ASTRONOMY FROM WIDE-FIELD IMAGING

Part Seven:

ARCHIVING AND DATABASES

THE FUTURE OF MEMORY: ARCHIVING ASTRONOMICAL INFORMATION

M.J. KURTZ
Harvard-Smithsonian Center for Astrophysics
Cambridge, MA 02138
U.S.A.

ABSTRACT. The function of astronomers is to capture and create information and send it into the future. The vehicle for this transmission of knowledge is the archive. The digital revolution is quickly rendering our paper and glass archives obsolete; great challenges exist in creating archive systems for digital data. We can only meet these challenges by substantial shared effort.

1. Introduction

The database equivalent of humans remembering their perceptions is the archive, where the basic measurements are stored as raw or reduced measures. This is the foundation for all future work.

Perhaps most digital data taken in the first 20 years of the digital era are irretrievably lost; bits are lying all over the floors of most astronomical centers. Many of the 4-meter plates from KPNO and CTIO are scattered in unknown places.

The mob which murdered Hypatia and burned the library at Alexandria did not understand that they were destroying something valuable to their posterity; WE DO NOT HAVE THAT EXCUSE.

Archives are long-term commitments to the future. They are the foundation for our future research; thus they are of substantial importance. They require both short term capital expenditures in their creation and long term operational funding. It is incumbent upon those responsible for the direction of our major institutions to include archiving among the policy areas which affect the long-term strategic goals of their organizations.

2. Three Current Archives

2.1 THE HARVARD COLLEGE OBSERVATORY / SMITHSONIAN ASTROPHYSICAL OBSERVATORY / CENTER FOR ASTROPHYSICS LIBRARY

John Harvard left half his money and all his books to form Harvard College. The first thing Congress agreed to do with the Smithson bequest which founded the Smithsonian Institution was to build a library. Thus the parent organizations of both the HCO and the SAO were founded, in part, as libraries. The actual observatory libraries were formed from existing subsets of the

H. T. MacGillivray et al. (eds.), Astronomy from Wide-Field Imaging, 331–337.

https://doi.org/10.1017/S0074180900047604 Published online by Cambridge University Press

main library collections after the founding of the HCO (1849) and after the SAO moved from Washington to Cambridge (1959). The CfA Library is an amalgam of the two observatory libraries, formed in 1992.

The HCO library has had major parts of two buildings built for it within the last 50 years; currently it occupies about the same amount of space as one of the CfA's scientific divisions, about the same as a university astronomy department.

The library has had stable and adequate staffing and funding over the years. Items have only been removed from the collection (weeded) after being reviewed by a committee of experts, and very infrequently (theft is a more important factor in controlling growth). The library is still expanding.

The result of more than a century of care is a data archive which actively supports the wide range of current and historical research in astronomy.

2.2 THE HARVARD COLLEGE OBSERVATORY PLATE COLLECTION

There has long been a large collection of astronomical photographic plates at Harvard; Henry Draper, for example, took the first photographic spectrogram at Harvard in 1872. The collection, which became inactive in 1989, contains over 500,000 plates.

In 1931 an especially strong and stable building was built to house the archive, which had outgrown the available space in the Observatory. When finished the plate archive occupied space amounting to about 25% of the total capital cost of the Observatory facilities in Cambridge. In a report to the Harvard overseers the Observatory director, Harlow Shapley, called the plate archive the "most valuable asset of the Harvard College Observatory".

The building was built to contain the next 50 years data, and could (can) be expanded. The archive has had stable funding and staffing through this time (save during World War II). Weeding was done once, in the mid 1950s, by a committee of experts. Not much was removed.

2.3 THE NATIONAL SPACE SCIENCE DATA CENTER

The NSSDC is a major NASA initiative; founded in 1967, it provides the archive for NASA's space science missions. Currently NSSDC holds several Terabytes of digital data.

NSSDC is a major financial obligation for NASA; it has a building and a staff, and is experiencing rapid growth as the space age begins. The data ingest rate is currently growing very rapidly, doubling every year. NSSDC is a very active archive; it handles now about 3,000 requests for data per month.

Weeding is done on an infrequent basis by committees of experts.

As a rapidly expanding archive using the new digital technologies, NSSDC must continuously deal with cost, service level, and technology issues.

2.4 COMMON ELEMENTS

Archives are expensive, long-term commitments; the commitment to the plate stacks was a building worth about a quarter of the total capital cost of the observatory and over 50 years of stable operating budgets. As the problem with weeding shows, once data is archived it is almost never discarded; thus its maintenance becomes essentially a permanent obligation.

What commitment does our new digital data require?

3. The NSSDC Cost Model

As a means of predicting the future budgetary requirements of the NSSDC, Klenk, Green & Treinish (KGT, 1990) have built a computer model to estimate the cost of archiving data from various NASA projects. The KGT model covers a wide range of different data types (NASA has more than just astronomy data) and could be modified to cover other types of archive situations; it is available as an MS-DOS program from Jim Green at NSSDC.

The KGT model defines four levels of service (LAS); some costs scale with level of service, some with amount of data, some with both. The four levels of service are:

1) Data holding.
   The data are kept in a 'shoe box', and it is the users' responsibility to find it and get it. When it rots, it's gone.
2) Traditional Data Center.
   Most current data at NSSDC are kept at this level. There is quality control of the input media (is it the right tape?), the media is maintained (if it rots, it is repaired), and there are rudimentary indices and general retrieval software (gets files).
3) Discipline Data Systems.
   This is the highest existing level. There is quality control of the input data (is what you expect on the tape?), the media is kept current, there are sophisticated indices and on-line meta-data, and there is specialized retrieval software (gets data by object).
4) Data Archive and Distribution System.
   This would have interactive, on-line data, science software, etc.

Figure 1, from KGT, shows the predictions of the model for a typical mission at each of the four LAS. Note that even LAS-1 (the 'shoe box') is not exactly cheap, that LAS-2 is somewhat less than twice LAS-1, and that LAS-3 is more than twice LAS-2. Note also that the difference in cost between LAS-2 and LAS-3 could pay for a 2.5 m imaging telescope, a ground based device capable of generating the 200 Gbyte/year dataflow shown in Fig. 1. Clearly choosing the level at which an archive operates is a crucial decision for the scientific and financial health of an organisation.

4. A Small Archive/Database Problem of Tomorrow

Within the next couple of years the Smithsonian Astrophysical Observatory and the Steward Observatory will replace the current MMT with a 6.5 meter single mirror telescope with a 1 degree field-of-view. Dan Fabricant and his collaborators are building a 300-fiber, large beam spectrograph (the Hectospec) for the new telescope.

In 30 clear nights a redshift survey to measure 150,000 spectra of galaxies to 19.5 R magnitude can be finished. Work is currently underway to accomplish this during the first spring of operation. Other groups are working on similar projects, both larger and smaller.

The data train which leads to the final redshift catalog is roughly as follows: images of the sky are reduced to catalogs of stars and galaxies with magnitudes and positions; a selection algorithm chooses the galaxies to be observed (and the guide stars); a set of instrument configurations for the multifiber spectrograph is created; observations are made; CCD frames of many spectra are

https://doi.org/10.1017/S0074180900047604 Published online by Cambridge University Press

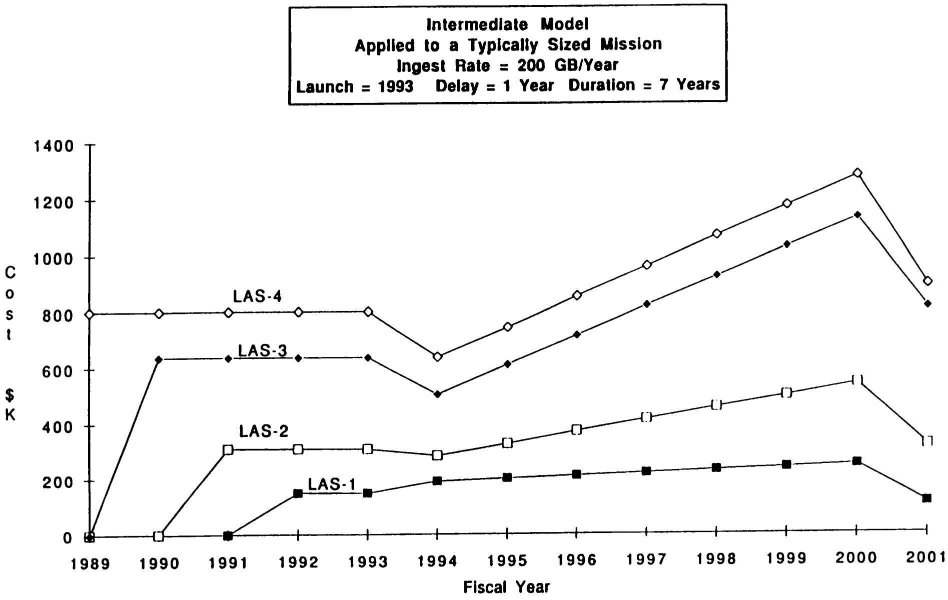


Figure 1. Life-cycle costs from KGT for a typical (200 GB/year) intermediate mission.

reduced to many 2-D spectra, which are further reduced to many 1-D spectra, which are finally reduced to redshifts.

At each of these steps there are data reduction programs, with variable parameters, and calibration measures. At each of these steps there is the possibility of error. Currently substantial human intervention is required to keep errors out of redshift catalogs, both at the reduction stage and at the cataloging stage. With the near total automation required to reduce and catalog these upcoming data, how will the integrity of the catalogs be maintained? It will be necessary to be able to check the whole data train end-to-end, and quickly, after the reduction, when questions arise. To do this the data must all be connected in an archive system which is integrated with the data reduction system. The archive system must be 'smart'.

5. Communications and Standard Interfaces to Archives

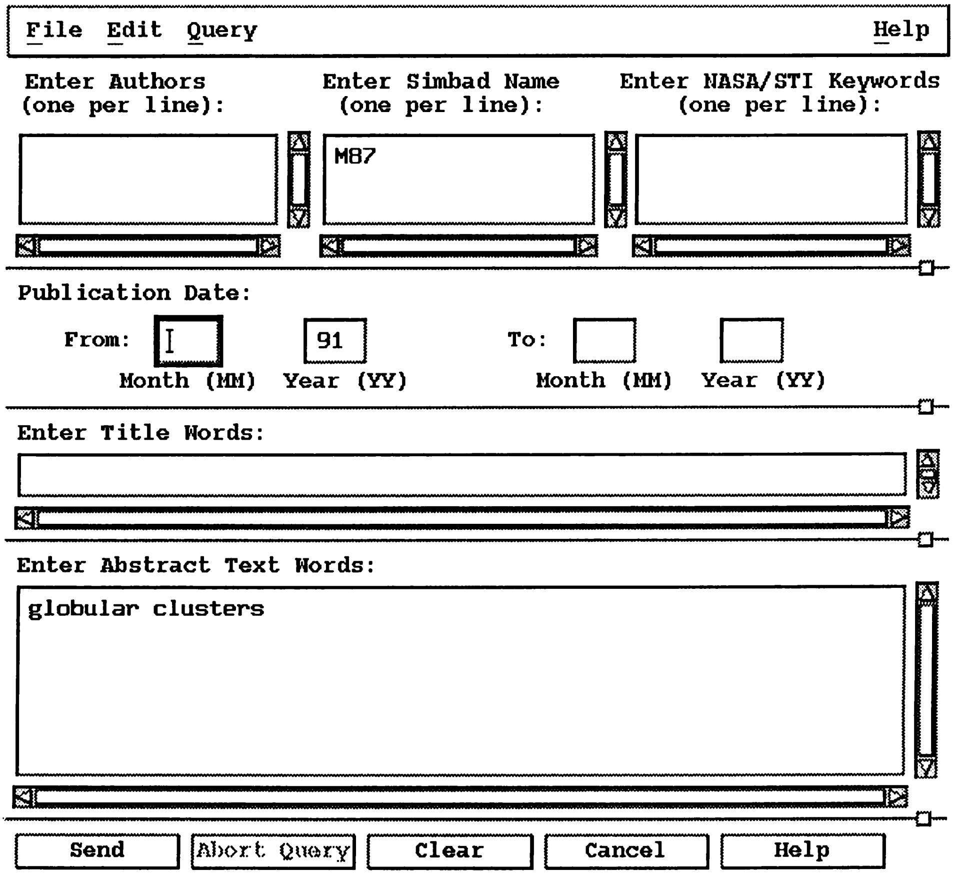
If we are to be able to take full advantage of our new, data-rich environment, we will need to communicate with the databases. Communication requires the use of standard interfaces to succeed; in the case of this article it is English on paper. For a database problem it might be SQL via TCP/IP. Just as the level of discourse between humans is greatly determined by the level of shared language and concepts, so will the level of discourse between databases be determined.

https://doi.org/10.1017/S0074180900047604 Published online by Cambridge University Press

As an example of why these interconnections are necessary, consider the Hectospec redshift survey program. Should there be a question about a particular spectrum, essentially all the data which led up to it, and perhaps some secondary data which could confirm the primary data, will need to be examined. These databases will be created by several groups, as part of different projects, with some not even in the same geographical location; if they are to be queried effectively they must be logically connected.

There are two projects now underway which are building the sinew which will connect database and reduction/analysis software to form an information system. They are the Astrophysics Data System (ADS) of NASA, and the European Space Information System (ESIS) of ESA.

The synergy to be achieved by combining our data in this way can only be guessed at, but must be substantial. An example of what is possible with a simple communication between two important databases is shown in Fig. 2. Here a user of the ADS Abstract Service asks for recent



Enter Authors (one per line):

Enter Simbad Name (one per line): M87

Enter NASA/STI Keywords (one per line):

Publication Date:

From: Month (MM) Year (YY) 91

To: Month (MM) Year (YY)

Enter Title Words:

Enter Abstract Text Words: globular clusters

Figure 2. A typical query using the ADS Abstract Service which combines data from the SIMBAD database and the NASA STI Abstracts database.

https://doi.org/10.1017/S0074180900047604 Published online by Cambridge University Press

papers which the SIMBAD bibliographic database says concern M 87 and which contain the words 'globular cluster' in the NASA STI database of scientific abstracts. As there are nearly 1000 papers in SIMBAD on M 87 and over 2500 in STI containing 'globular cluster', this is an otherwise nearly impossible task. As it is, one gets the few dozen papers on the M 87 globular clusters in a few seconds.

6. Intelligent Retrieval

To retrieve complex data from an archive one must be able to make complex queries. For text archives this might be a query such as "give me recent articles on the statistical study of the large scale structure on the universe", clearly a complex request. This request can now be made and answered within the ADS Abstract Service.

For non-textual data archives we need to develop standard languages to describe the data, so that different archive systems, or humans, can query an archive for complex data. A0V, K2III are examples of well-defined terms in a standard language. One needs to be able to query a spectroscopic database for all G dwarfs without having to give a precise definition of the numerical values which define the spectral types for that exact data.

Within the structure of a standard language there will be dialects, connotations, and shades of meaning. This will be due to several factors; one being different data. One can be certain that objects called G dwarfs on the basis of UBV photometry will not be exactly the same as objects called G dwarfs on the basis of detailed spectroscopy.

Slight differences in definition will result in dialects. It is unlikely that any of the sky survey groups will use exactly the same methods to determine the morphology of galaxies, so the connotations of spiral galaxy in the APM dialect will be somewhat different than in the COSMOS or Muenster dialects. These dialectical differences do not erase the general agreement on what a spiral galaxy is, they just make it a little fuzzy.

With standard interfaces and languages it will be possible for machines to talk to machines to decide to query a database, then machines can decide, based on the contents of several archives to query another one (or perform some action).

For example, in the Hectospec redshift survey, assume a machine decides that the quality of a redshift measurement is in doubt. That machine can then query the source catalogs and a confirming catalog, say the APS galaxy catalog and the COSMOS galaxy catalog. If those catalogs agree that the object was really a galaxy which met the selection criteria, then the data logs could be examined to see that some gross error positioning did not occur. If all this seems (to the machine) OK, then the image archives (at STScI) can be queried for pictures of the object, in a couple of colors. This information, along with the catalog information, the spectrum, and the results of the reduction process can then be collated into a report to be shown to a human. The human can then decide whether to redo the observation, remove the object from the sample, reduce the spectrum by hand, or do something else.

7. Conclusions

Aside from the oral tradition and from our buildings and instruments, archives are what we create as astronomers. Archives are the vehicle by which we communicate with the future. Until

https://doi.org/10.1017/S0074180900047604 Published online by Cambridge University Press

recently we have stored our knowledge in the form of paper books, paper logsheets and note books, and glass photographic plates.

We are now leaving this era and entering the era of digital information. Astronomers have been leaders in making this transition; most of our data has been taken using digital detectors for some time. While most of the developments necessary to handle and store digital information will come from outside of astronomy, there remains a great deal of astronomy specific work to do.

The new digital technologies are threatening to totally overwhelm our current ad hoc attempts to archive our work. Just as AIPS, MIDAS, and IRAF have formed the basis for a world wide effort to share the cost of developing reduction and analysis software for digital data, so must we have a shared basis for the development of archive software for digital data. The ADS can provide the necessary basis.

The costs of information storage and retrieval are huge; the effort must be shared or else our data will be lost.

Acknowledgements

I am happy to thank Martha Hazen, Donna Colletti, Michael Van Steenberg and Joyce Watson for discussions. Financial support was provided by the IAU, the Smithsonian Institution, and NAGW-1714.

References

Klenk, K.F., Green, J.L. and Treinish, L.A., 1990. 'A Cost Model for NASA Data Archiving, Version 2.0', Greenbelt: Goddard Spaceflight Center publication NSSDC/WDC-A-R&S 90-08 (KGL).

https://doi.org/10.1017/S0074180900047604 Published online by Cambridge University Press