LIBRARY AND INFORMATION SERVICES IN ASTRONOMY VI
ASP Conference Series, Vol. 433
E. Isaksson, J. Lagerstrom, A. Holl, and N. Bawdekar, eds.

Small Data Archives and Libraries

A. Holl

Konkoly Observatory, H-1525 P.O. Box 67, Budapest, Hungary

Abstract. Preservation is important for documenting original observations, and existing data are an important resource which can be re-used. Observatories should set up electronic data archives and formulate archiving policies. VO (Virtual Observatory) compliance is desirable; even if this is not possible, at least some VO ideas should be applied. Data archives should be visible and their data kept on-line. Metadata should be plentiful, and as standard as possible, just like file formats. Literature and data should be cross-linked. Libraries can play an important role in this process. In this paper, we discuss data archiving for small projects and observatories. We review the questions of digitization, cost factors, manpower, organizational structure and more.

This is a paper about data and archiving, about an electronic archive which is much thought about but not yet built. I begin with a story.

On the night of February 2, 1826, after a month of unusually favorable weather — meaning tiring observational work — Paulus Tittel (I use the name of the director of Gellérthegy Observatory, a Catholic priest, in Latin form) fell asleep at his desk. His candle burned down, and the papers burst into flames. He woke up and managed to extinguish the fire, but several books and stacks of paper were reduced to cinders and ash. Among them was a fully written paper on the fourth comet of 1825, to be sent to the *Astronomische Nachrichten* in Altona, together with all of the observational material recorded on this object. As we can learn from the description of the event by his pupil at that time, Ferenc Albert, they had no time to make a copy of the original observations, so their work was lost, with no chance for reproduction (Vargha & Kanyó 1998).

As a marginal note, a large part of the old book collection of Konkoly Observatory originates from Tittel's collection at the Gellérthegy Observatory. Most of the books have survived the conflagration, the siege of Buda in 1849 — thanks to Albert — and many perils since then.

Why do we need to archive observations? Naturally, we want to use them in publications. Those publications need to be verifiable. The editors and peer-reviewers, as well as the readers, must be able to go back to the original observations to check them. The eminent science journal *Nature* requires that supporting data be made available. We must recognize that, although most scientific papers lose their importance after some time, there is no time limit for this requirement — data should be kept indeterminately. Published data are by no means exhausted data. Observations could have secondary, serendipitous use. Or, to put it simply, information overlooked could later lead to a discovery, as in the case of a supernova patrol plate taken in 1968 at the Piszkéstető Station of Konkoly Observatory in which a supernova was found. Later, a new one was revealed by a different astronomer.

Observing logbooks, glass plates — these media might last for centuries if the paper is acid free, if the development of the plate was correct, if we manage to keep them safe from fire, damp, mishandling, dust etc. But the greater perils lie elsewhere: we can not maintain the equipment and expertise necessary for measuring the plates. Moreover, non-digital data can be difficult to access.

Digitization could be the answer. But are digital data preservable, are they kept permanently accessible? The dismal fact is that digital data are often less accessible and preservable than the old non-digital material.

Digital data might get lost because they are physically damaged. Ink might fade during the centuries, but the information storing capability of magnetic and even optical media is much shorter. The lifetime of a technological device could be even shorter; we might have the data intact on a tape cartridge only to find ourselves without a working tape drive. The third factor is data format: the data might be readable bit by bit, but without the proper software, still unintelligible to us.

My experience shows that often the reasons for unavailability or loss of digital data lie not in the technology, but in ourselves. The data is there, on a hard drive in a drawer, on a DVD mislaid somewhere. Observatories might have well-functioning libraries and plate vaults, but at the same time they lack a proper electronic archive. Data might even be on the web — but in peril. Data might still be lost, or might linger on the web unmaintained. Often digital data lack unique identifiers.

Large institutions and large projects have excellent archive facilities, while the situation is often more problematic at small observatories. What could be done at such places? Here is a to-do list I have compiled for establishing the electronic archive of Konkoly Observatory.

Electronic Archive as Part of the Organizational Structure

The electronic archive should be set up as an integral part of the observatory's organizational structure. A person in charge should be appointed and staff allocated. In these times, it is hard to create new jobs, so the head of the IT department or library could be entrusted with this task, and existing IT/library staff could be used for the operation — electronic archive responsibilities should become part of their job descriptions.

Budgetary needs should be estimated and the observatory budget readjusted accordingly. It does not necessarily mean a budget increase, we should note. For digitization, existing photography and/or microfilm units could be converted. Archive status should be reported in the annual report.

Regulation

The Director should regulate what should be archived, the rights of the original observer, and the length of the proprietary period. The ingest and operation should be regulated — librarians and IT specialists should create the rules of operation.

Hardware

The physical archive should be part of the IT infrastructure. Nowadays spinning disks provide the best media for storage as they are the cheapest and easiest to access and

replicate. (A comparison of the DVD and 1 TB HDD prices on a vendor website — at the time of the preparation of this article — shows that HDDs are half as expensive for unit storage capacity. Linear tape cartridges are cheaper.) My suggestion is to use an expandable storage appliance. Migration of the whole data content of the electronic archive should be possible with a simple `cp` command. Storage costs are manageable until the disk drive capacities for a given price grow exponentially.

What to Archive

In general, the research process could be viewed from the data perspective as depicted in Fig. 1.

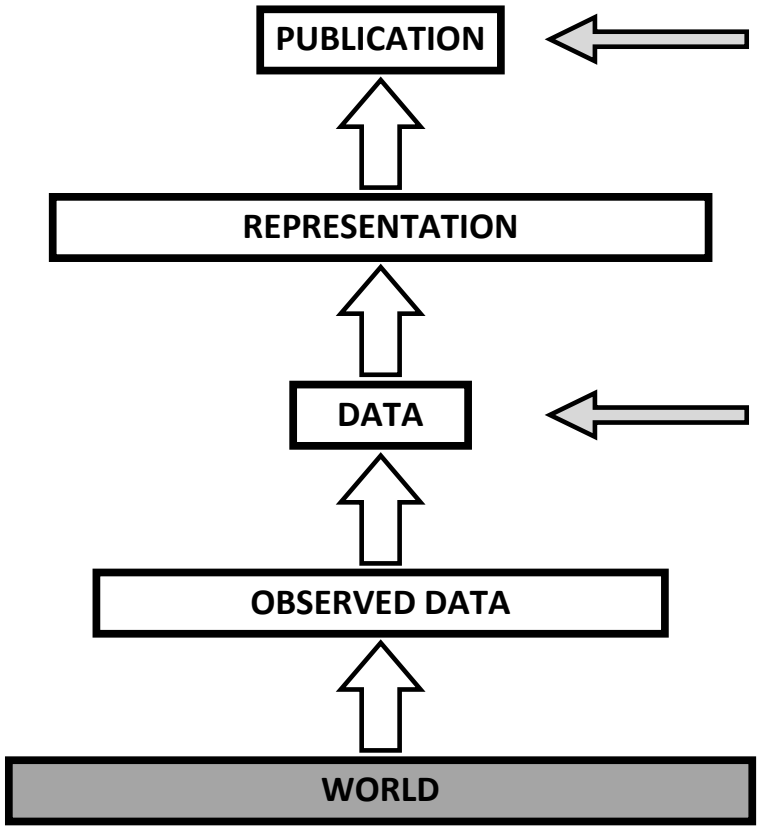

Figure 1. There are certain phases of research when archiving is easier. The box at the bottom represents the object or phenomena to be observed. On top of it are the raw observational data: they contain only selected information, plus noise, instrumental signatures. The next box depicts the reduced, cleaned data — here everything important is present, and this is the appropriate stage to archive the data. In the next phase of research this data is visualized and combined with data from catalogs and the literature and entered into complex databases. No new information is observed here, only the existing information is being combined, increasing the volume. The last phase is the publication — the information is concise, offering another opportunity to archive.

The important point here is that there are phases in the research process when all the important data could be captured and archived, at a minimal cost. Complex data structures could improve functionality, but are hard to maintain in the long term. These could be always rebuilt from the basic data set.

Our solution will be to set up a two-tier archive: one part will store the raw data coming from the telescopes. The camera pipeline will store everything in two copies: one copy will go to the archive, the other to the place chosen by the observer.

The other part of the archive will contain clean, reduced data, processed by the observer. This part will hold all additional types of data (e.g. observatory publications, documentation, digitized plates etc.), and will be publicly visible. It is an open question — it depends on our capabilities to expand the storage — whether the raw data of the material deposited in the public part would be deleted from the raw archive section or not.

Data Formats

FITS is an old and proven standard. The CCD images and scanned plates will be kept in FITS. For the documents, we will use PDF, or rather PDF/A as much as possible. The third choice is the simplest: plain ASCII text. Limiting the accepted data formats might cause inconvenience for the depositing researchers, but helps to keep the long-term costs and risks lower.

Metadata

Pipelines should be fixed to provide as complete metadata as possible. We intend to employ members of the IT/library staff for the maintenance of the FITS headers. The way metadata are kept within the file is a strength of the FITS format. For documents — observatory publications and others — we prefer to put the most important metadata into the text of the document — author's name, title, date, publication name, etc., should be visible on the printed document. Meanwhile, it should be present in the PDF metadata too.

We intend to keep our archive robust — most metainformation should be reconstructible from a heap of files. There will be databases facilitating search, but their contents should be regenerated from the files themselves.

Continuity — a Virtue

Librarians know well the benefits of publishing the LISA proceedings in a series. I think that "sporadic" articles have less impact than those published in a well-known journal. My suggestion is to keep the observatory's publication series running, and publish every important piece of information there. Even if it is not printed any longer, the framework of the series will give strength to the published information. Each published article should get a unique identifier known outside the observatory: a bibcode or (preferably and) a DOI.

Cross-linking

Datasets and data files need to have unique identifiers assigned (Eichhorn et al. 2007). A local naming system should be devised, and a name resolution service for the local data identifiers (PrivateId) should be set up. Research papers should contain unique dataset identifiers. If it is impractical in the body of the article, they should be placed in supplementary files, which could be published only in the electronic version (Holl, Kalaglarsky, Tsvetkov et al. 2006). Persistent identifiers of relevant information should be inserted in the FITS headers of the data files, like telescope and detector documentation published in some of the observatory publication series, or research papers based on the given dataset. Obviously, bibcodes (DOIs) of the latter could only be inserted *a posteriori*. This requires manpower: the work of a librarian.

Costs and Manpower

Small observatories probably can not allocate substantial budgets to archiving. Long-term archiving costs should be kept minimal. Raw data archiving should be done automatically and built in to the data acquisition pipeline. Science-ready data should be archived by the person or team who acquired it, after the publication of the primary paper(s), or before the proprietary period ends. In small observatories we could not expect to have complex reduction pipelines maintained centrally. Data should be reduced on a best-effort basis by the observer. Archival costs should be budgeted within the framework of the given research project and data should be archived in a standard format, requiring minimal migration costs for the long term.

The maintenance of the storage will be done by the IT staff. Hardware migration, as we already mentioned, should be simple. Using standard formats, like FITS or PDF/A, no format migration needs are foreseen for a long time. Metadata maintenance should be done by the IT/library staff.

Visibility

Visibility is beneficial for maintenance. Errors (either present or creeping in) are detected better if the data are used more. VO techniques could provide good visibility, and out-of-the-box free software is available to set up a VO archive. Textual documents should be published in established series and reported to ADS. Data identifiers should be referred to in papers. A possible way to expose documents is to set up an institutional repository — there are excellent free software tools for this purpose as well.

Incentives, Drivers, Benefits

What are the incentives for researchers to deposit their data? First of all, depositing should be mandated by the director. (Or perhaps by publisher, or funder. Funders should require archiving plans and accommodate archiving costs.) Annual reports should refer to publications and data. The latter should be freely available, either in external archives, or in the IR or institutional archive. Archive policies should safeguard the researcher's and observatory's interests with a suitable embargo policy. The archive could take the burden of storage off the individual researcher or the group, at

least to some degree. If there is no money for digitization, digitization/archiving should at least be done on demand. If depositing could be automated, the archive could grow faster. It is easy with the raw data. On the publication side, protocols like SWORD[1] could help. Provenance of the data should be maintained. If the observers and the institute are acknowledged and papers describing the observations cited, researchers' willingness to deposit will increase.

Maintenance and Backup

The electronic archive needs continuous care and maintenance, although that of low intensity. Mechanisms well-known in the management of server computers and services could monitor the operation of the storage and data services. Hardware migration should be performed if storage technology changes, and migration of the registered data formats present in the archive should be done when necessary — not too often if we choose the formats carefully. Printed documents were distributed in a large number of copies, and that ensured the safety of information. Electronic archives need off-site backups in addition to local ones.

The electronic data archive is one of the points where IT and Library meet. Data and publications are inseparable. Librarians are skilled in working with metadata; moreover, the concepts of long-term preservation are known to them.

Acknowledgments. The author is grateful for the support of the organizers and the FOL (Friends of LISA).

References

Eichhorn, G., Accomazzi, A., Grant, C.S. et al. 2007, in ASP Conf. Ser. Vol. 377, Library and Information Services in Astronomy V, ed. S. Ricketts, C. Birdie & E. Isaksson (San Francisco: ASP), 36

Holl, A., Kalaglarsky, D.G., Tsvetkov, M.K. et al. 2006, in Virtual Observatory: Plate Content Digitization, Archive Mining and Image Sequence Processing, ed. M.K. Tsvetkov, V. Golev, F. Murtagh & R. Molina (Sofia: Heron Press), 374

Vargha, D., Kanyó, S. 1998, ...csillagkoronák éjféli barátja: Tittel Pál élete és működése, (Budapest: Akadémiai Kiadó)

[1] http://www.swordapp.org