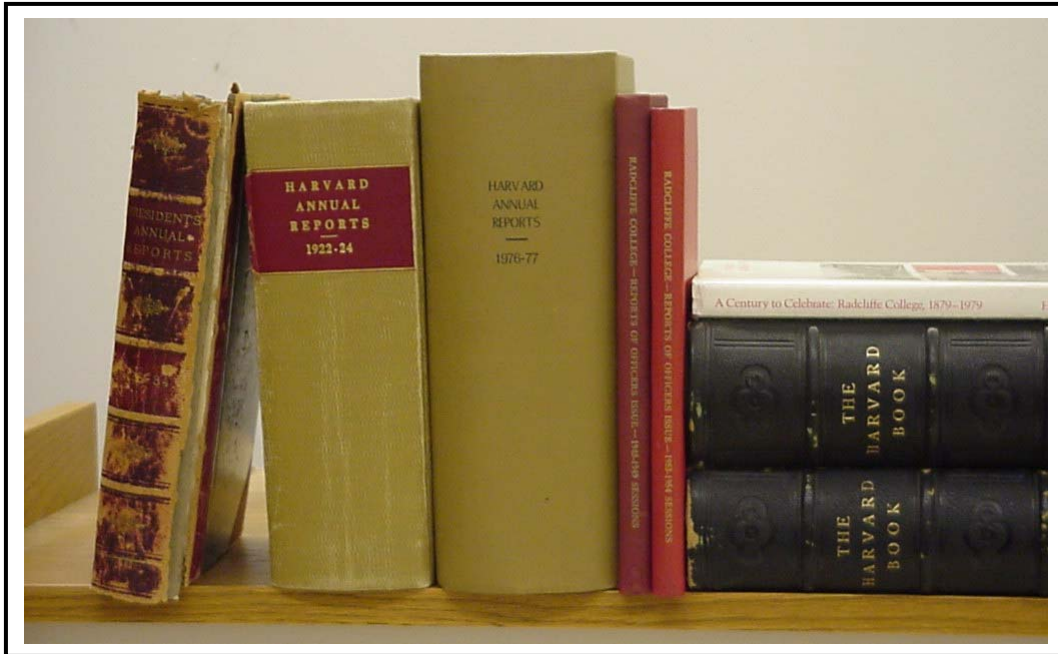


The Harvard-Radcliffe Online Historical Reference Shelf
LDI Round One Project



Completion Report
April 23, 2003

Submitted by
Kate Bowers
Jane Knowles
Robin McElheny

Table of Contents

Preface	2
1. Introduction	3
2. Project history	4
3. Project goals	5
4. Preliminary work	5
5. Creating digital content	6
• Conversion Workflow chart	11
6. Building the metadata	12
7. Storing and delivering the content	15
8. Building the Reference Shelf	18
9. Access to the Reference Shelf and its contents	18
10. Evaluating the Reference Shelf	18
11. Lessons we learned	20
12. The future of the Reference Shelf	21
13. Additional information	
• Project statistics	23
• Project participants	25
• LDI project budget	26
• Publicity timeline	27

Preface

When we – Jane Knowles, Robin McElheny, Kate Bowers, and other archivists at Harvard and Radcliffe – first thought of the Online Historical Reference Shelf, we knew what we wanted but we had no idea of how to achieve it. Thanks to the expertise, knowledge, persistence, and optimism of the University Library’s LDI staff, nearly four years later we now have a resource on which we rely every day. Particular thanks are due to Wendy Gogel, whose superb project management skills and patience helped us bring this project to a successful completion.

This completion report is long, detailed, and dry, but we decided to include all of the documentation that we had compiled throughout the project as a convenience to anyone, ourselves included, who may need to find out what we did and how we did it. We hope that other archivists and librarians planning similar projects will find this information to be useful.

Note: All photographs in this report were taken by Lewis Day, Cataloging Assistant in the Harvard University Archives.

1. Introduction

The Harvard University Archives and the Radcliffe Archives are charged with the responsibility for collecting records and documenting the history of their respective institutions. Harvard is the oldest institution of higher education in America and the model for many North American colleges and universities. Radcliffe (before it merged with Harvard in October 1999) was one of the oldest and foremost colleges for the education of women. Both have played a significant part in the development and direction of higher education in America. As a result, their archives are resources for a broad spectrum of research and scholarship.

The Harvard-Radcliffe Online Historical Reference Shelf (HROHRS) project, through the innovative technology of the Library Digital Initiative, offered an opportunity to make available to the local academic community and the research public a sampling of documents that have framed the history of Harvard and Radcliffe and are essential for the study of higher education in America.

Harvard and Radcliffe archivists envisaged the Reference Shelf as a multi-phased project to provide access to heavily-used historical resources encompassing core aspects of life at Harvard and Radcliffe, from students and alumni, faculty and curriculum, to athletics, extra-curricular life, philanthropy, the built environment, and administration. In the first phase of the project, three categories of reference works were included:

- The annual reports of the Presidents of Harvard and Radcliffe
- Narrative histories: *The Harvard Book* (1875) and *A Century to Celebrate, Radcliffe College 1878-1978* (1979)
- Key archival documents

and links to other online historical information.

We hoped that additional resources, such as the 1930 *Quinquennial Catalogue of Harvard University* and the 1968 *Radcliffe College Alumnae Directory*, would come in a later phase. These would provide searchable access to comprehensive lists of Harvard faculty and graduates from 1636 to 1930 and Radcliffe students from 1879 through 1968, including years of study, academic degrees, and degree dates. Other material would be linked as it became available.

The Harvard-Radcliffe Online Historical Reference Shelf (HROHRS) went live for the first time on September 7, 2001. Nearly 17 months later, on Monday, January 27, 2003, Phase I of the Reference Shelf reached completion. This first phase encompassed the following tasks:

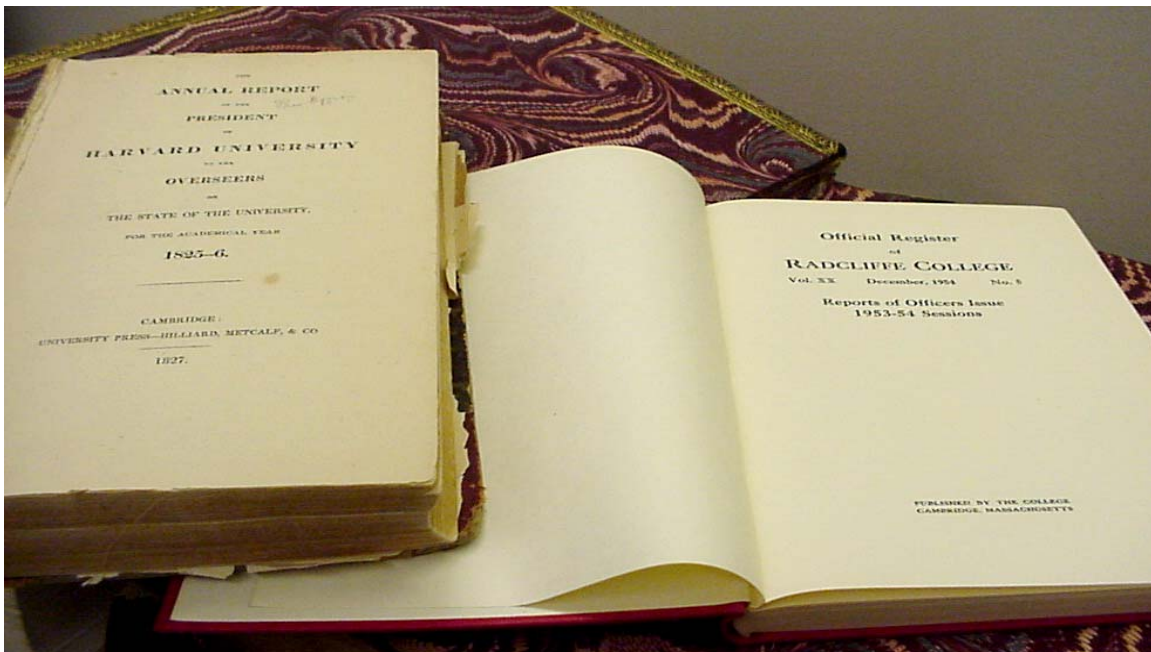
- Creating digital content for the Reference Shelf
- Building the Reference Shelf web site (<http://hul.harvard.edu/huarc/refshelf/>)
- Linking the Reference Shelf to the new digital content and to existing digital content elsewhere at the University
- Conducting a user evaluation of the Reference Shelf

The project, which began in March 1999, took 46 months and cost \$40,917 in direct project funds (see Section 14, Additional Information, for details). It should be noted that most of the project expenses, including project management, systems development, and programming, were incurred as indirect expenses.

2. Project history

The seeds for this project were sown in the early 1990s, when the Harvard University Archives, faced with multiple crumbling copies of the annual reports of the presidents and treasurers of Harvard, determined that reformatting was the only way to ensure long-term preservation of the informational content. A similar situation existed at Radcliffe.

In addition to their deteriorating physical condition, access to the reports through HOLLIS was difficult because of twenty six different title entries, and the reports were only available at the Harvard and Radcliffe Archives. Yet these sources provided the most detailed view of life at both institutions, from the travails of faculty and administrators coping with unruly students, to the establishments of new academic disciplines and fields of research.



At the time, microfilm was the gold standard for preservation reformatting, so the Harvard Archives carefully assembled three complete sets of reports – two for ready reference use in the Archives’ reading room, and one to be used for microfilming. At the same time, the Radcliffe Archives was assembling its own set of annual reports for microfilming. Money, however, posed a considerable obstacle to proceeding with these projects. In 1993, the Harvard Archives received a cost estimate of nearly \$16,000 to prepare a master microfilm negative of the Harvard reports alone. In spite of universal agreement that the annual reports were vital sources of information on the history of Harvard and Radcliffe, no funds were available for these projects, so the microfilm was created.

3. Project goals

By 1999, after the University Library launched the Library Digital Initiative, archivists at Harvard and Radcliffe were aware of the promise that digital access offered, both as a way to improve resource discovery and as a way to improve access to the content of important historical sources. We submitted our proposal for the Harvard-Radcliffe Online Historical Reference Shelf with three goals in mind:

- To establish a convenient on-line source of historical information about Harvard and Radcliffe – the Online Reference Shelf itself
- To create new digital content for the Reference Shelf
- To establish links from the Reference Shelf to existing digital content

We also wanted to provide three types of functionality for the new digital content – browsing, full-text searching, and printing.

Rather than providing electronic access to particularly old or unusual materials (“hidden treasures”), we wanted to enhance access to materials that researchers routinely consult – true reference sources.

For several reasons, the annual reports were prime candidates for digital conversion. As primary sources on the history of Harvard and Radcliffe, they are the starting place for much of the research that takes place in the Harvard and Radcliffe Archives. Digital reformatting could expand access to these information-rich sources by providing full-text searching capability. Online availability would also provide round-the-clock access for the broader University community, even for researchers in remote locations. In addition, the Archives already had complete sets of the reports set aside for reformatting. These copies could be disbound for efficient scanning.¹

In addition to the annual reports, we chose to digitize *A Century to Celebrate*, a narrative history of Radcliffe, to complement *The Harvard Book*, a narrative history of Harvard that was digitized in a 1997 pilot project by the Weissman Preservation Center.

As a corollary to the project, we planned to address long-term preservation of the annual reports by making copies of the digital files on silver halide microfilm. The microfilm was to be paid for by the Harvard and Radcliffe Archives with non-LDI funds.

The Harvard-Radcliffe Online Historical Reference Shelf project was approved for LDI funding in March 1999.

4. Preliminary work

On receipt of funding, we wrote up a detailed plan of work and schedule, with the following steps:

- Preparing the annual reports and narrative history for scanning

¹ In preparing our project proposal, HCL DIG scanned two Harvard reports, one from the mid-19th century and one from the mid-20th century, to see whether the disbound pages could survive an automatic document feeder. Neither report sustained any damage.

- Upgrading the catalog records in HOLLIS for the annual reports and narrative histories
- Scanning the pages to create digital images of each page
- Processing the page image files with Optical Character Recognition (OCR) software to create searchable text files
- Testing the resulting text files to determine if the searching success rate was acceptable
- Re-keying those text files for which the searching success rate was not acceptable
- Marking up the text files for online delivery to users
- Building the Reference Shelf
- Linking the various historical sources to the Reference Shelf
- Evaluating the effectiveness of the Reference Shelf

Concurrent with project work, LDI staff would build the storage and delivery systems required to support online delivery of the historical sources.

5. Creating digital content

In the course of this project, we created four categories of digital content - the Harvard and Radcliffe annual reports, narrative histories, founding documents, and an architectural survey of Cambridge. We relied on two production methods to create the digital content, although each category required certain adjustments to the work flow. For an illustrated overview of our methods, see the Conversion Workflow chart at the end of this section.

Production method #1

We developed this method to handle printed pages of text, beginning with the Harvard and Radcliffe annual reports. It involved the following steps:

- Preparing the volumes for scanning by collating and disbinding them
- Scanning the pages using a flatbed scanner with an automatic feed to create 1-bit (bi-tonal) 600 dpi TIFF master images of all pages
- Creating searchable text files from the page images using OCR software

In June and July 1999, a project assistant working at the Harvard Archives collated the reports, counting pages and flagging missing pages, color plates, and fold-outs. After the reports were disbound by staff in the College Library Conservation Services, the College Library Digital Imaging Group (HCL DIG) scanned them to create the TIFF page images (electronic “photocopies” of each page). Researchers see lower-resolution delivery versions of these files when they browse the annual reports on line.

TABULAR VIEWS.

[For the purpose of presenting the subjects in the order believed most favourable for reference and comparison, the Statements **FIRST** and **SECOND**, in the vote of the Board, have in the following Views been transposed.]

I. [II.] THE STATE OF THE DEPARTMENTS. [See Remarks at the end of the Table.]

1	2	3	4	5	6	7	8					
Name of each Department.	Name of each Instructor in each Department.	Class.	No. of Exercises to each Class.	No. of Lectures to each Class.	No. of obligatory Exercises.	No. of obligatory Lectures.	No. of optional Exercises.	No. of optional Lectures.	No. of written Exercises.	Whole No. of Exercises.	Whole No. of Lectures.	Whole No. of written Exercises.
Latin,	} Rev. George Otis, A. M. College Prof. John Fessenden, A. M. Tutor.	{ Senior,	130	130	} 658		
		{ Junior,	125	125			
		{ Sophomore, Freshman,	176 227	176 149			
Greek,	} Rev. J. S. Popkin, D. D. Eliot Prof. George R. Noyes, A. M. Tutor.	{ Senior,	130	130	} 658		
		{ Junior,	125	125			
		{ Sophomore, Freshman,	176 227	176 149			
Hebrew and other Oriental Languages,	} Sidney Willard, A. M. Hancock Prof.	{ Th. Stu.	86	. . .	86	. . .	130	} 276		
		{ Senior,	130	130			
		{ Junior,	60	60			

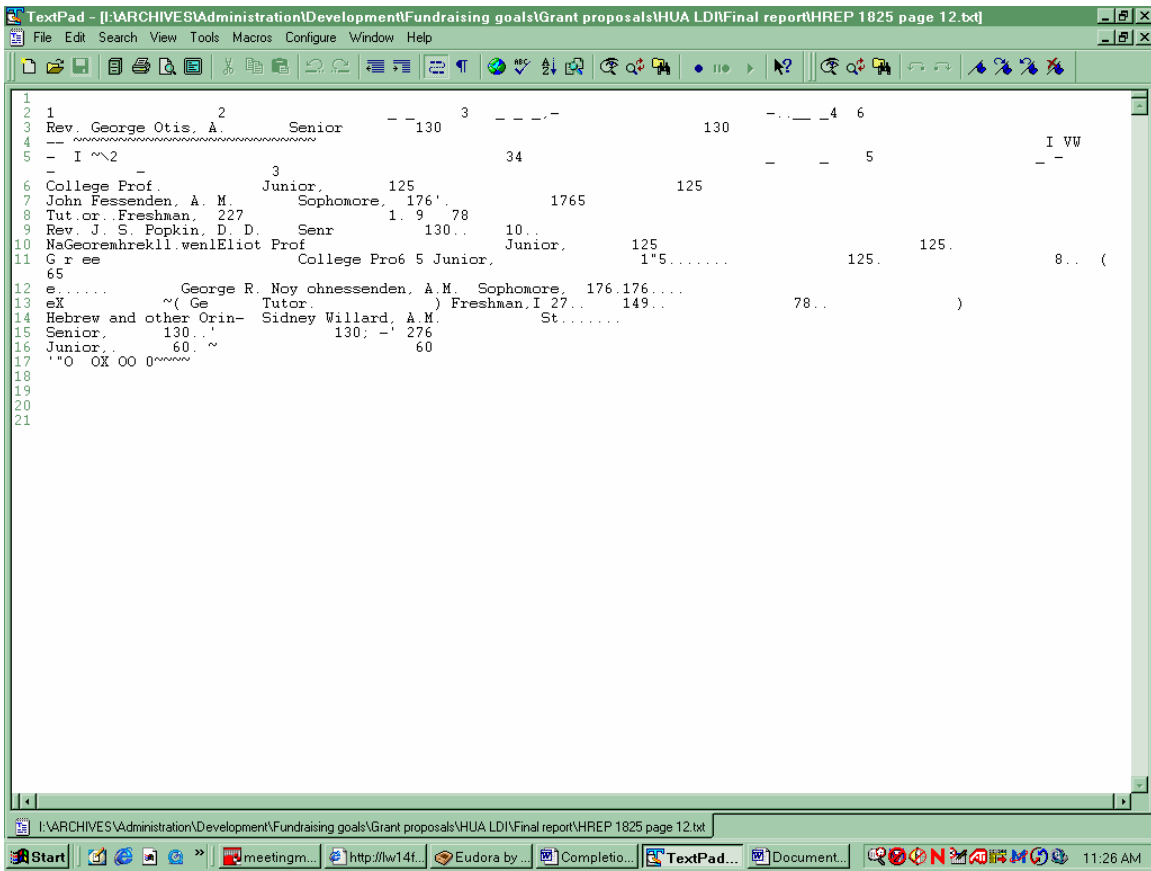
12

Page 12 of 1825/26 Harvard president's report. Note that text is oriented in three directions.

As part of the scanning process, HCL DIG compiled administrative and technical metadata documenting when, where, and how the digital files were created, as well as file formats, sizes, and ownership. In addition, HCL DIG created spreadsheets with structural metadata to accompany the page image files. This included file sequence numbers, printed enumeration (the page numbers that appear on the paper originals), and feature codes indicating image orientation and special formats, such as foldouts and two-column indexes.

In early August, HCL DIG began sending the bi-tonal page image files and corresponding metadata spreadsheets to the University of Michigan Digital Library Production Service for OCR processing² to create searchable text files. When researchers search the annual reports, the search engine uses the text files, although the searchable text remains hidden behind the page images on-screen.

² The University of Michigan Digital Library Production Service was chosen as the contract vendor for OCR processing because it had developed OCR software that is specifically designed to handle older typefaces and page formats.



Corresponding text file for page 12 of 1825/26 Harvard president’s report

We had assumed from the outset that the non-standard sizes and styles of typefaces used by printers in the 19th century would make the earlier annual reports unsuitable for OCR conversion. Thus we planned to conduct a search-and-retrieval test to determine whether the quality of the text files created by OCR conversion was high enough to support useful full text searching. If not, we would need to hire a vendor to re-key the annual reports in order to create the searchable text files.

By the end of October 1999, 179 reports had gone through the scanning and OCR conversion process, and were ready for the search-and-retrieval test. The test was based on a random sample of 2000 page images (5% of the total in the first batch of reports), with the sample set weighted to pages from pre-1879 reports. Working from the printed reports, staff at the Harvard and Radcliffe Archives entered a word or phrase in a prototype search engine and noted whether the electronic image of the page from which the word was chosen was retrieved. As it turned out, the success rate of retrieval was high enough (97%) to preclude the need for rekeying text.

Production method #2

We developed a second method to digitize half-tone photomechanical illustrations (printed versions of photographs) as well as color illustrations and covers. It involved the following steps:

- Scanning half-tone photomechanical illustrations in the bound volumes using an overhead digital camera to create 300 dpi 8-bit grayscale TIFF master image files
- Scanning color illustrations and covers using an overhead digital camera to create 300 dpi 24-bit color TIFF master image files
- Deriving 100 dpi JPEG versions of the grayscale and color master files for online delivery
- Deriving bi-tonal page images from the TIFF master page images for OCR processing³

We used this method to upgrade an existing digital version of *The Harvard Book*, a two-volume history published in 1875. It contains pages of text and line art (small, high-contrast illustrations) interleaved with full-page half-tone photomechanical illustrations. HCL DIG had scanned it in a 1997 pilot project, so bi-tonal page image files already existed. We adjusted our digitization work flow as follows:

- Created searchable text files from the bi-tonal page images using OCR software
- Re-scanned the half-tone photomechanical illustrations in the bound volumes using an overhead digital camera to create grayscale TIFF master image files
- Substituted grayscale TIFF masters for existing bi-tonal versions of the photomechanical illustrations in each volume
- Derived JPEG versions of the grayscale TIFF master files for online delivery

A Century to Celebrate, a one-volume history of Radcliffe published in 1979, contains numerous half-tone photomechanical illustrations interspersed throughout the text. Thus we varied our production methods as follows:

- Scanned the bound volume using an overhead digital camera to create 300 dpi 8-bit grayscale TIFF master images of all pages and derived 100 dpi JPEG versions of these files for online delivery
- Derived bi-tonal page images from the TIFF page images for OCR processing to create searchable text files

We varied this production method once again to scan selected color covers and tables in the annual reports, from which we created 300 dpi 24-bit color TIFF master images and derived 100 dpi JPEG versions for online delivery.

Production method #3

In January 2002, the Harvard Archives acquired Christopher Hail's *Cambridge Buildings and Architects* (CBA), a four-volume historical survey organized by street address. Anticipating a high level of interest in the printed volumes, we asked Chris if he would give us an electronic version for the Reference Shelf. He agreed, offering us four text files and 351 digital photographs.

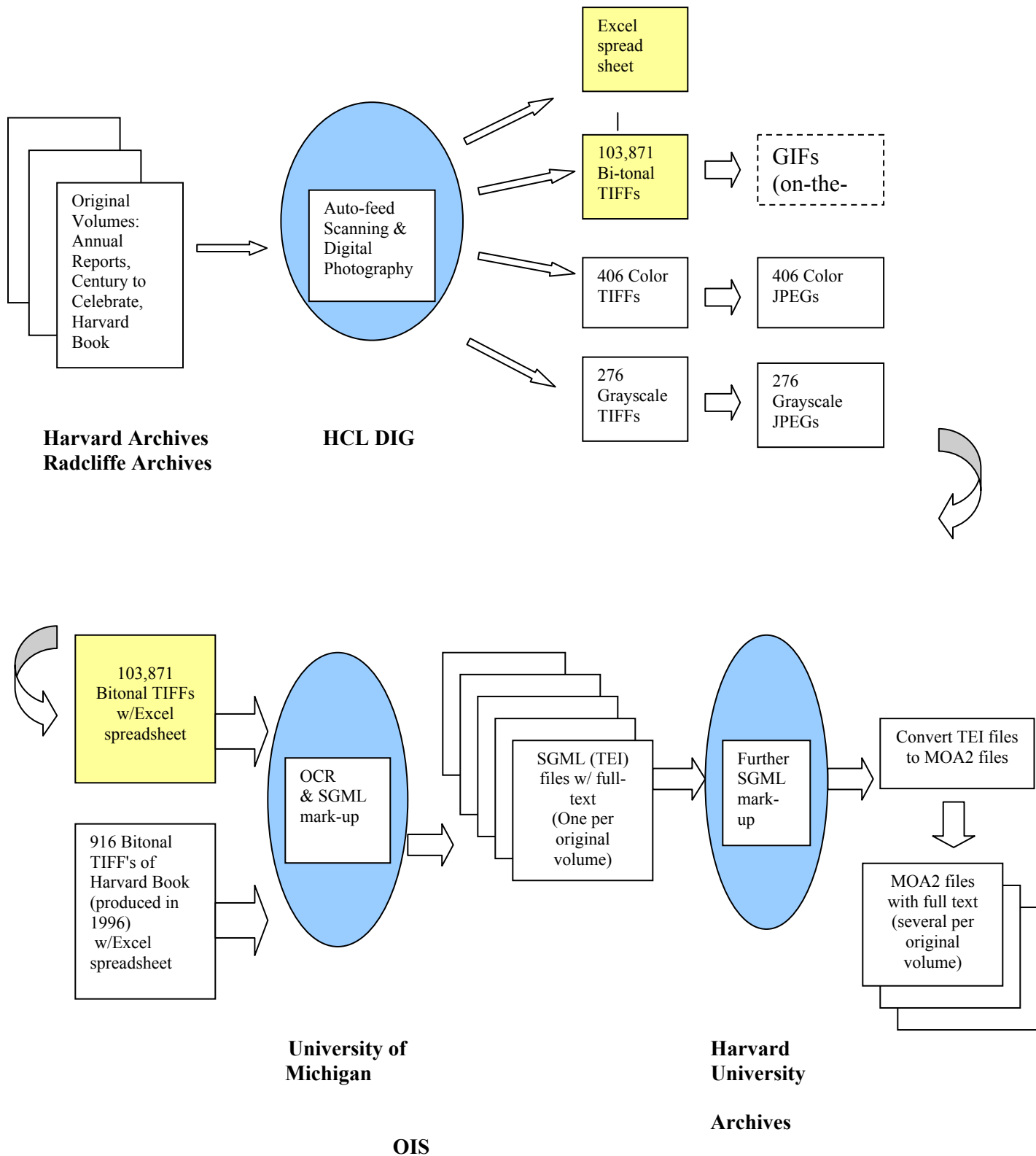
³ Optimal OCR processing depends on bi-tonal image files.

The CBA in electronic form presented us with an interesting dilemma. The four-volume version consisted of a report from a database that Chris had enhanced as a word-processing file. Ideally, we would have chosen to present the survey in its original form as a searchable database, but no database delivery system was immediately available.⁴ As an interim solution, we opted to present it as a text report marked up in HTML.

To prepare the survey for web delivery, we developed a third production method. In the summer of 2002, we hired a web designer, who marked up the text files, linked the text files to the corresponding digital photographs, and designed the CBA web pages.

⁴ At this time, the TEmplated Database system was in the early stages of planning.

Online Historical Reference Shelf: Conversion Workflow



6. Building the metadata

Metadata in various guises is critical to finding and using information. In libraries, we rely on descriptive metadata, such as catalog records, to determine which books we want to read. We rely on administrative metadata, such as call numbers, to locate the books on library shelves. We rely on structural metadata, such as tables of contents, indexes, and page numbers, to navigate through the books to find the information we need.

Descriptive, administrative, and structural metadata serve the same roles in the digital world. In the analog world, publishers supply the structural metadata for books, while librarians supply the descriptive and administrative metadata. In the world of digital reformatting, librarians must supply the structural metadata as well.

Cataloging the content

Concurrent with scanning and OCR processing, we upgraded the catalog records for the paper originals and cataloged the digital content, including the Reference Shelf itself, in HOLLIS. This proved to be a critical step in the project because the LDI delivery systems rely on bibliographic metadata to provide online access to the digital content.

Cataloging upgrades included the addition of path names that would eventually link to the digital content, but we suppressed them from the online public access catalog until the content and the delivery systems were ready. To facilitate resource discovery in HOLLIS, we also added standard title added entries, including “Annual Reports of Harvard University” and “Annual Reports of Radcliffe College,” to each of the 26 bibliographic records for the annual reports.⁵

Creating administrative and structural metadata

As noted above in Section 5, the creation of administrative, technical, and structural metadata for the annual reports, narrative histories, and founding documents began in the scanning phase at HCL DIG. Structural metadata was assembled in spreadsheets that accompanied the page image files when they were sent to the University of Michigan for OCR processing.

Digital Library Production staff at the University of Michigan incorporated the structural metadata from the spreadsheets into the OCR text files to create text files encoded with Standard Generalized Markup Language (SGML). This began the process of establishing page-turning and other navigational functionality in the digital volumes.

File mark-up

Concurrent with digitization, we planned for the file mark-up phase, during which we would add more structural metadata to demarcate sections within each volume. An electronic table of contents based on this metadata would allow researchers to navigate section-by-section through a report, rather than having to move page-by-page through the volume or continually having to return to an online list of search results in order to move

⁵ Without these standard titles, it is difficult to find all of the annual reports in HOLLIS. Their titles changed many times over the course of their 175+ years of publication, so that researchers must remember 26 separate titles ranging from the straightforward *The President's Report* to the less-obvious *Issue Containing the Report of the President of...*

directly to relevant sections. We decided to base the mark-up on an exact replication of the printed tables of contents, with the creation of electronic tables of contents where no printed ones exist (in the early Harvard reports, for example).

In February 2000, we hired a project mark-up assistant to enter the additional structural metadata in the basic SGML files created by the University of Michigan. The project assistant used a greatly modified version of the Text Encoding Initiative Document Type Definition as the mark-up schema⁶. LDI staff created file header templates that included bibliographic descriptions of each report (based on the catalog records described above), as well as some pre-defined administrative metadata. These templates, into which the mark-up assistant entered new structural metadata and volume-specific descriptive metadata, such as publication date, helped to reduce the data entry and ensured consistency.

As the first digital candidates for the Reference Shelf, the annual reports presented a set of unique challenges for mark-up. The internal structure of the reports varied throughout their publication history, so that some volumes have sections and others have sub-sections. With the digital delivery system still under development during the mark-up phase of our project, we were not sure whether these varying levels of hierarchy could be successfully reproduced online. In the end, we opted to flatten the internal structure to one level throughout the reports. Thus we created intra-volume navigation and online tables of contents that are consistent from volume to volume, with a result that is practical for the user and relatively faithful to the original structure of the reports.

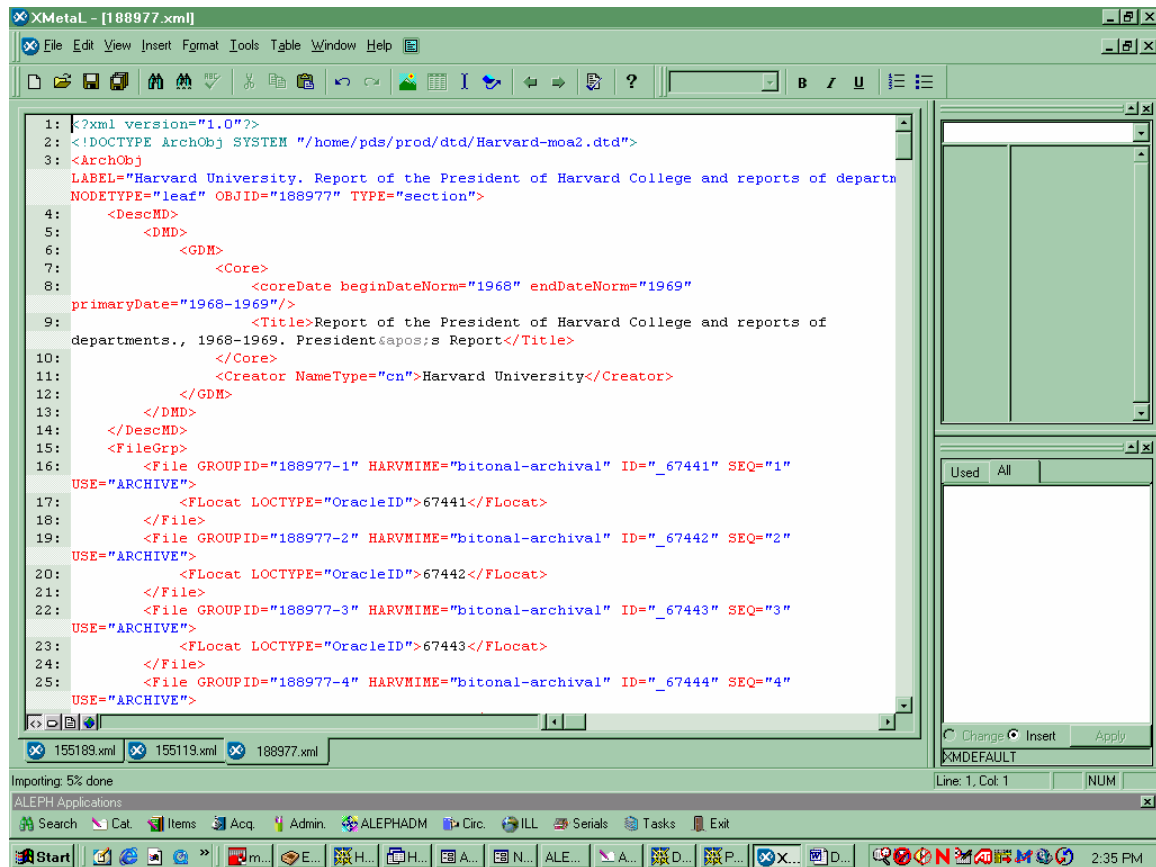
The project assistant marked up front matter, back matter, and sections, as well as specifying the roles of certain pages, such as title pages of volumes and first pages of sections. This careful section-by-section review turned up a new category of content for the Reference Shelf. One section of the Harvard *Treasurer's Statement* of 1835 had been distributed and bound in with the *Statement* but was, in fact, a separate publication. The *Constitutional articles and legislative enactments relative to the Board of Overseers and the Corporation of Harvard University* contains a compilation of all laws from 1642 to 1814 concerning Harvard, including the text of Harvard's charter of 1650. We decided to classify this publication as "founding documents." To find all of this legislation as separate records would require many hours of research in the Massachusetts state archives. Not only is the published compilation now available online, it is also keyword-searchable.

By the time file mark-up was completed in October 2002, LDI staff had determined that the delivery systems for digital content would be based on the Making Of America 2 (MOA2) DTD rather than on the TEI DTD. One reason for this revised approach is the flexibility of the MOA2 framework, which can accommodate source materials in a broad range of formats, from single-sheet handwritten documents and photograph albums to published serials.

⁶ The TEI DTD was developed as a way to facilitate the delivery and sharing of electronic texts created by reformatting printed publications. It provides a framework that can accommodate the internal structure of an electronic text document, such as chapters in a book.

LDI staff was able to convert the TEI files to MOA2 files, so no additional mark-up was needed, but this shift in approach meant that the initial functionality of the delivery systems was somewhat different from what we originally had in mind. Some of the indexing and display features that we wanted would not be immediately available.

The MOA2 files, identified as “nodes,” recreate in electronic form the hierarchy of relationships between the titles, volumes, sections, and pages in the printed originals. Citation nodes correspond to bibliographic citations and link them to their component volumes. Intermediate nodes correspond to volumes and link them to their component sections. Leaf nodes correspond to sections and link them to their component pages.



MOA2 “leaf node” file for the “President’s Report” section of the *Report of the President of Harvard College and reports of departments, 1968-1969*. Note that this file contains links to the corresponding bi-tonal page images, with file i.d.s and sequence numbers.

File management

In addition to mark-up, the project assistant reviewed the electronic page files for quality control, in the course of which she found problems with 45 files and original volumes. These problems included the following:

- Incomplete hard copy scanned, resulting in incomplete electronic files
- Complete hard copy sent for scanning but electronic files incomplete

- Incorrect page sequence in electronic files
- Inadequate OCR of a few selected pages
- Electronic files with incomplete or incorrect metadata

After consulting with project managers and LDI staff, the project assistant was able to correct some problems herself. She re-keyed text for which the OCR was inadequate, corrected some metadata, and located replacement hard copy for scanning. Other problems, involving the insertion of new files in established page sequences and the replacement of incorrect files with corrected versions, were more difficult to solve because they required coordinated efforts by HCL DIG staff and the project assistant.

Resolution of these problems called attention to the complexity of managing so many interrelated files and the need for a file management phase of the project. As a result, we hired the mark-up assistant for another five weeks in November 2000 to assemble an authoritative set of image and text files and ready them for transfer to the Digital Repository.

7. Storing and delivering the content

In addition to the World Wide Web, the Reference Shelf relies on four storage, access, and delivery systems developed by LDI staff: the Digital Repository Services (DRS), the Name Resolution Service (NRS), the Page Delivery System (PDS), and the Full-text Search Service (FTS).

Development of the four LDI systems occurred simultaneously with the creation of digital content, so that we had to adjust our project methodologies along the way. As first-round project participants, on the other hand, we were able to influence the design and functionality of the systems. With Kate Bowers as a member of the Digital Collections Systems Steering Committee, we expect to have a continuing voice in the adjustment and enhancement of the systems.

File storage

The electronic versions of the annual reports, narrative histories, and founding documents are stored in the DRS. They consist of two file types: high-resolution TIFF master images of text pages, illustrations, and covers, and lower-resolution JPEG delivery versions of illustrations and covers. (The delivery versions of text pages are not stored in the DRS but are created on-the-fly.)

Now that the files are stored in the DRS, we have reached a tacit agreement that we will not create copy the files onto microfilm but will rely on the DRS for long-term preservation instead.

The architectural survey, which is stored on a web server, consists of HTML-encoded ASCII text files and color JPEG versions of photographs.

File delivery

At the project's outset, we planned to use two delivery systems: a page-turner service that would allow page-by-page browsing in the online volumes, and an indexing service that would support searching in the OCR text files with links to the corresponding page image files for online viewing. LDI staff developed both of these systems, building a printing function into the page-turner service.

The annual reports, narrative histories, and founding documents are delivered via the Page Delivery System. To ensure quick delivery of page images on-screen, the PDS delivers compressed GIF images, created on-the-fly from the TIFF master files, of the text pages in the annual reports, *The Harvard Book*, and the founding documents. For the same reason, the PDS delivers compressed JPEG images of the color illustrations and color covers of the annual reports and continuous-tone illustrations in *The Harvard Book*. Due to the number of photographic illustrations interspersed throughout the text in *A Century to Celebrate*, the PDS delivers JPEG images of all pages in the volume.

The architectural survey is delivered via the World Wide Web.

Full-text searching

All of the Reference Shelf resources that are delivered via the Page Delivery Service – the annual reports, narrative histories, and founding documents - can be searched using the Full-text Search Service. The FTS search capability is based on bibliographic citations, or titles, thereby allowing searches within one citation at a time. In other words, when a researcher selects a particular citation and links to it, s/he can search within all of the volumes covered by that citation and *only* within those volumes.

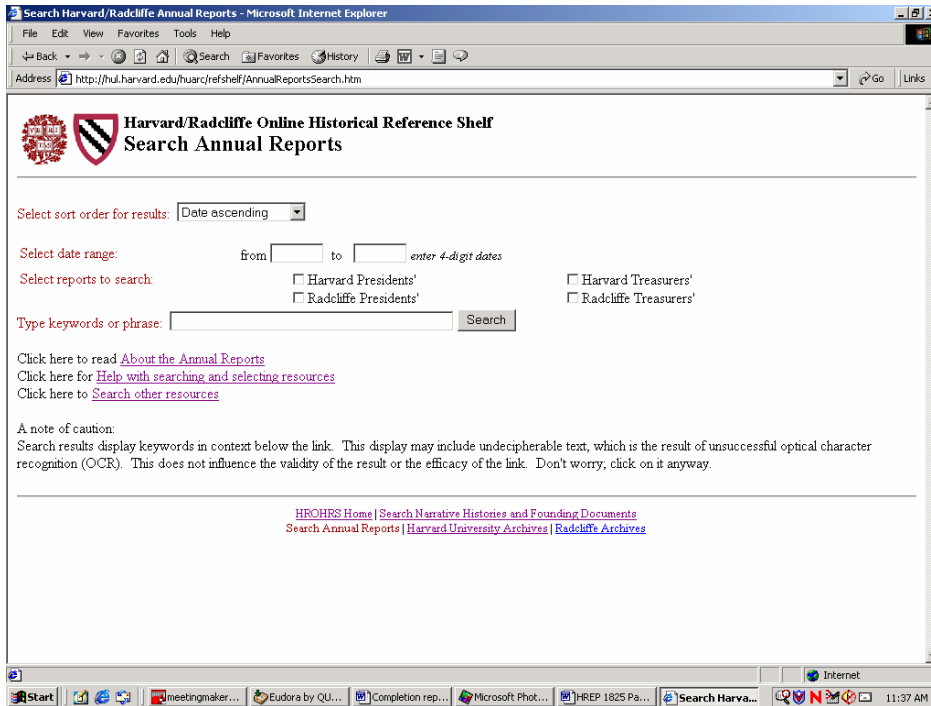
In addition to the citation-specific searches built into the PDS and FTS, we wanted to give researchers two additional methods of searching, both of which are important to historians and users of periodicals:

- The ability to search among multiple citations simultaneously
- The ability to limit searches by date range

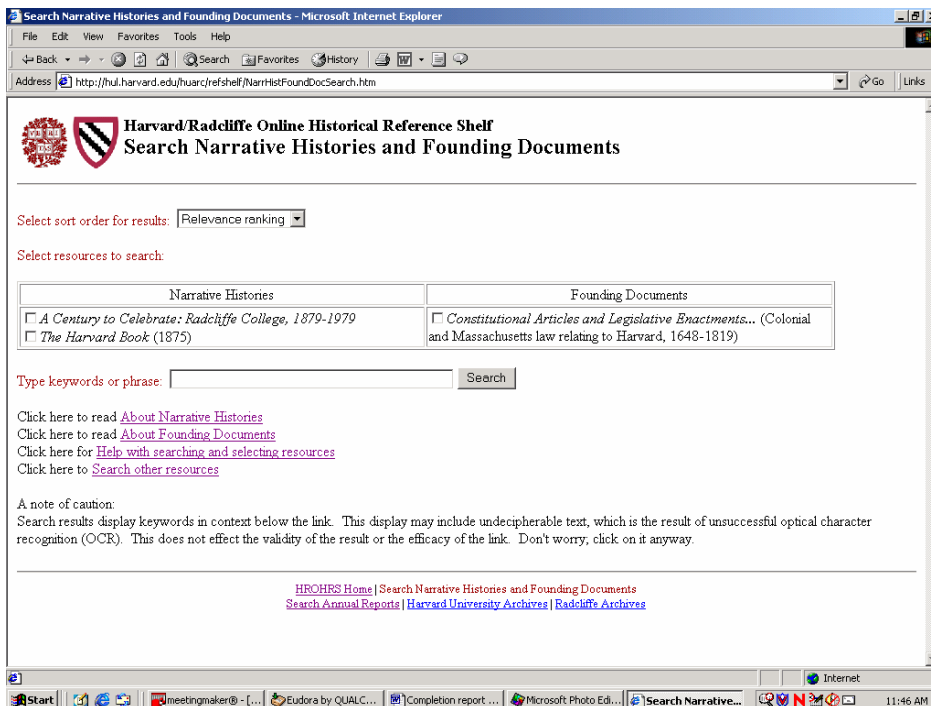
The dates available for searching and sorting in the FTS are the publication dates in the descriptive metadata added to the text files during the mark-up phase. Whereas the publication dates of volumes of the annual reports are contemporaneous with the events that they document, this is not true of the publication dates of the narrative histories or founding documents. Searching all resources at once, without an understanding of these differences, would misdirect the user, defer pertinent results to later years in the date sort, or falsely disqualify legitimate results.

As a way to ensure that a researcher would understand the nature of his/her search results, we created two search interfaces outside of the PDS: one for contemporaneous accounts (such as annual reports and anticipated additions such as diaries) and one for documents whose dates of publication are not necessarily the same as dates of coverage or enforcement (such as narrative histories, memoirs, and laws). We do not provide date

range searching for those documents in which the publication dates do not correlate with coverage dates.



Search interface for annual reports



Search interface for narrative histories and founding documents

Printing

The Page Delivery System includes a printing function that allows researchers to print a page or a section of a volume by creating PDF files on-the-fly.

8. Building the Reference Shelf

At the start of the project, LDI staff planned to build a dynamic web page server, a system that would provide a centralized web delivery framework, into which librarians could insert content without having to build separate web sites for each purpose. We planned to use the LDI system to build and support the user interface for the Reference Shelf. In July 2001, the University Library Council decided to delay development of the system for at least another year, so we proceeded to design the Reference Shelf in the standard way, as an assemblage of static web pages.

Radcliffe and Harvard archivists agreed upon basic page hierarchy, content, look, and feel, with simplicity and low maintenance as overarching goals. One archivist designed the original pages. A second archivist added content, search capabilities, and supporting pages such as “help” and “about.”

The Reference Shelf home page offers a list of categories of resources, each of which is linked to an intermediate “splash” page, with detailed descriptive information about the resources, or directly to the resource. Simple additions to the Reference Shelf can be made by constructing a page for the new resource, establishing a link to the resource using the Name Resolution Service, and updating one of the Reference Shelf pages.

9. Access to the Reference Shelf and its contents

The Reference Shelf is publicly available on the web at <http://hul.harvard.edu/huarc/refshelf/>. It can also be reached from the University Archives’ web site, from the “E-Resources” list on the Harvard Libraries portal, and from its bibliographic record in the HOLLIS catalog (HOLLIS number 008418732).

The Reference Shelf web site provides links to all of the resources, each of which can also be accessed from their bibliographic records in the HOLLIS catalog.

10. Evaluating the Reference Shelf

After the Reference Shelf first went live in September 2001, we worked with SurveyTools Corporation to design and host a questionnaire that was posted on the Reference Shelf home page. We were ready to post the survey form in March 2002, but opted to delay the evaluation until the beginning of the fall semester in September 2002, when we anticipated heavy use of the Reference Shelf and a good response rate.

In spite of a fluorescent green label announcing the survey at the top of the Reference Shelf home page, constant exhortation by Archives reference staff, and an extended evaluation period (September 2002 through January 2003), the return rate was dismal. We received a total of 18 survey forms (10 in electronic format, 8 on paper), as well as two independent e-mail messages regarding the Reference Shelf.

What the users say

Of the 18 survey respondents, 11 rated the Reference Shelf as “Very useful.” Four respondents urged the addition of more online content, including course catalogs. One person complained about the printing function, finding it “slow and cumbersome.” Another person suggested image rotation, commenting that “the Treasurer's Reports would be a little easier to read in landscape rather than sideways, but this is a small matter compared to the bulk of the very useful materials.”

A researcher at Notre Dame sent the following comments to the reference staff at the Harvard Archives:

“...[Y]our suggestion that I take advantage of [the] newly digitized collection of annual reports of the president and treasurer are proving already fruitful. A quick search for “Dante” on a selected range of reports returned a number of citations relating to course offerings on Dante. Though I didn’t mention during our conversation, such data is actually quite relevant to my project since I am interested in examining the relationship between building collections in Dante and Italian literature and the development of the modern languages curriculum at Harvard. And while I know that such information could be found in bulletins and annual reports and [I] even have access to some of these at Notre Dame, the ease with which I can cull references from your online full-text database is tremendous – a great time-saver! Among the treasurer’s reports in the same I was quickly able to find a reference to one of my “heroes” Theodore Koch, ‘of the Class of 1893, now engaged in cataloging the Dante collection of Cornell University, has presented to our Dante collection many of the Cornell duplicates, while the Dante Society has put another hundred dollars in our hands from which to still further increase our Dante collection.’ Again, it is nice to be able to find such corroborating evidence so easily and quickly.”⁷

What the statistics say

We looked at web server statistics from January through December 2002, to determine whether the Reference Shelf is fulfilling its mission.

- Is the Reference Shelf useful?
We assume so. Sustained use appears to reflect academic activity during the school year. The average level of activity hovered at 450 to 500 hits per week, with a mid-semester peak of 600 to 700 hits per week from mid-September through mid-October 2002.
- Do users take advantage of 24/7 access?
Over the course of each week, the Reference Shelf received peak use in the middle of the week (Tuesday through Thursday) but 15% to 25% of total use occurred on the weekend. We were unable to determine the level of after-hours activity from the available statistics.

⁷ Christian Dupont, e-mail message to Harvard Archives Reference, November 6, 2001.

- Does the Reference Shelf serve a Harvard audience?
According to statistics based on domain names, access from Harvard domains during 2002 ranged from 84% in January to 15% in June and rose to 88% in October.

High-use months = mid-semester months with high use from Harvard domains, which suggests that many of the students who use our collections for course work also use the Reference shelf.

- Do users benefit from the enhanced access offered by full-text searching?
A review of the web logs reveals that the Reference Shelf search pages are the most frequently consulted pages after the Reference Shelf home page. In addition, web logs suggest that researchers use the Reference Shelf search pages rather than the search functionality provided in the Page Delivery Service.

Unanswered questions

- How do people find the digital content associated with the Reference Shelf? Do they go to the Reference Shelf first, or do they begin with a search in HOLLIS?
We are unable to answer these questions from current data.
- Who is using what?
We do not know which user groups are using which resources. This would be useful to know in order to prioritize additions to the Reference Shelf.

11. Lessons we learned

Collate twice, digitize once!

With more careful preliminary assessment of the original sources, we could have anticipated problems and designed alternative work flows to handle missing pages, torn pages, oversize illustrations, and other anomalies. This would have reduced some of the complexity of the file management in later phases of the project.

Never judge a serial by its cover!

When reviewing a publication as a candidate for digitization, it sometimes makes sense to take a fresh look at it. Cataloging rules may not provide the best foundation on which to base resource discovery in a full-text system. (Serial cataloging rules, for example, have varied over the years and not in response to user behavior.) It is more important to understand your users' needs and behavior than it is to accommodate the requirements of an on-line cataloging system.

Electronic files require expert management!

With a project this large (involving more than 200,000 separate files), vigilant and consistent oversight is required. During the mark-up phase, when much of the file-by-file quality control took place, we realized that a subsequent file management phase was needed to resolve problems with file sequences, file names, and file replacement.

Electronic resources ≠ printed counterparts

Careful scrutiny of the original sources during digitization led us to revise our original intention of creating strict electronic facsimiles. In some cases we determined that conventions followed to produce the paper originals were unnecessary or confusing features in the electronic versions. In other cases, time and budget constraints precluded the replication of certain features of the paper sources.

The differences between paper resources and their electronic counterparts were also highlighted during the metadata phase of the project. As the delivery and search systems evolved with their reliance on bibliographic metadata, we learned that bibliographic descriptions of paper documents can conflict with the type of descriptions preferred for delivery of electronic documents.

In the end, the electronic versions are truly new publications, with some look and feel of the originals.

12. The future of the Reference Shelf

Aside from a few technical problems to be addressed in the near term, the Reference Shelf project was a success. Not only did we accomplish our immediate goals of building the Reference Shelf, we also developed several production models for creating the digital content and preparing it for online delivery. We hope to be able to expand the scope and functionality of the Reference Shelf, depending on the available resources.

Residual issues

- A few of the grayscale page images in *The Harvard Book* are out of sequence. Harvard Archives staff will correct this problem.
- Image rotation is needed in the Page Delivery System so that tables and images in a landscape format can be displayed in the proper orientation. This is a top priority on the PDS enhancements list and should be taken care of soon.
- Printing from the Page Delivery System is cumbersome, particularly if you want to print only one page of a work, rather than a complete section.

Additional content and functionality

We had planned on additional content from the beginning of the project, with two directories, the *Quinquennial Catalogue of Harvard University, 1636-1930* and the *Alumnae Directory of Radcliffe College, 1968*, at the top of our list. While researchers consult the annual reports for *what* and *when*, the directories supply information about *who* taught and studied at both institutions. Both resources are starting points for research on individuals. They define an individual's connection to the institution, and contain basic biographical information. The *Quinquennial* includes the names of all presidents, fellows, overseers, and graduates of the College and all the faculties; the *Radcliffe Directory* includes all alumnae, living or dead, who attended the college even if they did not graduate. The biographical database could also be supplemented by biographical files, Faculty of Arts and Sciences memorial minutes, Sibley's *Harvard Graduates*, and Radcliffe's unpublished "Memorial Biographies."

- Leighton Parks (*see Hon. 1900*), Preacher to the University 1891-1894
- Derric Choate Parmenter (*see A.B. 1913*), Instr. in Physical Training 1919-1920; Instr. in Hygiene 1919-1920, 1921-1925; Instr. in Physical Education 1920-1923; Instr. in Industrial Medicine (S. of Public Health) 1924-1929
- Milman Parry, Instr. in Greek and Latin 1929-; Tutor 1929-
- Harry Snow Parsons (*see D.M.D. 1892*), Instr. in Mechanical Dentistry 1893-1895, 1902-1906; Instr. in Operative Dentistry 1895-1896, 1914-1921
- Talcott Parsons, Instr. in Economics 1927-; Tutor 1927-
- Theophilus Parsons (*see A.B. 1769*), Fellow 1806-1812
- Theophilus Parsons (*see A.B. 1815*),
- Endicott Peabody (*see Hon. 1904*), Preacher to the University 1899-1902
- Francis Greenwood Peabody (*see A.B. 1869*), Overseer 1877-1882; Lectr. (Div. S.) 1880-1881; Parkman Prof. of Theology 1881-1886, 1893-1894; Plummer Prof. of Christian Morals 1886-1913; Dean of the Divinity S. 1901-1906; Preacher to the University 1905-1906
- Francis Weld Peabody (*see A.B. 1903*), Asst. Prof. of Medicine 1915-1920; Assoc. Prof. of Medicine 1920-1921; Prof. of Medicine 1921-1927
- Oliver Peabody (*see A.B. 1745*), Librarian 1748-1750
- Robert Swain Peabody (*see A.B. 1866*), Overseer 1888-1899, 1907-1913; Lectr. on Architectural Design 1905-1906; Lectr. (Gr. S. of Appl. Sci.) 1911-1912

Page from the Harvard *Quinquennial Catalogue*

Another phase of the project might focus on digitizing course catalogs and back issues of the *Harvard Alumni Bulletin*, *Harvard Magazine*, and the *Radcliffe Quarterly* to complement current online versions. We also hope to add unpublished narratives of life at Harvard and Radcliffe, such as the diary of John Langdon Sibley, Harvard librarian, who kept a detailed record of daily life in Cambridge from 1846 to 1882.

As we add content to the Reference Shelf, we also anticipate the need for two kinds of enhanced functionality:

- cross-resource searching, allowing researchers to search simultaneously in text delivered via the web and text delivered by the Page Delivery System (such as the ability to search for information about a University building in *The Harvard Book* and the *Cambridge Buildings and Architects* architectural survey)
- delivery of structured documents, such as diaries, as text rather than as page images (as in the case of John Langdon Sibley's diary, which has been transcribed from the original handwritten manuscript volumes)

The addition of new content and new functionality will depend on the availability of funding. The Harvard and Radcliffe Archives may be able to find money to digitize single-volume reference publications such as the *Quinquennial Catalog* and *Radcliffe Directory*, or small batches of manuscripts such as student essays, but long runs of serial publications, such as the Harvard course catalogs, are beyond the reaches of our annual operations budgets, as are complex documents that will require extensive analysis and mark-up. An ongoing digital library program to fund reformatting projects and to provide technical expertise would help to expand the Reference Shelf.

13. Additional information

Project statistics

Project developers: Harvard University Archives & Radcliffe Archives

Source of funding: Harvard University Library Digital Initiative

Project duration: June 1999 - January 2003

Project participants: 22 people in 5 library departments, 1 independent content provider, 3 vendors

Storage, access, and delivery systems in use:

- DRS (Digital Repository)
System in which digital versions of the annual reports, narrative histories and founding documents are stored
- NRS (Name Resolution Service)
- PDS (Page Delivery System)
System in which the electronic reports, narrative histories, and founding documents are browsed
- FTS (Full-text Search Service)
System behind the PDS search interface and the HROHRS web search interface
- HOLLIS (Harvard OnLine Library Information System)
Harvard's on-line library catalog in which bibliographic records for most resources on the Reference Shelf, as well as the Reference Shelf itself, include direct links to the resources
- WWW
Harvard/Radcliffe Online Historical Reference Shelf web site provides links to all electronic resources

Total number of resources: 16

- 4 series of annual reports
- 2 narrative histories
- 1 link to facts and figures
- 1 founding document
- 2 links to music resources
- 5 links to serial publications
- 1 architecture resource

Total number of bibliographic records for Reference Shelf content: 31

- Harvard Presidents' Reports: 7
- Harvard Treasurers' Reports: 4
- Radcliffe Presidents' Reports: 13
- Radcliffe Treasurers' Reports: 2
- Narrative histories: 2

- Founding documents: 1
- Music resources: 1
- Architecture resource: 1

Total number of published volumes digitized for the HROHRS project: 467

- Harvard presidents' reports – 164 vols.
- Harvard treasurers' reports – 163 vols.
- Radcliffe presidents' reports – 85 vols.
- Radcliffe treasurers' reports – 51 vols.
- Narrative histories – 3 vols.
- Founding documents – 1 vol.

Total number of digital files: 224,988

Annual reports = 221,794 files

- 103,871 bi-tonal TIFF files (images of each page of text)
- 103,871 ASCII files (OCR versions of text pages)
- 406 color TIFF files (1 map, 374 charts, 31 color covers)
- 13,646 structural metadata files (one for each volume and section of the reports)

Narrative histories = 2,719 files

- 916 bi-tonal TIFF files (images of text pages in *The Harvard Book*)
- 160 color TIFF files (images of all pages in *A Century to Celebrate*)
- 1,076 ASCII files (OCR versions of text pages in *The Harvard Book* and all pages in *A Century to Celebrate*)
- 115 grayscale TIFF files (continuous-tone photographic illustrations in *The Harvard Book*)
- 275 JPEG files (delivery versions of continuous-tone illustrations in *The Harvard Book* and all pages in *A Century to Celebrate*)
- 177 structural metadata files

Founding documents = 93 files

- 38 bi-tonal TIFF files
- 38 ASCII files
- 17 structural metadata files

Architectural survey = 382 files

- 31 HTML text files
- 351 color JPEG files (color photographs)

Total file size: 16.6 gigabytes

- Harvard reports: 12.83 gigabytes
(images – 12.46 gigabytes; text – 0.37 gigabyte)
- Radcliffe reports: 2.13 gigabytes
(images – 2.10 gigabytes; text – 0.34 gigabyte)

- Narrative histories: 1.64 gigabytes
(Harvard Book – .6 gigabytes; Century to Celebrate – 1.04 gigabytes)
- Founding documents: .003 gigabyte
- Architectural survey: .1 gigabyte
(images – .09 gigabyte; text – .01 gigabyte)

Project participants

Department	Name/title	Project role
Harvard University Archives	Kathryn Bowers, Processing Archivist	Project management
	Patrice Donoghue, Collection Management Archivist	Hard-copy inventory control
	Sarah Gray, Project Assistant	Mark-up and file management
	Kara Lewis, Project Assistant	Document preparation
	Robin McElheny, Associate University Archivist for Programs	Project management
<hr/>		
Radcliffe Archives	Joanne Donovan, Audiovisual Assistant	Web site planning
	Glynn Edwards, Manuscripts Processor	Web site design
	Jane Knowles, Acting Director, Schlesinger Library	Project management
	Katherine Kraft, Acting Radcliffe Archivist	Project management
<hr/>		
HUL Library Digital Initiative	Stephen Abrahms, Digital Library Program Manager	Full-text Search Service development
	Steve Chapman, Preservation Librarian for Digital Initiatives	Project planning, work flow development and OCR management
	James Coleman, Digital Library Projects Manager	DRS development and project specifications
	Wendy Gogel, Projects Liaison	Project management
	Julian Marinus, Programmer/Analyst	DRS development and file deposit
	Clare McInerny, Programmer/Analyst	PDS development and metadata conversion
	Catherine Palmer, Technical Assistant	OCR quality assurance
	Mackenzie Smith, Digital Library Projects Manager	LDI systems development and functional specifications
<hr/>		
HCL Digital Imaging Group	William Comstock, Manager	Scanning work flow management
	Stephanie Mitchell, Photographer	Color scanning and file management
	David Remington, Photographer	Color scanning and file management
	Mingtao Zhao, Computer Assistant	Bi-tonal scanning and file management

Project participants (cont'd)

HCL Conservation Services	Nancy Schrock, Chief Collections Conservator	Document preparation
Independent content provider	Christopher Hail, author, <i>Cambridge Buildings and Architects</i>	Preliminary file mark-up
Vendors	Brainwashed	File mark-up and web page design
	Surveytools Corp.	Development and hosting of evaluation questionnaire
	University of Michigan Digital Library Production Service	OCR conversion of page image files



Some of the project participants (left to right): Patrice Donoghue, William Comstock, Mingtao Zhao, David Remington, Robin McElheny, Stephen Chapman, Kathryn Bowers

LDI project budget

Expense category	Cost
Project assistants - salaries + fringes	
Scanning	\$19,965
OCR file conversion	13,502
HTML file mark-up + web design	1,074
User evaluation design + hosting	1,990
Total expenses	\$40,917

Publicity timeline

March 25, 1999	Announcement in <i>Harvard University Gazette</i> , Vol. XCIV, No. 23
April 13, 1999	Presentation at LDI Brownbag Luncheon, Session II: LDI Grants
November 1999	Announcement in <i>Harvard University Library Notes</i> No. 1289
November 14, 2000	Presentation at Harvard/Radcliffe Archivists Group fall meeting
September 1, 2001	“The Case of the Reappearing Annual Report,” paper delivered at annual meeting of the Society of American Archivists, Washington D.C.
Fall 2001	“Radcliffe’s History: Read All About It” in <i>News from the Schlesinger Library</i>
November 2001	“Now online: The Harvard-Radcliffe Online Historical Reference Shelf” in <i>Harvard University Library Notes</i> No. 1304
March 19, 2002	Presentation at <i>Service to Remote Users</i> forum sponsored by the HUL Professional Development Committee
September 2002	“H-R History Online” in <i>Harvard Magazine</i> , September-October 2002, Vol. 105, No. 1
November 2002	Included in showcase of web initiatives at <i>Harvard, Internet & Society</i> conference sponsored by The Berkman Center for Internet & Society, Harvard Law School