Library and Information Services in Astronomy IV
July 2-5, 2002, Prague, Czech Republic
B. Corbin, E. Bryson, and M. Wolf (eds)

Preserving Observatory Publications: Microfilming, Scanning...What's Next?

Donna J. Coletti

John G. Wolbach Library & Information Resource Center,
Harvard-Smithsonian Center for Astrophysics
60 Garden Street Cambridge, MA 02138, USA
dcoletti@cfa.harvard.edu

Abstract. Since 1996, the John G. Wolbach Library & Information Resource Center at the Harvard-Smithsonian Center for Astrophysics has participated in a preservation project, funded by the U.S. National Endowment for the Humanities and carried out at Harvard University's Weissman Preservation Center, to preserve the history of science. More than 2,000 volumes of Wolbach Library's 3,000 volume collection of historical observatory publications from around the world have already been preserved on microfilm. A follow-up project to convert the collection to digital format was begun in the year 2000.

Meanwhile, Harvard University unveiled its Digital Repository Service (DRS) offering state-of-the-art storage and retrieval of digital collections. DRS goes further than our previous projects by offering full-text searching, page turning capability, color plates, strict metadata requirements, persistent links using universal resource names, reformatting as necessary, and perpetual storage. Harvard also offers the service and guidance of preservation experts from its state-of-the-art Imaging Service and Preservation Department. In anticipation of the LISA IV meeting in Prague, Wolbach Library issued a challenge to Harvard, "Show us what you can do with a brittle volume from the Observatory Publication collection." Harvard accepted. The result is included in this paper and links are provided to allow the reader closer scrutiny of the final product.

1. Introduction

The John G. Wolbach Library & Information Resource Center is home to the combined collections of the libraries of the Harvard College Observatory and the Smithsonian Astrophysical Observatory. In addition to providing information services to the Harvard-Smithsonian Center for Astrophysics (CfA) community of over 900 employees, the Library is Harvard University's Astronomy Library serving undergraduate, graduate, and post doctoral researchers and faculty. Among the Library's holdings is one of the world's most complete collections of historical observatory publications. These 3,000 volumes, which are in frequent use, contain seminal articles describing research and results of work conducted at observatories throughout the world during the 19th and 20th cen-

turies. More than 90% of the volumes that comprise this collection were printed on acidic paper; 35% are so brittle that their use has had to be restricted or completely curtailed.

In response to preservation concerns, a series of projects has been initiated, not only to preserve and digitize the content, but to also broaden and optimize access.

2. Project goals

2.1. Short-term, long-term, and ultimate goals

In 1996 the situation was grim. Some of the volumes in the collection were so brittle that content had already been lost, and more was being destroyed with each subsequent use. The short-term goal was to save the content without destroying the original volumes. This was accomplished by preservation on microfilm, the only accepted standard for preservation and the only format supported by the large U.S. agencies at the time the project was funded by the U.S. National Endowment for the Humanities (NEH) in 1996.

The long-term goal was to make digital images of the collection available in the NASA Astrophysics Data System (ADS). Volumes on microfilm are scanned and images of the pages are available free to anyone with a web browser on a computer connected to the Internet. But problems using the image files have caused re-evaluation of this approach.

Now, the ultimate goal is to go beyond digital page images to create virtual volumes that can, at least, be used in ways similar to their print counterparts and can, at best, add value when the volumes are reborn digitally. Searchable tables of content, searchable full-text, full-color high-quality illustrations, and foldouts are necessary. Persistent links, universal resource names, reformatting of the digital volumes as necessary, and perpetual storage have become basic requirements.

3. Project descriptions

3.1. Microfilming

Harvard has been microfilming materials from its collections for over 70 years. In 1995, the Weissman Preservation Center at Harvard University turned its focus to the deteriorating collections in the Harvard science libraries. The Wolbach Librarians proposed to preserve the history of astronomy and were invited to participate. The astronomy proposal was combined with those of three other Harvard science libraries and a larger proposal, titled "Preservation of the History of Science in America," was submitted to the NEH.

Since then, the Wolbach Library has participated in two additional proposals resulting in two additional two-year grants. The final astronomy proposal was submitted on July 1, 2002 and, if funded, will continue the Wolbach Library participation through 2005.

3.2. Benefits of the microfilming project

- Original volumes are not destroyed.

  Preservation Center staff members, trained to handle brittle materials, oversee the project and, as a result, minimal or no damage has been done to the original volumes while they are shipped, processed, cataloged, and filmed. Very few volumes have had to be burst; bindings are left intact whenever possible. The master microfilm will be housed in a vault at Harvard for safekeeping. Original volumes will be transferred to Harvard's state-of-the-art depository to be recalled as necessary.

- Microfilm lasts a long time.

  The life expectancy of the microfilm produced by this project is 500 years.

- Gaps are being filled.

  Our grant requires a title to be complete (i.e. no missing volumes) before it can be submitted for filming. Missing volumes must be replaced; unpublished materials are being identified and noted in the cataloging record. We are grateful to colleagues in libraries throughout the world who have participated by loaning replacement volumes or providing photocopies to fill the gaps in our collection.

- Cataloging is performed by professional librarians.

  Machine-readable cataloging (MARC) records are uploaded into the Online Computer Library Center (OCLC) database and searchable in WorldCat (`http://www.oclc.org/ worldcat/`). The WorldCat database, the most consulted database in higher education, contains over 47 million cataloging records with information on 852,239,000 locations of specific items. Content spans more than 4,000 years and is provided in 400 languages.

3.3. Disadvantages of using materials on microfilm

Very few librarians or patrons are champions of microfilm. Specialized equipment restricts use to the library and is often difficult and slow to use. The film can be easily damaged by improper or careless use; illustrations, plates, and photographs reproduced on microfilm are of low quality with no grayscale or color.

4. Converting microfilm to digital

The original goal of this project was to digitize the collection. However, at the time our project began, digital projects were not being considered or funded because the format was considered unstable. By 1999, scanning technology had matured to the point that a second project, to digitize from the microfilm, could begin. Also, the price per frame to digitize had decreased and, though grant

funding was still not available, Wolbach Library, the U.S. Naval Observatory (USNO) Library, and the ADS were able to contribute enough funds to start the project.

The Wolbach librarians relied heavily on the expertise of Steve Chapman, Preservation Librarian for Digital Initiatives at the Harvard University Library (HUL) for guidance. Steve's knowledge of libraries and digitization benefited the project tremendously. Alberto Accomazzi of the ADS helped write a request for proposals to insure the resulting digital files would be compatible with the ADS.

ADS wanted to include observatory publications and had been seeking donations of hardcopies to burst and scan. Wolbach Library had a nearly intact collection but could not sacrifice it. Digitizing from the film would allow ADS to have the information and Wolbach Library would have a delivery system for the information.

Wolbach Library contracted with Preservation Resources, a Division of OCLC. Preservation Resources delivers TIFF (Tagged Image File Format) files on CD-ROM. The files are checked by Wolbach Library staff for quality before being forwarded for loading into the "Scanned Historical Literature Service" of the ADS.

4.1. Advantages and disadvantages of scanning from microfilm

The advantages of digitization for this project are fairly obvious: the information is available online, world wide access is free, desktop delivery is convenient, and electronic format can be easier and more convenient to use than microfilm. The images of text produced from the microfilm are very high quality.

Scanning from microfilm allows rapid digitization of brittle materials, which, because of their deteriorating condition, could not be incorporated in the regular work flow of the large digital library projects. Working from microfilm allows content of rare materials to be digitized without destroying the original volumes. But there are disadvantages as well.

The digital volumes in the "Scanned Historical Literature Service" of the ADS are difficult to use for a variety of reasons. There are no searchable tables of content or searchable lists of plates. Full-text searching is not available; paging through images to find needed pages from thumbnail images is inconvenient and time-consuming.

Because plates, photos, drawings, and illustrations on the microfilm were reproduced in black and white without grayscale, a significant amount of detail was lost. The quality of the scanned plate could not be better quality than what was on the microfilm, so the plates in ADS provide only a glimpse of the original image (see Figure 1).

Currently, the page numbers listed under the thumbnail images in most of the volumes that were scanned and loaded into the ADS are not the same as the page numbers in the original volumes. The problem can be better appreciated with a real-life example that illustrates the importance of using metadata, i.e. structured and specific descriptions of objects, collections of objects and content in objects that are used to locate objects and content. Librarians have always created and relied on metadata but they call it "cataloging."

4.2. Sample search of observatory publications in the ADS

We received a request from a researcher at a remote site who needed a list of individuals who had made financial contributions to the young Harvard College Observatory. We bragged that, to save time, he could find the information himself using the digital volume in the ADS. We even gave him the volume number, VIII, and the page number, 11, of the information he wanted. Before long he called us back to let us know he could not find what he needed.

So we decided to try the search ourselves. We went into ADS → Browse Library → Scanned Historical Literature → Annals of the Astronomical Observatory of Harvard College → (volume) 008 → page 11. Page 11 turned out to be a blank page. So we began clicking on "next page" moving forward looking for a REAL page number. Finally on about the ninth mouse click we found a page number on the image...page 4, which ADS listed as thumbnail (page) 19. It is important to understand that the page number in ADS is not necessarily the same as the page number in the original volume. Think of the ADS page number as the "image" number. Once we knew that page 4 in the original volume was ADS thumbnail 19, we could plan our attack. If thumbnail 19 was page 4, and we wanted page 11 then we needed to move 7 pages forward (4+7= page 11 so 19+7=image 26). Image 26 turned out to be page 11. This demonstrates how time consuming and frustrating it can be to use a digital volume without adequate metadata. Actually this was a simplistic example because we knew the volume and page number of the needed information, which usually is not the case.

Navigating through thumbnail page images is slow and tedious and users often give up. Researchers try once or twice and then come back to the library. At this point it is faster to use the print version.

What is the problem and what is the solution? The answer to both questions is metadata. The solution is to plan and create lots of metadata, descriptive cataloging to help patrons find information on the web, and metadata tags to allow searchers to navigate through the volume once it has been located. Metadata is time consuming to plan and create and expensive to produce, so often it is left undone. Most searchers learn very quickly that just because something is digitized and on the web there is no assurance that it can be retrieved simply or at all.

5. Project status

As of July 2002, 2,000 volumes (250 titles) have been filmed at Harvard and approximately 1,230 volumes have been digitized by Preservation Resources. A list of titles filmed and scanned is available at Wolbach Library web site cfa-www.harvard.edu/library; click on "Special Projects."

6. Digital Repository Service of the Harvard University Library

As the Wolbach projects were progressing, the Harvard University Library announced its Digital Repository Service (DRS) created to house, deliver and preserve materials born, or reborn, digitally. The DRS is overseen by librarians who

understand the importance of in-depth cataloging (metadata), cross references, and multiple access points. In preparation for the LISA IV Conference, Wolbach Librarians asked the DRS staff to show what could be done with an old volume from the observatory publications collection. DRS staff took volume VIII of the Annals of the Observatory of Harvard College and transformed it into a robust, simple to use, digital volume. The results were presented at LISA IV and are available on the web (`http://nrs.harvard.edu/urn-3:FCOR.WOLBACH.AAO:42725`).

6.1. Benefits of the Digital Repository Service

- Persistent links.

  Every item that resides in the DRS must be fully cataloged and have a persistent link (URN) to insure it will be permanently "findable". The volumes we digitized and added to the DRS have links from the Harvard Online Public Access Catalog (HOLLIS) and from OCLC/WorldCat.

- Plates.

  Staff at the Digital Imaging Lab at HUL decide the best way to treat plates, depending on whether they are drawings, photographs, foldouts, engravings, etc. and depending on whether they are color, black and white or grayscale. The sizes of some plates were manipulated. Plate 21, "A Key to the Moon" by Beer and Madler, was too small to be easily read so an enlarged version was inserted into the digital volume. The difference in black and white and grayscale illustrations in DRS compared to ADS is startling (see Figure 1.). The plates on the left are from the DRS and can be viewed in color on the Internet. Plates on the right are from the ADS and were scanned from the microfilm with no grayscale.

- Full-text searching.

  "Dirty" (uncorrected) optical character recognition was performed on the volume to allow full text searching. A search tool that allows sorting of results by relevance, alphabet, or date is available. Boolean searching is an option as is full text keyword searching.

- Page delivery service.

  Page delivery allows the digital volume to be used in ways similar to the print counterparts. Tables of content can be examined, the book can be viewed page by page, or individual pages can be requested.

Repeating the earlier search for contributors to the early Harvard College Observatory using a combination of traditional and digital library tools is fast and simple. The title can be searched in OCLC/WorldCat. The searcher does not have to know beforehand that the volume resides in a certain section (Scanned Historical Literature Service) of a particular virtual library (the ADS).

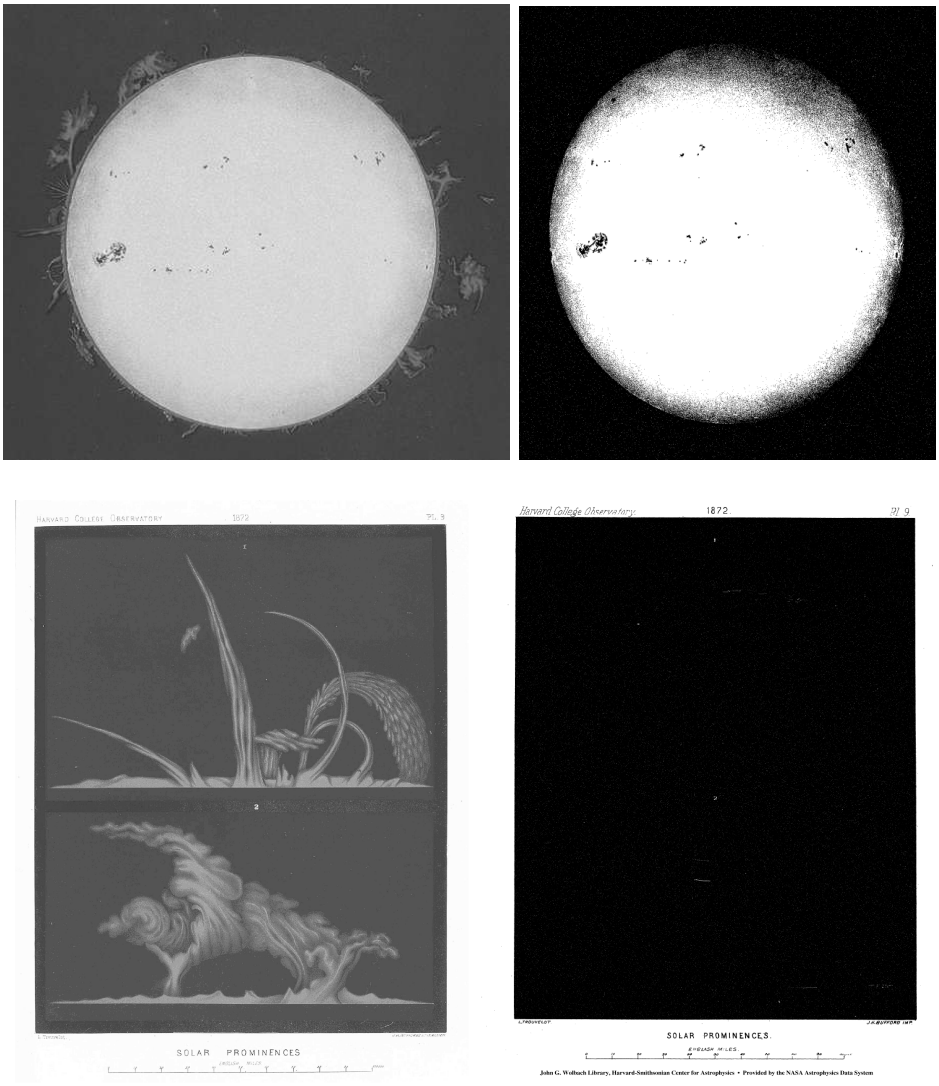

Figure 1. Plates on the Left as viewed in the Digital Repository Service of the Harvard University Library; on the right are the same plates as seen in the "Scanned Historical Literature Service" of the Astrophysics Data System

Because there are multiple access points (OCLC/WorldCat, Harvard's Digital Repository, Wolbach Library's web page, Harvard University's online catalog HOLLIS) the volume is easy to find and maneuver through.

6.2. Searching from OCLC/WorldCat

Sample search from OCLC-WorldCat. As an example, open the "Basic Search" page of WorldCat → type title "Annals of the Astronomical Observatory of Harvard College" → title radio button → Search. You receive a list of titles. Selecting the entry with the computer disk icon and you enter Harvard's DRS with a variety of options for navigating the volume. Tables of content allow traditional use; keyword or boolean full-text searching is available.

7. For more information

Digital Repository Service/Digital Initiatives, Harvard University Library:

preserve.harvard.edu/digital/

John G. Wolbach Library & Information Resource Center:

cfa-www.harvard.edu/library

Digital Repository Service, Annals of the Astronomical Observatory of Harvard College, Volume VIII:

nrs.harvard.edu/urn-3:FCOR.WOLBACH.AAO:42725

NASA Astrophysics Data System:

adswww.harvard.edu

8. Acknowledgments

This project was funded in part by grants PA-22937-97, PA-23289-99, and PA-23711-01 to the Harvard University Library from the U.S. National Endowment for the Humanities. And by funds from the USNO Library and the Astrophysics Data System.

Thanks to Will Graves, Wolbach Library, Harvard-Smithsonian Center for Astrophysics, Brenda Corbin, USNO Library, Staff of the Harvard University Library Preservation and Imaging Services, and Weissman Preservation Center. Particularly Steve Chapman, Bill Comstock, and Maggie Hale. Thanks also to the staff of the Astrophysics Data System, particularly Alberto Accomazzi.

Additional thanks to librarians throughout the world who have contributed library materials to this project.