Digitization of Historical Astronomical Literature

Brenda G. Corbin [a] and Donna J. Coletti [b]

[a] *U.S. Naval Observatory Library, Washington, DC, USA*

[b] *John G.Wolbach Library, Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA*

Abstract

In the late 19th and early 20th centuries, much of the important research in astronomy generally appeared in observatory publications which were distributed by observatories throughout the world. Much of the data is still in demand (observations of variable stars, cometary observations, solar phenomena, etc.), but complete sets of these publications are held by few libraries. The information is often not easily accessible because significant portions of the material are brittle and in danger of being lost. The U.S. Naval Observatory (USNO) Library, and the Harvard College Observatory's Wolbach Library (HCO) hold almost complete sets of these series in their collections. Digitization of these historical collections will preserve them for posterity, improve access, reduce retrieval time, and reduce the risk of the information disappearing through disintegration, mutilation, or theft. A pilot project is currently in progress.

1 Importance of Digitizing Older Materials

1.1 Preservation

A large part of the important research in astronomy during the late 19th and early 20th centuries was reported in great detail in exhaustive articles in the main publication series of observatories throughout the world. Although journals were also important during this period, the reports of these research projects containing descriptions of instruments, detailed tables of findings and explanations of the course of the research were too long to be published in journals. A significant amount of this data is still in frequent use today, e.g. observations of variable stars, star catalogs, cometary observations, solar data, etc. Requests for this information are regularly received by the libraries holding these collections. However, many of the volumes are quite brittle and literally

Preprint submitted to Elsevier Science *29 August 1995*

falling apart. In essence, some of this important information is in danger of being lost.

Why are some of these volumes in such brittle condition? Many factors contribute to this, such as storage in poor climatic conditions and heavy use. The major problem however, is the acidity of the paper on which the volumes were printed. The increasing demand for paper in the late 19th century resulted in mass production techniques which began using wood pulp rather than the 100% rag fiber of earlier days. The paper producers also began using an alum-rosin compound which allowed the paper to receive the printer's ink more efficiently. The alum compound eventually breaks down into acids which eat the chemical bonds holding the paper together, and the paper destroys itself. This process is quite visible in many of the books published since 1850 in which the pages are yellow and brittle, and are easily broken when bent.

1.2 Access to Observatory Publications

As newer astronomical institutions have been established, it has been difficult for their libraries to retroactively obtain copies of observatory publications, and thus astronomers have not had ready access to data from these older series. As a result, libraries that have copies of these materials regularly receive interlibrary loan and other requests for photocopies. In some cases, photocopying would further damage the already brittle volumes, and of course these books cannot be loaned. Security is also a concern, especially in large university libraries. Theft and mutilation of the volumes have occurred. During an incident which occurred in 1993 at the Harvard College Observatory Library, 35 volumes were mutilated by a collector who used the plates in his artwork. Digitization of the observatory publications, including the plates, would not only preserve the information, but also make it easily accessible to the entire astronomical community.

2 History

2.1 Earlier Proposals for Archival Preservation

In 1979, it was suggested (Corbin, 1979) that the Physics-Astronomy-Mathematics Division (PAM) of the Special Libraries Association (SLA) investigate the possibility of archival microfilming for the entire set of observatory publications. Proposals for such a project to the National Science Foundation and other institutions were considered. After some initial inquiry, it was determined that funds were not available without the firm support of a university.

Another approach was taken in 1983 (Vrba, 1983) when the Historical Astronomy Division (HAD) of the American Astronomical Society (AAS) was asked to join in supporting this preservation project. Astronomical institutions throughout the U.S. were asked to pledge funds to support the archival microfilming of the observatory publications. Several institutions agreed to contribute funds, but some noted the recent introduction of optical disk technology, and suggested this might be a better solution to the preservation problem. The decision was made to wait for further development of the optical disk technology rather than pursuing archival microfilming.

A complete overview of the optical disk technology project was given in 1987 (Corbin, 1988) with a proposal that it be undertaken in cooperation with the Smithsonian's National Air and Space Museum. However, technical difficulties and lack of funds once again stalled the efforts on this project.

In early 1994, the High Energy Astrophysics Division (HEAD) at the Center for Astrophysics (CFA) included the digitization and preservation of observatory publications as part of a much larger proposal to the National Science Foundation (NSF) for the "National Astronomy Research Library" (NARL) (Murray, 1994). After the proposal was submitted the NSF decided that the Smithsonian Institution, which supports the CFA, was not an educational institution and the proposal was not funded. However the Astrophysics Data System, a part of the HEAD, continued to support the idea of this project.

2.2 Current Digitization Technology

It is clear that digitization technology is currently accepted as an archival preservation medium (Kenney, 1995) as demonstrated by projects at Cornell, the University of Michigan and other large universities worldwide. Once scanned, in addition to being electronically accessible, the digitized files can be used to produce a variety of formats such as CD-ROM, microfilm, and printed copies.

3 Initial Efforts

3.1 Two Volumes Digitized

The Smithsonian Astrophysical Observatory Annals, volume 1 (1900) was removed from the binding, and the pages were form fed and scanned at a resolution of 400 dots per inch (dpi). Mechanical feeding is economical, but may not be gentle enough for older, brittle materials to survive the process.

The Harvard Observatory Annals, volume VIII (1876) was scanned at 600 dpi by the Harvard Preservation Department on a state of the art "Documents on Demand" (DOD) system. This Xerox system uses a PC with a gigabyte hard drive, a Xerox scanner and a Docutech printer, all run under Xerox software. The volume was removed from the binding, and each page was placed by hand on the scanning surface to make sure the edges were properly aligned (registered). The plates, many in color, were also scanned. To achieve a higher resolution of the plates, some were also photographed, and the slides were digitized. This scanning of the slides resulted in better reproduction than the direct scanning of the plates.

For both volumes the table of contents, lists of plates and indices were keyed in manually. This was done because optical character recognition (OCR) programs would have been more time consuming and unreliable than manually keying the data. These files were then linked to the bitmapped pages to allow for hypertext searching.

Both volumes are now available on the World Wide Web (WWW) through the Astrophysics Data System (ADS). These publications may be seen at the following address: http://adsbit.harvard.edu/books/.

4 Concerns and Obstacles

4.1 Insufficient Finding Aids

Locating specific information in observatory publications can be difficult as it was not the custom to include abstracts and indices, and standardized keywords were nonexistent. It would be advantageous to include a searchable abstract from information provided in the introduction of each publication. The abstracts could be keyed-in or run through an optical character recognition program. Preparation of these abstracts would be quite time consuming however, and the persons preparing them would have to be familiar with the field of astronomy.

Searching of entire volumes by performing optical character recognition on the bitmaps is not feasible for a variety of reasons. The process continues to be error prone and is even more so on older type fonts and discolored paper.

4.2 Search Speed

Although many view the WWW as a panacea for storing and viewing full text documents, retrieval time can sometimes be slow as each page constitutes a separate file and must be downloaded individually. It is frustrating for the user who might wish to scroll through a chapter much as one pages through a book or scrolls through microfilm. The speed of the network is based on many things such as the computer used and its speed and also the limitations of the browser (Netscape vs. Mosaic). There is no current solution to this problem, but it will most likely be solved as the technology evolves. At present, by far the most preferable way of obtaining information from these volumes is still to pull a copy from the shelf and turn the pages.

5 Plans for the Near Future

Funding sources are being sought for the pilot project to digitize the complete set of the Harvard College Observatory Annals, a total of 120 volumes with approximately 33,000 pages. When funding is identified, a proposal will be written jointly by the Harvard University Preservation Department and the Astrophysics Data System. The Annals will be made available on the WWW through the Astrophysics Data System. The ADS will also offer continuing support as they maintain and refresh the digital files.

6 Long Term Plans

6.1 Preserve Observatory Publications Worldwide

The scope of the long term project is much larger – to archivally preserve the complete publications of all observatories worldwide. This would mean scanning approximately 3000 volumes. Other institutions are planning similar or even much larger projects. Cornell University hopes to scan 100,000 volumes published between 1860 and 1960 in their "Making of America" project (Kenney, 1995).

6.2 Copyright Issues

Since the long term project will encompass many series other than the Harvard College Annals, the question of copyright must be addressed. Most observatory publications are no longer covered by copyright restrictions, but one must be absolutely certain of this before mounting this digitized information on the World Wide Web. Contact will be made with the Copyright Office of the Library of Congress to check the U.S. publications, and with the proper offices within other countries to obtain similar information.

6.3 Beyond the Pilot Project

Reformatting the 120 volumes of the Harvard Annals is a pilot project. If it is successful, funding will be sought for the larger project to reformat all of the existing observatory publications. The almost complete collections held by the Harvard College Observatory and the U.S. Naval Observatory will initially be used. For publications not available at either USNO or Harvard, colleagues at other institutions will be asked to loan these volumes with the understanding that the pages will be removed from their bindings.

The administrations of both the Harvard University Library and the Astrophysics Data System have demonstrated a strong interest in both the pilot and long term projects to preserve and provide access to past research in astronomy by digitizing the observatory publications. If this project does succeed, it will guarantee that the important astronomical research appearing in these volumes will be available in the 21st century and beyond.

Acknowledgement

The authors would like to thank the following people for offering information and advice concerning this project: Astrophysics Data System: Stephen S. Murray, Guenther Eichhorn, Carolyn Stern-Grant and Alberto Accomazzi. Harvard University: Lisa Biblo and Robin McElheny of the Preservation Department; Chris Kupec, Widener Photographic Services; Dale Flecker and Katharina Klemperer, Office for Information Systems. Cornell University Library: Anne Kenney and Steve Worona. Smithsonian Air and Space Museum: David H. DeVorkin.

References

[1] Corbin, B., 1979 Physics-Astronomy-Mathematics Bulletin, **7**, 21.

[2] Corbin, B., 1988, in Debarbat S. et al, eds, IAU Symposium 133, Mapping the Sky, Kluwer Academic Publishers, Dordrecht, 1988, 115.

[3] Kenney, A., 1995, private communication and data from website http://www/englib.cornell.edu/ccr/digital-lib.html#Big

[4] Murray, S., Eichhorn, G., and Kurtz, M., 1994, Proposal to National Science Foundation for National Astronomy Research Library P3153-1-94. Smithsonian Astrophysical Observatory.

[5] Vrba, F., Corbin, B. and DeVorkin, D, 1983, Bulletin of the American Astronomical Society, **15**, 996.