Small Data Archives & Libraries

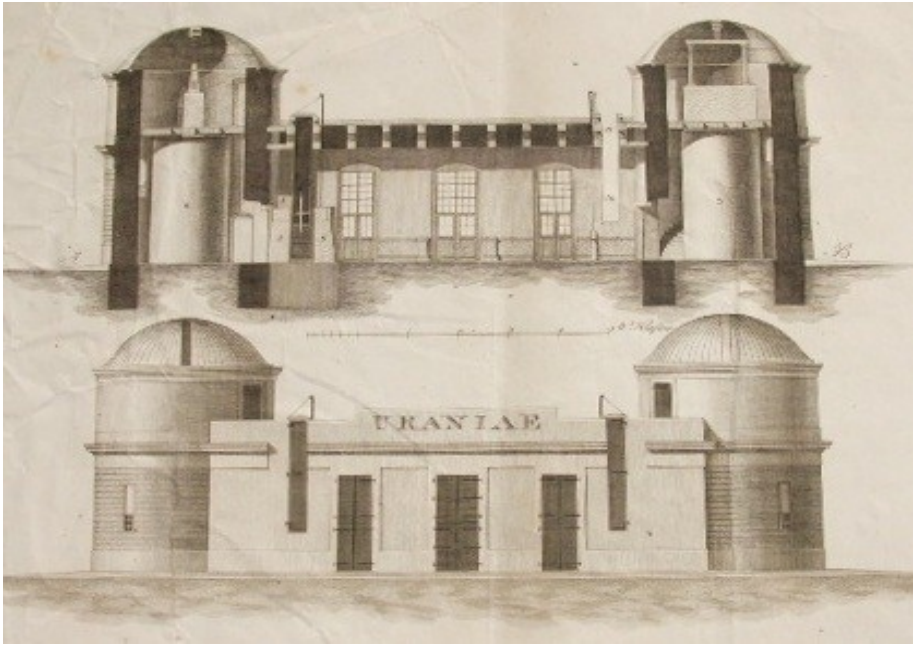
András Holl

Konkoly Observatory, Budapest, Hungary

Library and Information Services in Astronomy (LISA VI) - IUCAA & NCRA - Pune, India, 2010

How can we ensure that small-project data will survive? First we have to ask the question: why preserve? Preservation is important for documenting the original observation, and re-use could be possible for non-stationary phenomena. Observatories should set up their electronic data archives and archiving policies. VO compliance is desirable, but even if it is not possible, some VO ideas could be applied. Data archives should be visible, data kept on-line. Meta-data should be plentiful, and as standard as possible, just like file formats. Literature and data should be cross-linked. Libraries can play a role in this process.

In this paper we discuss data archiving for small projects & observatories. We review the questions of the digitization, costs, manpower, organizational structure and more.



Gellérthegy Observatory, 2nd of February, 1826
observations of P. Tittel consumed by fire (Vargha et al., 1998)

- Why archive?

    Data: use, verification, re-use

    Storing raw data: backup, chance for re-processing

- Old observational material: relatively robust - NOT risk free!

    difficult to measure & access

- Digitization of old material

    digital data are easy to copy error free

- Perils in the digital domain

  media, recording technology & format lifetime

- Present practice in small observatories

  unorganizedness!

  we moved to digital, not learning the new requirements and forgetting the old ones

- Set of recommendations compiled for the future digital archive of Konkoly Observatory

- ***Electronic archive: part of the organizational structure***

  person in charge
  staff (might be existing IT, library & photo/microfilm)
  budget
  yearly report

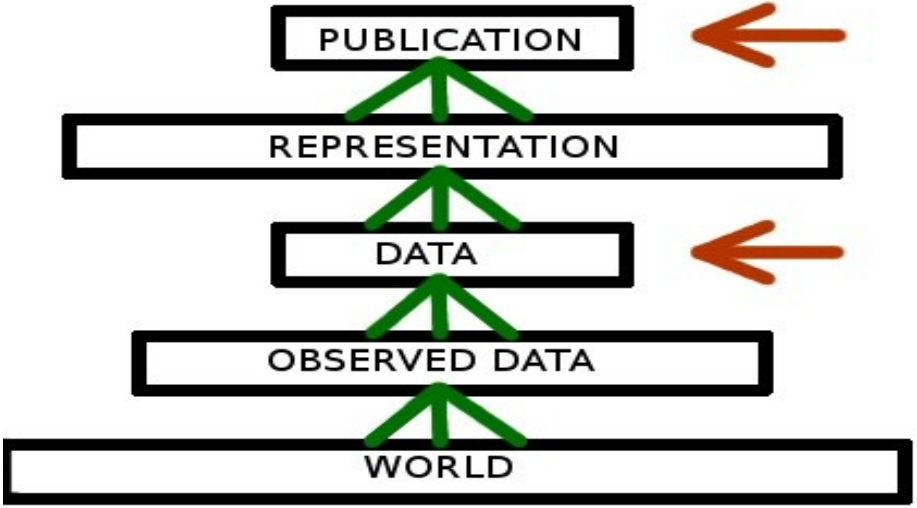
- ***Regulation***

  what to archive?
  proprietary period
  operational rules

- Hardware

part of the IT infrastructure
spinning disks: cost is competitive, good accessibility
easy migration (`cp -R /oldarchive /newarchive`)
expandable storage appliance
manageable till the growth is exponential

- What to archive?



- ...what to archive?

two-tier archive:
raw data & science ready data + documents, etc.
raw data: archiving by data acquisition pipeline
science ready data: by the researcher

- Data formats

FITS, PDF/A, TEXT
limited set of accepted formats

- Metadata

Images - fix the data acquisition pipeline!
Documents: metadata within (like in FITS)
both in human readable way (like in journals)
and in PDF metadata and/or: packaging?
Robust storage: all info in a heap of files
should be able to re-generate everything else

- Continuity

Keep the observatory serials running
everything important should be published there -
- even if only electronically bibcode / DOI !

- Cross-linking

Datasets need unique identifiers (Eichhorn et al, 2007)
Local name resolution service
Bibcodes to FITS headers - task for the librarian?
Research papers should list identifiers of data used
(Holl et al., 2006)

- Costs and manpower

Tight budgets now - need for manageable long-term costs
Raw data: pipeline archiving
Science-ready data: archiving by the scientist / team

- ... costs and manpower

Archival cost should be included in the project budgets
(Hungarian Scientific Research Fund - support?)
Use of standard formats should keep long-term
costs down (FITS, PDF/A)
Storage maintenance: IT staff
Metadata maintenance: library staff?

- Visibility

Good for maintenance
Documents: published & reported to ADS
Free software: VO data publication & OA repository

- Motivation

Mandate (institutional, publisher's, funders)
Reports
Safeguards (embargo)
Digitization on demand / archival on demand
Gratis storage
Automated deposit (pipeline, SWORD for publications)
Provenance carried all the way
Acknowledgements, citations

- Maintenance and backup

continuous, albeit low-level work
storage and service monitoring
timely hardware migration
occasional format migration
off-site backup

*

The electronic archive is where IT & library meet.
Data and publications are inseparable
Librarians are familiar with long-term archival and metadata

References

Eichhorn, G., Accomazzi, A., Grant, C.S. 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)