th Century American history, with the majority of resources from the antebellum period through reconstruction. MoA aims to preserve and make accessible through digital technology a significant body of print materials in United States history and 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 project is a collaborative effort within the University Library, involving staff from Collection Development, Preservation, Technical Services, and Digital Library Services. Primary responsibility for the production of the MoA system lies with the Digital Library Production Service (DLPS), a unit within the Digital Library Services division.

MoA was born out of a major collaborative endeavor between the University of Michigan and Cornell University with funding from The Andrew W. Mellon Foundation. This initial effort is known as MoA I. At the conclusion of MoA I, in 1996, the U of M collection contained approximately 1,600 books and 50,000 journal articles, a total of over 630,000 pages[2]. 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 would be most readily applicable to research and teaching needs.

[2] The Cornell Making of American collection, http://moa.cit.cornell.edu/moa/, exists as a separate body of materials, currently totaling 907,750 pages (967 monographs and 955 serial volumes).

MoA4, begun in 1998 with funding from the Mellon Foundation, added almost 8,000 volumes to the MoA collection – over 2,500,000 pages of monographic content. MoA4 converted the vast majority of the 1850-1876 U.S.-imprint, English language materials in the U of M Buhr remote shelving facility (these volumes had been removed to remote storage either because of low use or deteriorating physical condition). This content represents a significant percentage of U.S. imprints of this period. The body of material converted through both MoA I and MoA IV, and freely available over the Internet, is so substantial that it has transformed both the perception of the size of collections on the Internet and the ways in which libraries are currently thinking about conversion activities.

The Making of America, both in its initial phase and since the MoA4 extension, 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: in the year 2000, materials previously unused and in storage were searched an average of 120,000 times a month, and users displayed more than 500,000 pages each month. 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. 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. 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.

Further, 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), who have begun to adopt the U of M deployment strategy in their preservation efforts. MoA serves as a model for conversion that accommodates both scanned images with automatically generated OCR and carefully prepared (proofread and fully encoded) materials, where journals can coexist with monographs, and where preservation and access are equally well supported.

In summary, the Making of America represents a powerful marriage of preservation and access, bringing new vitality to preservations efforts and ensuring that the information in these resources is available to vast new audiences. The collection is a boon to the public; the project is a model for digital library practitioners.

Conversion methods and process

Content Selection

Readers familiar with the creation of digital library collections will immediately note that the process outlined above contains no mention of the selection phase of the project. One of the theoretical and practical principles of MoA4 was a careful avoidance of laborious selection. At this time, most full text digital collections are thematically organized; they are subject-specific and have carefully articulated requirements for items for inclusion. While such projects often result in attractive subject integrity and thus in high quality collections, they are also time consuming and costly, requiring careful identification and physical examination of possibly thousands of volumes, and case-by-case decisions about the appropriateness of those volumes for the project. Such collections also depend on a specified notion of an audience with a particular subject interest and research needs.

While MoA1 depended upon manual selection by professional librarians of items for conversion, MoA4 employed a more cost effective selection model in order to establish a more generalizable and automatic method for identifying materials suitable for conversion. By establishing and using broad selection criteria rather than making volume-by-volume decisions, libraries can considerably reduce the amount of time and effort spent on selection. In addition, such criteria help to avoid the often-ineffective process of trying to predict future use with imperfect information about the potential users of the collection. These criteria were that the items be:

- Monographs (including pamphlets)
- U.S. imprints
- English language
- Published between 1850 and 1876
- Currently housed in the remote shelving facility
- Embrittled condition

In this simplified selection effort, volumes were retrieved by clerical or hourly staff and moved automatically into the conversion process. This method relied upon the high quality of existing processes in the Library: it trusted the decisions of collection development in the original acquisition of the volumes and the decisions of selectors in moving items to remote storage (decisions that have already taken into account imprint, circulation and condition). By drawing upon these established processes and reducing the human costs, the automatic selection was a streamlined and cost-effective practice.

Although this was largely a wholesale automatic selection of volumes based on project criteria, the process did build in some provision for review to avoid unwitting destruction of volumes of significant artifactual value. To accommodate this review, the Preservation staff responsible for preparing volumes for disbinding were advised of the criteria for books to be removed from the production stream (volumes containing significant illustration or photographic plates, first editions of notable authors, signed copies, and some volumes in intact original bindings with paper in relatively good condition). These volumes were set aside for review. Collection specialists and staff from Special Collections reviewed these every few weeks and removed a small number of volumes from the production stream. The subject specialists report that this took no more than an hour or so of their time each month; the process had no significant impact on the preparation workflow.

Basic Principles of Conversion

The items in the MoA4 project underwent a simple and automatic conversion process. Each volume was collated page-by-page to ensure that the volume was complete. Missing pages and significant deterioration in the condition (other than expected high levels of embrittlement) were noted on a processing sheet. The processing sheet also contained a unique ID. Each of the volumes was disbound,[3] the processing sheet attached and the volumes were sent to an outside scanning vendor.

[3] Michigan's experience indiciates that the highest quality images is obtained from disbound material. Material that is scanned in bound form often requires a great deal of post-scanning enhancement in order to flatten out the curvature of the volume and to remove gutter shadow. Overhead scanners also provide lower levels of resolution and therefore, lower quality images. All the volumes converted in this project were brittle, and their remaining life span was short. Very few of the volumes to be converted in this project had artifactual value. However, project staff watched for volumes that have special bindings or that have potential artifactual value and referred these titles to the

Pages were scanned as preservation-quality (600dpi) image files, and the Preservation Department undertook a quality control process for the images, using a roughly 5% sample. Images were inspected to ensure completeness, and to assess the legibility and placement of the images. Page image files were then processed to generate OCR and simple SGML that enable search and navigation and that sit "behind" the image files which are still the primary means of access to the content. This automatic generation of the OCR makes it possible for the project to take advantage of advances in OCR technology as they become available: when better recognition becomes possible, the OCR can be automatically regenerated at very little cost.[4]

By using these methods these costs are constrained to those a funding agency could justify in large-scale conversion: expensive and detailed handwork such as proofreading or keyboarding, and careful low-level encoding, are excluded. U of M supports an ongoing and parallel activity in its Humanities Text Initiative (HTI) where (as a result of user demand and through the allocation of appropriate resources) individual titles are more fully processed, with correction of the OCR and fuller encoding, and then re-submitted to the system.

All of these conversion activities took place within the context of an established physical and digital library infrastructure. It is important to keep this in mind when applying the benchmark costs noted in this report. Almost all projects are predicated upon some existing facility, staff, equipment and expertise – while this is an obvious point it must still be actively taken into account in project planning and costing. At the low-tech end, for example, a project like MoA4 required a number of book trucks for retrieval and interim shelving of its many volumes. In the context of the University of Michigan Library remote shelving facility, there were more book trucks than a project twice this size would need, so there was no need to account for them in budgeting. Or consider another example: project staff needed bar code scanners in order to check out the books. At the beginning of the project these were actually at a premium at the storage facility and project staff sometimes had to wait until the end of the day, when the work of the Buhr staff was done, to check out the MoA4 books. When the Preservation department became aware of this bottleneck, they purchased more bar code scanners. Preservation saw this as part of its standard operating costs and thought that the scanners would provide greater efficiency for all its projects, so these costs are not accounted for in the MoA4 costs. Another project might need to budget specifically for such a purchase.

These examples may seem trivial, but it is important to identify the assumptions at work when the cost of a project is projected. The biggest assumption at work in the MoA4 projects is the existence of an established digital library system for storing the material and for making it accessible. This includes servers, considerable disk space, a search engine, staff to deploy the materials and to develop the software for search and retrieval. All of these elements are in place as part of the daily work of the University of Michigan Digital Library Production Service and deploying MoA4 was just one of many DLPS projects. Although this report concludes with an attempt to describe those components, the costs indicated here do not take into account that level of system building.

---

appropriate subject specialist for review. Volumes that could not be disbound (but which are strong enough to withstand the conversion process) are converted in house when necessary.

[4] U of M has performed OCR on the first MOA materials twice, each time improving the quality of the resulting text substantially. The most recent OCR has been statistically determined to be approximately 99.8% accurate for nearly all significant pages (Bicknese).

Some notes on interpreting costs

In order to explain costs and thus make them meaningful in budget calculations for future projects, it is important to understand the steps in the conversion process, the tools used and the human labor involved. This section begins with a discussion of the representation of costs for the MoA4 project and some explanation of points of variation in the costs. It then outlines the steps involved in the conversion process, annotated with fuller descriptions of the methods involved in each step. This section also includes a description of the project workflow so as to illustrate the relationships between the steps. These steps are described within the local context of the University of Michigan University Library. In all likelihood many of the methods can and will be applied at other institutions, but there are some points where there might be considerable variance in local practice.

Each of the steps in the digitization process constitutes a category of cost, and this report endeavors to break down the costs into portions that are meaningfully specific. The categories can obviously be broken down, but this analysis attempts to aggregate at a meaningful level of detail. Activities that are linked are grouped together, and information is generated at a level of specificity that can help planners of similar projects gauge the need for resources.

This report uses the page as the unit for representing the costs. While page length varies widely between volumes, the total number of pages in the project remains constant. The cost of each page carried within it a considerable number of factors – human labor from the project team, the costs of hardware and software, the costs incurred from outside contractors and so forth. In representing costs this report attempts to both break those factors out so that they can be assessed separately if need be and to combine them so that total costs of conversion and implementation can be ascertained.

There are two substantial variables that need to be taken into account in applying these costs to other projects. The first is that there may exist considerable variation in local practice that will have an impact upon costs. For a fairly simple example, consider the ways in which the location and organization of the materials to be converted might affect the costs. Some collections may be co-located and complete. Such a collection, if selected for conversion, could be easily retrieved for preparation. In the case of the MoA4 materials, the diffuse locations throughout the remote shelving facility (the major organizing principle in Buhr is volume size), combined with labeling practices that varied over time, at times slowed retrieval considerably. The project staff was able to retrieve about forty books an hour, totaling about 187 hours of labor on this phase. Since staff involved in this activity averaged about $13.00 an hour, this cost the project a total of about $2,430. Although this is only a small fraction of a cent on a per-page basis spread over a project as large as MoA4, one can see how an accumulation of such variations in local practice could substantially affect costs. Decisions about which project activities to do in-house and which to outsource are more complicated examples of such potential variability. The MoA4 project chose to outsource only the scanning component of the project, trusting its local capabilities for all other tasks. Theoretically, almost all of the components could be out-sourced, most notably creation of pagination structure and feature identification, OCR and SGML creation. Costs would of course then vary according to the vendor market.

The second, and perhaps more obvious, variable is that of differences in the local labor market. The appendix to this report contains position descriptions indicating the levels of staff that UM thought appropriate for the project and the salaries associated with those levels. Costs for a project elsewhere will need to be adjusted, depending on differences in the experience and cost of the staff involved.

As will become obvious, the sum total of the components steps reported here is less than the actual cost of the digitization process. This will be particularly evident in the costs of the steps of the preparation process. The costs of the individual components reflect the amount of time and labor involved in those components when the preparation staff was working consistently and efficiently on that part of the process. Since the staff was constantly multi-tasking, however, and moving from one stage of preparation to another (when the work room got full, books had to be collated and disbound, shipments had to go out on monthly deadlines and work had to be adjusted accordingly, and so on) the efficiencies and economy of scale that might take place from concentrating on one piece of the process until it was completed were impossible to obtain. Not only was that efficiency impossible, it was probably also undesirable as the variety of the prep tasks was important in breaking up often very repetitive and sometimes tedious work and in keeping the staff "fresh." And of course, staff members go on vacation, get sick, need time to recover from particularly big pushes to meet deadlines and otherwise behave in perfectly human ways that interfere with maximum productivity.

Finally it is also worth noting the variation in costs across the duration of the project. Early missteps in tools and methods were corrected and staffing levels adjusted. By the end of the project there was considerable efficiency in both the preparation and the post-processing, resulting in much lower per-page costs in the last months than in the first.

This handbook uses several different techniques to represent some of the variations discussed here. First, it reveals the formulas used for calculating each cost, so that variables can be "plugged in" to calculate costs for other projects. Second, three cost snapshots are reported: real per-page costs for the entire project; per-page costs based on three months of work near the end of the project when routines were established and staff trained; and finally, per-page costs for the month with the highest overall productivity in terms of pages prepped, scanned and OCR-ed. The range of these costs should indicate what it is possible to achieve as well as illustrate some of the investment needed in a period of ramp-up and training

Overview of digitization activities:

The following table provides an overview of the activities in the digitization process, along with the tools and staff necessary for each step:

Activity	**Tools required**	**Staff involved**
Retrieval of volumes from storage		Preservation prep staff
Charging out of volumes		Preservation prep staff
Identification, collation and repair		Preservation prep staff
Disbinding		Preservation prep staff
Removal of covers	Electric paper cutter and blades	Preservation prep staff
Packing and shipping		Preservation prep staff
Scanning and CD burning		Outside vendor
Metadata creation	Imagetag software	Preservation staff w/ trained hourlies
Quality Contol	Tiff viewer, Imagetag	Preservation staff w/ trained hourlies
OCR and SGML generation	OCR software	OCR operator

Details of digitization activities:

The tables below report on productivity levels and costs for each of the steps in the digitization process, as determined by time studies. Each table is then followed by a description of the activity.

Retrieval and charging out of volumes

volumes retrieved per hour	40	
hours for retrieval	188.6	
total cost for retrieval	$ 2,439.95	Calculated w/ average hourly salary for full time prep staff (w/benefits)
cost per-page	**$ 0.0009**	

volumes charged out per hour	40	
hours for charging	188.6	
total cost for charging	$ 2,439.95	Calculated w/ average hourly salary for full time prep staff (w/benefits)
cost per-page	**$ 0.0009**	

Volumes were retrieved using printouts generated from the MARC records that met the criteria for selection for the project. Volumes in remote storage are accessed using a location number. This number was located and confirmed before the physical volume was pulled from the shelves. Because of the age of these volumes, not all location numbers were in the automated catalog, and staff would resort to the shelf list, adding considerably to the time required for retrieval.

Volumes were charged out to the Preservation department. For the most part this is a very simple process using the Circulation computers and bar code scanners at the remote shelving facility. At times, the age of the items and their location in remote storage also slowed the charge-out process. Since some of these materials had not circulated in thirty years, circulation records had never been created in the automated catalog. For these volumes, staff had to create circulation records; some circulation records were also incomplete (no location numbers, bar codes matched to incorrect records) and staff needed to complete or correct the records as they worked.

Identification, collation and repair/disbinding/packing

volumes collated per hour	3	
hours for collation	2514. 67	
total cost for collation	$ 32,532.63	Calculated w/ average hourly salary for full time prep staff (w/benefits)
cost per-page	**$ 0.01**	
covers removed per hour	30	
hours for removing covers	251.47	
total cost for removing covers	$ 3,253.26	Calculated w/ average hourly salary for full time prep staff (w/benefits)
cost per-page	**$ 0.0017**	
packed per hour	21	
hours for packing	359.24	
total cost for packing	$ 4,647.52	Calculated w/ average hourly salary for full time prep staff (w/benefits)

cost per-page	**$ 0.0017**

The preparation of the volumes sent for digitization included collating each volume to note missing pages, followed by disbinding. Disbinding included removing the book covers from the text block and carefully removing the cloth and glue holding the text block together. Finally, the spines of the books were cut with an electric paper cutter. Preparation staff members were trained by conservation staff on the use of this cutting so as to obtain the cleanest, straightest cut possible.

As the books were prepped, a processing sheet was generated for each volume. This sheet was generated from a Microsoft Access database into which the unique identifier for each volume had been preloaded, along with call numbers – the IDs and call numbers had been extracted from the MARC records. This database contained a field into which any special processing notes,

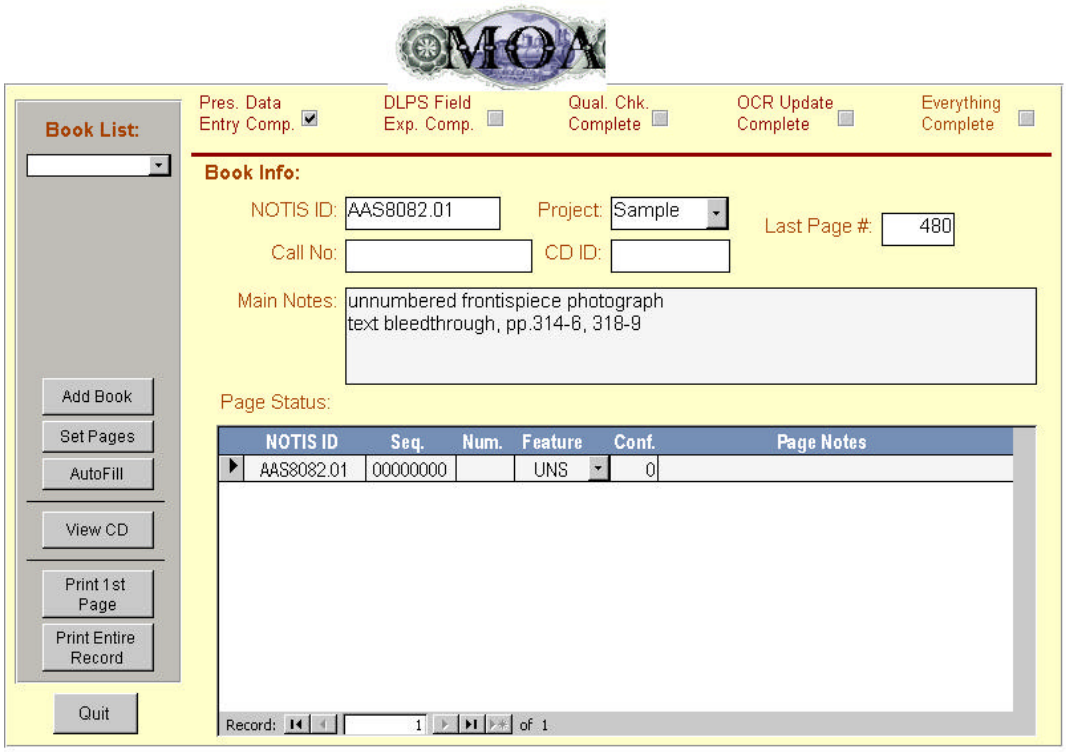


Figure 1: MoA Preparation Database.

such as records of missing pages or of substantial damage, could be entered (*see Figure 1)*. The record for each volume was printed in a report form to serve as the processing sheet. This sheet was tucked into the book as it made its way through the prep process and ultimately was shrink-wrapped with the bound volume prior to shipping. Although generally an automatic process, this could be slowed by multiple volumes with the same ID. For these, the project staff created additional records, using two-digit suffixes on the ID to indicate volume number.

As a last preparation step before shipping, volumes were shrink-wrapped with their processing sheets. In the initial shipment, staff tied the volumes with preservation string (the ubiquitous "pink ties") but the scanning vendor reported some disturbance of contents during shipping leading to stray pages and processing sheets. The books were packed into plastic totes with

snap-down lids and shipped via Federal Express. Because the scanning facilities were located in Mexico, the shipments were routed to a broker, selected by the scanning vendor, in Arizona who saw them through customs and to the scanning facility.

Issues with missing pages

Originally, staff intended to order replacements of missing pages through interlibrary loan as the missing pages were discovered and to hold those books back from the scanning queue until the replacements had been inserted. Early on the prep staff reported that this involved considerable overhead in making and tracking the ILL orders, in storing and tracking the volumes and in replacing the pages. The project team decided to note missing pages during the prep process and defer their replacement until after the volumes were scanned. The list generated by the prep staff was supplemented by missing pages discovered during scanning and reported by the vendor and by missing pages discovered during QC, resulting in a fairly exhaustive inventory of replacements needed. Of the 7,547 volumes digitized for the MoA4 project, about 1.5 % were identified as needing replacements; this is a percentage only indicating *missing* pages, not pages that were so damaged as to obscure the text. (Our current strategy to acquire missing pages calls for replacements to be ordered through ILL, scanned locally and then inserted into both the master directories stored on CD and in the online system. This will involve some down stream costs not represented anywhere in this report, but the costs associated with this strategy are less than generating the replacements upfront.)

Scanning and CD creation

An outside vendor was contracted to perform scanning and burning to CD of the page images. A scanning project of this size generated considerable interest in the vendor community and over forty imaging companies were offered the opportunity to review the RFP and bid on the project. Twenty-eight vendors expressed interest in bidding on the MoA4 scanning, and sample volumes were prepared and sent to these vendors. Fourteen bids were received, ranging from four dollars to ten cents per-page. University Purchasing, in consultation with DLPS, selected three finalists based on cost and quality.

The University of Michigan evaluated vendor proposals using the following guidelines:

- Understanding of and compliance with the requirements of the RFP
- Excellence of response
- Successful processing of the test materials
- Ability to meet technical and managerial requirements
- Ability to meet the required scanning production within the required project timetable
- Careful handling of fragile original materials
- Fair pricing of the proposal relative to other proposals received
- Qualifications and experience, evidenced by customer references, resumes of key personnel, etc.
- Guarantee of work, and the nature and extent of vendor support
- Financial stability and other various business issues

Evaluators assigned a numerical “Range of Compliance” number from 0 to 5 (with 0 as “Not Compliant”) to each. This “Range of Compliance” rating provided proposal evaluators with a

clearer insight into the overall strengths and weaknesses of the bidding vendors. After this evaluation, the contract was awarded to Digital Imaging Technologies (DIT), a Lason Company, as the rate of thirteen cents per page with a slightly higher rate for occasional pages requiring special treatment.

Scanning specifications were detailed in the RFP. Vendors were required to use a flatbed raster scanner to scan all pages as bitonal (one bit) images (one page per image file) with no gray scale or color scanning at a true resolution of 600 dpi. Because of the age of the material, it was anticipated that all pages would be manually placed on the platen and that an automatic document handler could not be used. Images were created in TIFF G4 format; specified metadata was captured in the TIFF header (see appendix for TIFF header elements) and written to gold CD-R. CD and directory names comply to ISO 9669 standards. CDs were then shipped to the Library while the vendor retained the original material pending quality control.

There is a small cost related to scanning that is reflected in the detailing of MoA4 costs that could and should be avoided by future projects. The original RFP for the project did not specify a brand or quality of CD-R to be used by the vendor. After almost 200 CDs had arrived at the University of Michigan, project staff began to experience difficulties loading and reading some of the CDs. Moreover, some of those that had loaded successfully, when tested a second time, failed to load. These problems were traced to a single brand of CD (the vendor had used various brands). When UM communicated this problem to the vendor staff, they did their own testing and experienced similar problems. UM requested re-mastered CDs (and all future CDs) on gold CD-R.[5] The vendor agreed to do this at the cost of materials and labor -- $15.00 a CD, for a total of $2,490. Although this is a relatively small cost spread over the project, it is still one that can be avoided by more stringent up-front specification.

Quality control and metadata creation

Creation of page level metadata to aid in navigation and quality control of the images were done simultaneously, although they were two distinct processes. The original intention of the project was to capture structural metadata in concert with collation. This was to involve some minimal pre-conversion activity: notation of pagination structure and identification of special features, such as title pages, tables of contents, and indexes. Early in the project, this caused the preparation staff to fall drastically behind in preparation of the shipments. Working with the physical volumes at the same time as entering data into a Microsoft Access database proved physically awkward and slow. Since maintaining the vendor's per page scanning costs was dependent on providing a steady and predictable volume of work, the process needed to be rethought. It appeared that there would not be a substantial difference in the information captured if the metadata creation phase was moved to the end of the process by working from the scanned images of the pages rather than the physical volumes. Like capturing metadata at the point of collation, this had the advantage of combining with another process (in this case QC of the images) and even added some checks to the QC process (see below). In addition, it proved more physically efficient to be working on two computer-focused activities at the same time than to be moving back and forth from the print volumes to the computer.

The metadata process itself was a fairly simple task. Page numbers were matched up to sequence numbers in the image directory. That is, the first image is not necessarily paginated as Page 1. It may be blank, or a title page, or page I, or, in the world of 19th century printing at

[5] Although there is not yet a NISO projected life span for gold CD, Kodak claims that its laboratory testing guarantees a life of 200 years. This would mean complete refreshing every 100 years and testing every 10. NISO planned to publish its standard sometime in 2000, but it does not yet seem to have appeared.

least, many other things. In order for the user to navigate the text in meaningful ways and to make use of tables of contents or indices, the proper pagination must be identified. In the case of a book with pages numbered straight through from 1 to 350, for example, this process can be done very quickly. In the case of a large book with plates and multiple pagination schemes (roman numerals, Arabic numerals, multiple parts), this process can be time-consuming and painstaking work. Originally this work was done in a Microsoft Access database with some special configuration done to enable auto-filling between page numbers. As mentioned above, the database and the data entry process were slow and mired down the preparation process. After the decision was made to capture this information post-scanning, Cornell University shared a software tool called ImageTag that enabled pagination and feature identification in tandem with viewing the pages images in one application *(see Figure 2).*

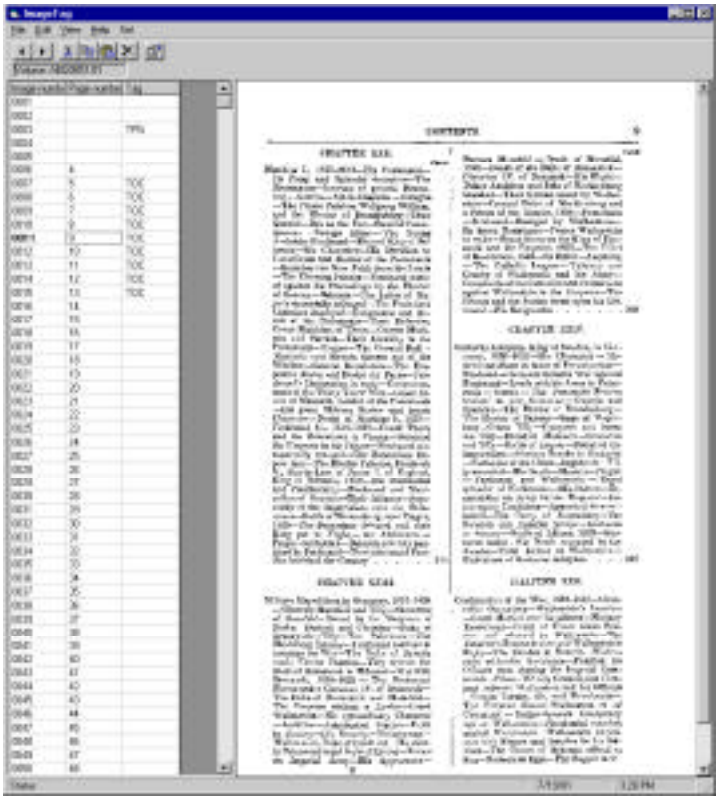


Figure 2. Pagination and feature identifcation viewed in tandem with page image via ImageTag.

ImageTag is essentially an interface to an Access database, so the information is still stored in Access, but the process of entering the data is considerably speeded. In addition, since the tool automatically displays the page image that corresponds to where the user is in the page sequence, opportunities for human error are considerably reduced. Pages containing special features are identified at the same time as the pagination structure. In the case of MoA4, these were the features deemed useful in basic navigation of the text (title pages, tables of contents, lists of illustrations, indices), but this would vary depending on project needs and the ImageTag menus can be rebuilt to enable customized choices of feature identification.

Quality control of the images was conducted simultaneously with the capture of the structural metadata. Criteria for QC of the page images were laid out in the RFP[6]:

> For images consisting of text/line art, all of the following requirements must be exhibited when examining a 600 dpi paper printout without magnification:
>
> - full reproduction of the page, with skew under 2% from the original (100% of all pages);
> - sufficient contrast between text and background and uniform density across the image, consonant with original pages (100% of all pages);
> - text legibility, including the smallest significant characters (100% of all pages);
> - absence of darkened borders at page edges (98% of all pages);
> - characters reproduced at the same size as the original, and individual line widths (thick, medium, and thin) rendered faithfully (98% of all pages);
> - absence of wavy or distorted text (98% of all pages).
> - Magnification may be used to examine the edges and other defining characteristics of individual letters/illustrations. Under magnification the following text attributes are required for 98% of all pages:
>   - serifs and fine detail should be rendered faithfully;
>   - individual letters should be clear and distinct;
>   - adjacent letters should be separated;
>   - open regions of characters should not be filled in.

Quality control methods specified in the RFP asserted that staff would perform quality control by random sampling of each CD (usually a 5% sample). Technical targets and a sample of the digital images would be viewed on-screen at full resolution using a high-resolution monitor. Staff members were to identify missing/incomplete pages, pages out of sequence, and pages skewed, and evaluate the image quality of text and illustrations.

Sampling issues in QC

The sampling method was chosen in accordance with emerging practice[7] and as the most cost effective way of gauging quality across a batch of images. By its very nature, sampling is not

---

[6] These criteria were drawn form the RLG Model Request for Proposals for Digital Imaging Services, available at http://www.rlg.org/preserv/RLGModelRFP.pdf

[7] See for example, the Library of Congress statement on Quality Review (http://lcweb2.loc.gov/ammem/prpsal/rfp9618e.html):

> E.4 LIBRARY QUALITY REVIEW OF DIGITAL IMAGES
> Images, file names, and directory names shall be inspected and evaluated by the Library. The Library's inspection and evaluation shall be in accordance with sampling procedures described in the American National Standard, general inspection level II (ANSI/ASQC Z1.4-1993 and ANSI/ASQC S2-1995). The sampling lot size will be determined by the task order production rate, and will be in accordance with the ANSI standard.
>
> Images - These shall be evaluated and inspected in accordance with the standards indicated in E.4 above, with random sampling lot sizes determined by the task order production rate, except for those specifications requiring 100% accuracy as indicated. A batch will be rejected if, in a given random sample lot size, more than one digital image per 200 sample images is found to be missing, duplicated, illegible, or otherwise defective. This constitutes a 99.5% accuracy rate. In this event, the entire batch will be returned to the contractor for corrections.

The National Archives and Records Administration reviews 10 images per directory or 10% whichever is more and rejects a batch if more than 1% is defective in some way. Less than that, they reject individual images but not batches. See http://www.nara.gov/nara/vision/eap/digguide.pdf

100% reliable, and in particular it is imperfect at turning up missing or duplicated images. (Duplicated images, while not inherently problematic since no content is lost, can still cause significant problems with retrieval. Every word of the OCR text of those images will appear twice in the online system resulting in a false density of search matches, and may cause confusion to the user who is paging through a volume and encounters a duplicate page. The user may even think the browser is "stuck" on an image.) Therefore, sampling may also result in some downstream costs that cannot be accounted for in the model detailed in this handbook. Problems discovered when the material is already online will need to be remedied – missing pages located and scanned, badly scanned ones replaced, duplicate images removed and sequential file names replaced.

In practice, inspection of the images ranged from five to nearly one hundred percent. Early on, staff inspected a very high number of images to look for problematic patterns. It quickly became apparent that image quality was consistently very high, with only occasional problems with skew and blurring. The majority of problems consisted of either missing pages or pages scanned twice, with some cases in which pages were scanned out of order. Once this was determined it became less of a priority to inspect the quality of individual images and more of a priority to determine if all the pages were present and in the right order. This was easily diagnosed with the ImageTag software because it reveals mismatches between page numbers and page sequence. Verifying correct pagination also had the effect of insuring a visual check on several images in each directory. If this check indicated problems with image quality, those directories could be singled out for further inspection.

Perhaps the most important aspect of the MoA4 quality control process was the early and consistent rejection of problematic images and directories. The project agreement with the vendor specified that as soon as an error was encountered on a CD, that CD would be "returned" to the vendor with the expectation that all subsequent errors would be identified and corrected. In practice, there was no need to return the CDs since the vendor had the master files for all the directories; spreadsheets were exchanged almost daily that indicated problem CDs and the nature of the problem and the vendor's comments on the problem (such as identifying a missing page as truly not present in a volume rather than not scanned). In the early months of the project this complete correction of errors did not always take place. That is, the reported error would be corrected but not subsequent errors. Michigan staff were doggedly persistent in their rejection of problematic images. A single CD was rejected as many as four times, each time resulting in a re-mastered CD created at DI's expense. Over time, this consistently high standard paid off. By the middle of the project CDs only needed to be rejected once; by the end there were very few errors, indicating better QC checks before the CDs left the imaging facility.

Although it was not incumbent upon Michigan to report the nature of the error (merely that there was one), it became useful to indicate the problems. In some cases this led us to a greater understanding of Michigan's materials. Incomplete books that had slipped through the prep process were often caught in QC. Missing pages were sometimes misidentified as a vendor error, and the DI staff would check against the original. By the end of the project, the DI staff had become quite adept at identifying such problems during scanning and notifying project staff on delivery of the images.

In addition to image quality control, at the outset of the project incoming CDs were also checked for completeness and correctness of the directory structuring and file naming. In practice, once a few CDs had arrived and these structures had been checked, there was little need to continue a manual check. Several steps in processing, such as OCR, use of ImageTag and data loading "broke" if directories were incorrectly named, thus building an automatic check into the process.

As another quality and completeness check, the contents of the checkmd5.fil were also sampled and inspected to ensure that data were entered as specified in the proposal. MD5, or Message

Digest 5, is a mathematical technique that distills the information contained in a file into a single large number, such that given an input file and its corresponding message digest, it should be computationally infeasible to find another file with the same message digest value. MD5 is a commonly used mechanism to ensure authenticity. For each MoA4 volume prepared by the scanning vendor, a checkmd5.fil was prepared, containing the following information: 1) each page image filename, as stored on the CD-R; and 2) an MD5 checksum value calculated on the given page image file. As the final step of data loading, once page image files had been transferred to the production server(s), DLPS load routines generated MD5 checksum values on the files and checked them against the vendor-supplied checksums in checkmd5.fil. If the values matched, then the files were determined to be identical and the load process had succeeded. If there was (even one) mismatch, the load routine assumed an error occurred in transmission, and signaled the data loader to reload the CD-R. Repeated failures indicated a problem with the source image or CD-R encoding, and were cause for follow-up with the scanning vendor.

Preservation staff members were responsible for both the QC and the metadata creation process. In practice, this meant repurposing the staff from the preparation process. Because of the length of the project and the quantity of the material prep and metadata creation /QC were happening in parallel (CDs flowing in as books continued to flow out) and that staff were reassigned depending on the demands of the moment. This made it useful to supplement the full time staff with trained hourly workers who were able to maintain the QC workflow when prep staff needed to work intensely on preparing shipments for the vendor. Toward the end of the project the prep staff was devoted almost entirely (apart from a few clean up tasks) to metadata creation and QC.

OCR and SGML generation

Number of pages per year	3,450,496
Annual cost	$ 123,105.33
Cost per-page	$ 0.04

(Note: The OCR and SGML generation took place in the context of an established OCR operation that was processing materials other than the MoA4 materials. Thus the costs are based on a higher capacity (and therefore lower per pages costs) than is used to calculate projects costs elsewhere. The appendices to this report contain a fuller explanation on assessing reliable OCR costs and the conditions necessary to sustain low rates.)

Finally, automated Optical Character Recognition (OCR) was generated from the pages images, and the OCR associated with the appropriate page image using a simple form of SGML. The OCR software technology currently used by DLPS is a voting system, providing the MOA system with a significantly higher quality of cCR) was 0.7pthe eQ2e eypted O75 sewherthe MOA

[illegible]RET BT 10.125 228.uptra-R encinclud pre75 Tcologcta creatitain a f CD-R encorkun

Costs of conversion: Real and Possible

In order to understand both the actual costs involved in the MoA4 project and to understand what costs might be possible to achieve in a similar endeavor, the table below represents costs in four different ways (moving from the left hand column to the right):

Column 1: The actual per-page costs as measured over the life of the project and as calculated in Spreadsheet I, Appendix I.

Column 2: The per-page costs in a hypothetical most productive month, Spreadsheet II, Appendix I. This is a hypothetical month since the highest production levels for preparation and OCR (the two most variable activities) did not coincide; the OCR had to wait to kick into high gear until September when the scans came back from the most productive preparation month in July. This "month" is therefore constructed out of July and September on the hypothesis that with sufficient production levels sustained, these costs would be possible to consistently achieve.

Column 3: A three-month average cost calculated over June, July and August of 2000, one year into the production phase of the project, Spreadsheet II, Appendix I. These costs reflect the production levels and efficiencies that are possible to achieve and sustain with a complete and fully trained staff with routines firmly in place.

Column 4: The sum of the component costs discussed in the preceding section, Spreadsheet III, Appendix I. As discussed in the notes on interpreting costs, the sum total of the component steps is less than the actual cost of the digitization process. It is interesting to note, however, that despite the time and money lost in moving from one activity to another, as well as the normal distractions of the work day and human lives, the sum of the component costs is indeed the same as that achieved during the project's hypothetical best month. This would indicate that given a sufficient flow of materials and trained staff, these inefficiencies are negligible to overall costs and can be minimized.

Per-page costs

	total project	**most productive month**	**3 month average**	**if measured by component activity**
prep	$ 0.05	$ 0.03	$ 0.04	$ 0.02
shipping	$ 0.01	$ 0.01	$ 0.01	$ 0.01
qc and metada creation per-page costs	$ 0.01	$ 0.01	$ 0.01	$ 0.006
OCR and sgml generation	$ 0.04	$ 0.02	$ 0.02	$ 0.04
scanning	$ 0.13	$ 0.13	$ 0.13	$ 0.13
process management	$ 0.01	$ 0.01	$ 0.01	$ 0.01

total	$ 0.25	$ 0.21	$ 0.22	$ 0.21

Side by side comparison of fours costs: total for project, three month average one year into project, best month, and sum of component

Even though the costs reported above and reflected in greater detail in the attached spreadsheets represent a careful inventory of the tasks and activities involved in the production of MoA4, there are, inevitably, some costs not directly accounted for that should be considered in project planning. For example, the preparation of the RFP for scanning vendors consumed several days of the project manager's time near the beginning of the project. While a significant effort, this is a one-time cost that is insignificant over the duration of the project. Furthermore, future scanning projects can capitalize on this investment, modifying sections as appropriate, without having to start from scratch.

Another example of a cost that may appear hidden in this report is the cost of managing the books upon their return from the vendor. In the case of MoA4, shipments have trickled back in over many months and been dealt with on an ad-hoc basis by available Preservation staff. Currently routines are being established for a more standardized processing of these books and for their removal to storage pending disposition decisions. These are only two examples of costs not captured in this report. Astute readers and those who have been involved with similar projects will no doubt discover others; it is our hope that they will share that information with the digital library community and thus augment the findings reported here.

A technical infrastructure to deliver digital library texts: Issues of equipment, implementation and long-term viability

This handbook is predicated upon the assumption that the costs of conversion of book and journal materials from paper form have largely stabilized and can be tracked and studied. The companion costs of implementation in an online system are much less stable and still vary wildly depending on local contexts and rapidly changing hardware and software development. For example, the computer and RAID (redundant arrays of independent disks) awarded to DLPS three years ago (as part of A Sun Academic Equipment Grant) that store and serve the MoA materials was estimated at a value of $250,000. DLPS is currently setting up a mirror environment (see below) with more storage for about $37,000. In such a rapidly evolving and shifting economy of digital technology, it is difficult to establish even benchmark costs.

That said, it is still apparent and important that readers of this handbook will want to understand the context in which these materials are put online. This includes the human labor and technical processes involved, the computer equipment that stores and delivers the materials and the methods by which DLPS assures the long-term viability of the digital materials. What follows is an attempt to describe those elements. It should be read with the proviso that it is a description of routines in place in an organization tuned to this type of production. Any similar conversion project will need to design or adapt their own local processes to make the digital materials accessible.

Data Loading

Once the textual materials have been converted to bitonal pages, it is integral to the access component of the project that the images be loaded from the CDs to the production servers. In MoA4, following the OCR process, CDs were handed off to the data loading staff. CDs were

loaded using a Sun Sparc 20 with five CD drives, and a single CD took from thirty minutes to two hours to load. At the beginning of the project, CDs were loaded serially, one at a time. It quickly became apparent that this would mean a considerable amount of time until all the data was online. More CD drives were purchased and load routines were modified so that five CDs could be loaded in parallel.

A programmer was required to set up two shell scripts to manage the load process. The first initiated the loading and reported errors back to the data loader. These errors generally indicated a problem with the media (thumbprints on the CDs were quite common) or with duplicate

mechanisms for page image-based books with associated OCR (such as those in the MoA4 project), and for SGML/XML encoded books and journals is available as part of the free Open Source software. More information on DLXS can be found at http://www.dlxs.org.

Hardware and Attention to Issues of Long Term Viability

In 1999, The University of Michigan Digital Library Services received a Sun Academic Equipment Grant to host publicly available digital collections in the humanities. This grant provided the University Library with an Enterprise 3500 server and 127.4 gigabytes of Sun StorEdge. The MoA4 materials are housed on this hardware. DLPS has also planned and will soon implement a mirrored instance of its collections. This will entail the use of several servers in a two geographically separate environments. Until recently, DLPS's only production environment was a staffed, secure, climate-controlled machine room approximately two miles from the location of DLPS staff. DLPS recently added a second machine room that is similar to the first in several ways, but is not staffed, and which is in the same building as DLPS staff (and thus two miles away from its primary machine room). The two sites will likely use Oracle-specific replication mechanisms for many services like authentication and authorization and data replication services for collections, and thus contain a mirror of each other. This will provide both better performance, by load balancing, and a fail-over mechanism to ensure constant availability of collections.

The image storage and access strategy for the converted texts in MoA4 privileges the TIFF G4 master image by storing it on-line in the access system. The access system itself depends upon the presence of the master image file to deliver information to the user. The master image resides always in the most current technologies and moves forward through technologies in the University's dynamic computing environment. This avoids problems of earlier digital efforts where the critical version of the page image was resident in an off-line medium requiring continuing refreshing. The page images are stored in RAID level 5.

DLPS maintains multiple copies of digital masters in a variety of parallel environments, a practice that further contributes to the long-term viability of the digital resource. The data stored using RAID technology are written to digital linear tape (DLT) on an at least a monthly basis; in the case of SGML/XML data and bitonal image files, the data on tape are indistinguishable from the master. SGML/XML text files and bitonal image files are stored redundantly, with the production version in the staffed machine room, and a secondary version in the local (but not staffed) machine room.

Appendix 1: Cost Calculation Spreadsheets

Spreadsheet I: Real Total Costs Per Page

Real Total Costs Per Page					
		Jun-00	Jul-00	Aug-00	3 month average
real prep costs					
number of months worked on prep	58.5				
average monthly salary	$ 2,156.19				
supplemental temps	$ 8,724.00	$ 3,130.00	$ 2,208.00	$ 801.00	$ 6,139.00
full time staff	$ 126,137.26	$ 6,468.58	$ 6,468.58	$ 6,468.58	$ 19,405.73
total prep staff	$ 134,861.26	$ 9,598.58	$ 8,676.58	$ 7,269.58	$ 25,544.73
prep per page costs	$ 0.06	$ 0.04	$ 0.03	$ 0.04	$ 0.04
shipping costs per page	$ 0.01	$ 0.01	$ 0.01	$ 0.01	$ 0.01
real qc staff costs					
number of months worked on QC	8				
full time staff	$ 17,249.54				
supplemental temps	$ 12,756.77				
total staff	$ 30,006.31				
hardware/software	$ 9,820.00				
total qc costs	$ 39,826.31				
per page costs	$ 0.02	$ 0.01	$ 0.01	$ 0.01	$ 0.01
OCR and sgml generation costs per page	$ 0.04	$ 0.02	$ 0.03	$ 0.02	$ 0.02
scanning costs per page	$ 0.13	$ 0.13	$ 0.13	$ 0.13	$ 0.13
process management costs per page	$ 0.01	$ 0.01	$ 0.01	$ 0.01	$ 0.01
total real costs per page	$ 0.27	$ 0.22	$ 0.22	$ 0.22	$ 0.22

using established rate.
See appendix.

Spreadsheet II: Per-Page Costs

1 year into project

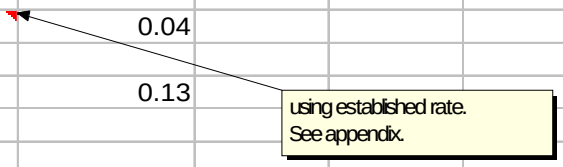
Per page costs	total project	3 month avg*	most productive month	if measured by component activity
prep	0.06	0.04	0.03	0.02
shipping	0.01	0.01	0.01	0.01
qc and metada creation per page costs	0.02	0.01	0.01	0.01
OCR and sgml generation	0.04	0.02	0.02	0.04
scanning	0.13	0.13	0.13	0.13
process management	0.01	0.01	0.01	0.01
Total	0.27	0.22	0.21	0.22

using established rate.
See appendix.

Spreadsheet III: Component Costs

MOA4 Costs by component activity				
	based on time studies			
Prep costs				
volumes retrieved per hour	40			
hours for retrieval	188.6			
total cost for retrieval	2439.94733			
cost per page	0.001044283			
volumes charged out per hour	40			
hours for charging	188.6			
total cost for charging	2439.94733			
cost per page	0.001044283			
volumes identified, collated and repaired per hour	3			
hours for collation	2514.666667			
total cost for collation	32532.63106			
cost per page	0.013923773			
volumes disbound per hour	40			
hours for disbinding	188.6			
total cost for disbinding	2439.94733			
cost per page	0.001044283			
covers removed per hour	30			
hours for removing covers	251.4666667			
total cost for removing covers	3253.263106			
cost per page	0.00198911			
total prep pre packing	0.019045732			
	43105.73616			
Packing				
packed per hour	21			
hours for packing	359.2380952			
total cost for packing	4647.518723			
cost per page	0.00198911			
total prep after packing	47753.25488			
cost per page after packing	0.02043811			
Shipping				
supplies	1875.85			
air freight	16749.6			
total shipping cost	18625.45			
shipping cost per page	0.007971582			

QC and metadate creation				
pages per hour	5000			
hours for qc	467.2962			
Hardware				
computer	2069			
monitors	5000			
total hardware	7069			
Software				
imagetag dev.	2000			
imaging professional	751			
Total software	2751			
Staff cost	6045.483114			
total cost for qc	15865.48311			
cost per page	0.006790333			
OCR and sgml generation costs per page	0.04			
per page scanning costs	0.13			
process management				
project managers salary w/ benefits	29490.92			
student hourly assistant	2751.9			
process management costs per page	0.013799736			
total per page costs based on component approach	0.21899976			
total pages in project	2336481			



Appendix 2: The Variability of Optical Character Recognition (OCR) Costs

The cost of capturing character-level information from page images is variable, depending on tools and methods. It is important to note that in some cases, only manual "keyboarding" can provide reasonably adequate capture and retrieval, which then of course creates a much higher cost for performing retrieval. But even when OCR is an option, there are several variables that, taken into account, can create widely divergent costs for different projects. In his chapter on "Enhancing Access to Digital Image Collections," Wilkin treats many of these issues; of particular value to this discussion is the section "OCR, Corrected OCR, and Keyboarding: Costs" (pp. 110-116), and the sidebar by Kenn Dahl (pp. 111-112).[8]

Foremost among the variables in OCR cost is the choice of an OCR software package. The OCR software used at the University of Michigan in the Making of America IV project is PrimeOCR from PrimeRecognition. PrimeOCR was chosen for production operations at the University of Michigan Library because it provides demonstrably more accurate output in the OCR process. It is, however, many times more expensive than an off-the-shelf OCR package. TextBridge, for example, can be licensed for a few hundred dollars per machine, while PrimeOCR may cost more than ten thousand dollars for a multi-machine implementation like that at the University of Michigan.

Another variable is staffing. In earlier work on OCR, the University of Michigan Library's Digital Library Production Service (DLPS) was able to perform OCR on page images with a very small fraction of a programmer's time, who ran the OCR software package ScanWorx (from Xerox Information Systems), the UNIX "sister" package to TextBridge, as part of his system administration responsibilities. DLPS currently employs a full-time OCR operator who manages OCR jobs, ensuring effective hand-off of output and a level of quality control impossible in the earlier staffing configuration.

A third and final significant variable is hardware. In its earlier work on OCR, the University of Michigan Library "borrowed" cycles from a multi-processor server used for access to online resources, thus diminishing the overall performance of public access, but effectively avoiding hardware costs. A short-term project with TextBridge might, for example, use low-end workstation-class computers costing less than $1,000 each. The current production environment at Michigan is designed to ensure a high degree of uptime and reliability; consequently, our five licenses for PrimeOCR are all installed on multiprocessor server-class Intel computers, each costing as much as $3,000.

The unit cost of OCR is a balance of factors such as those described above. In 2000-2001, with the volume of OCR performed by the DLPS, the unit cost was less than $0.03 per page. The higher cost of software like PrimeOCR and more expensive computers like those chosen by DLPS, along with the higher staffing costs, were spread over more than four million pages of OCR. Lower demand (and thus lower throughput) would raise the cost of performing OCR. A small project would be well advised to outsource or to choose inexpensive software and to use a fraction of an existing staff member, while a project that performed more OCR might wish to pay the slightly higher cost to achieve markedly higher accuracy and sustain higher throughput. Each project will need to balance a number of variables in making implementation choices.

---

[8] Kenney, Anne R., and Oya Y. Rieger, editors and principal authors. *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. Mountain View, CA: Research Libraries Group, 2000.

Appendix 3: TIFF Header Specifications

TIFF Specifications

Michigan requires that the following information be recorded in the TIFF header:

Date and Time of Scan
Source Statement to read "University of Michigan University Library"
Image Description

The data in the Image Description tag should be structured as follows:

CD-R number [assigned by the vendor – must include project identifier]
NotisID
Image sequence number padded to eight digits

In addition, Michigan would prefer that generated TIFF files not make assumptions about default values of TIFF header fields that could affect the ability of TIFF readers to properly display the file. Ideally, each TIFF file should explicitly state the values of the following header fields (inability to provide this information will not necessarily result in rejection of the proposal):

ImageWidth
ImageLength
X Resolution
X Position
Y Resolution
Y Position
Resolution Unit
BitsPerSample
Compression
Orientation

Appendix 4: Preparation and QC Staff Descriptions

Classification
Information Resources Associate I

Duties
repare material for digital imaging including page by page collation; order copies of missing pages through Interlibrary Loan; input information about the volumes into the Project's database; pack material for shipment to commercial vendor; inspect images for quality control; assist with disposition of original materials; related duties assigned; reports to the Electronic Imaging Technician.

Desired Qualifications
Library experience; experience with databases or spreadsheets; knowledge about preservation or digital imaging.

Minimum qualifications
High school diploma and at least one year of office experience; demonstrated interpersonal and communication skills; demonstrated ability to perform detailed work quickly.

Salary range
$15,210 – 35,490

Appendix 5: Resources for Project Planning and Costing

AHDS Guides to Good Practice
Digitising History : A Guide to Creating Digital Resources from Historical Documents
http://hds.essex.ac.uk/g2gp/digitising_history/index.asp

NDLP Project Planning Checklist
Library of Congress, National Digital Library Program
http://lcweb2.loc.gov/ammem/prjplan.html

The New York Public Library's Planning Digital Projects for Historical Collections
http://digital.nypl.org/brochure/index.html

Scoping the Future of Oxford's Digital Collections
http://www.bodley.ox.ac.uk/scoping/
see particularly Decision Matrix/Workflows at http://www.bodley.ox.ac.uk/scoping/matrix.htm

RLG Tools for Digital Imaging
http://www.rlg.org/preserv/RLGtools.html

Appendix 6: MOA DTD

Please note that this DTD is an historical artifact, included to illustrate the nature of the MoA markup, not a current, working version. The DTD is edited and adapted on an ongoing basis. For the most current version or further information, contact umdl-help@umich.edu.

```
<!ELEMENT moa o o (tei.2)+ >

<!ELEMENT tei.2 - - (teiheader, text) >
<!ATTLIST tei.2     ana CDATA #IMPLIED >
<!ELEMENT teiheader - - (filedesc, profiledesc) >
<!ELEMENT filedesc - - (titlestmt, extent, publicationstmt, sourcedesc) >
<!ELEMENT titlestmt - - (title+, author?, respstmt*) >
<!ELEMENT title - - (#PCDATA) >
<!ATTLIST title            type  (245 | main | other | art)  "245" >
<!ELEMENT author - - (#PCDATA) >
<!ELEMENT respstmt - - (resp, name) >
<!ELEMENT name - - (#PCDATA) >
<!ELEMENT resp - - (#PCDATA) >
<!ELEMENT extent - - (#PCDATA) >
<!ELEMENT publicationstmt - - (publisher, pubplace, date?,
                        idno*, availability*) >
<!ELEMENT publisher - - (#PCDATA) >
<!ELEMENT pubplace - - (#PCDATA) >
<!ELEMENT date - - (#PCDATA) >
<!ELEMENT idno - - (#PCDATA) >
<!ATTLIST idno  type  (notis | rootid)  "notis" >
<!ELEMENT availability - - (p)+ >
<!ELEMENT p - - ( #PCDATA | pb )+ >
<!ELEMENT sourcedesc - - (biblfull | bibl) >
<!ELEMENT bibl - - ((author)*, title+, publisher?,
                pubplace?, date?, note?, biblscope+) >
<!ELEMENT biblscope - - (#PCDATA) >
<!ATTLIST biblscope     type CDATA #IMPLIED >
<!ELEMENT biblfull - - (titlestmt, editionstmt?,
                    publicationstmt, seriesstmt?, notesstmt?) >
<!ELEMENT editionstmt - - (edition) >
<!ELEMENT edition - - (#PCDATA) >
<!ELEMENT seriesstmt - - (title) >
<!ELEMENT notesstmt - - (note+) >
<!ELEMENT note - - (#PCDATA) >
<!ATTLIST note       type  CDATA  #IMPLIED >
<!ELEMENT profiledesc - - (textclass) >
<!ELEMENT textclass - - (keywords, classcode?) >
<!ELEMENT keywords - - (term+) >
<!ELEMENT term - - (#PCDATA) >
<!ELEMENT classcode - - (#PCDATA) >
<!ELEMENT text - - (front?, body, back?) >
<!ELEMENT front - - (div1)* >
<!ELEMENT body - - (div1)* >
<!ELEMENT back - - (div1)* >

<!ELEMENT div1 - - (bibl?, p+) >
```

```
<!ATTLIST div1       type  CDATA  #IMPLIED
                     decls CDATA  #IMPLIED
                id    ID     #IMPLIED >
<!ELEMENT pb - o (#PCDATA) >
<!-- ref = page image reference; cnf = OCR confidence for fall '98;
    ftr = feature, as in "TOC1"; seq = page sequence; n = actual page number;
    res and fmt should be self-explanatory -->
<!ATTLIST pb         ref NAME  #IMPLIED
                     seq CDATA #IMPLIED
                res CDATA "600dpi"
                fmt CDATA "TIFF5.0"
                     ftr CDATA #IMPLIED
                cnf CDATA #IMPLIED
                     n   CDATA  #REQUIRED >
```

Appendix 7: RFP

Included in the paper version of this report is the RFP used to award the scanning contact for MoA4. For the electronic copy, submitted to the Mellon foundation, please see the attachment that accompanied this handbook.