THE STATUS OF ARCHIVING ASTRONOMICAL DATA

Summary by: **R.E.M. GRIFFIN**

https://doi.org/10.1017/S1539299600012259 Published online by Cambridge University Press

THROUGH A GLASS DARKLY

The Status of Archiving Astronomical Spectra

R.E.M. GRIFFIN
Institute of Astronomy
Cambridge, UK

1 Converging Elements

The Legacy of History

Astronomical research comprises a curious mixture of team work and individualism. From the hardware point of view data are ends in themselves, while from a strictly scientific aspect their acquisition is but the first stage in the complicated process of building astrophysical models. This dichotomy is reflected in a polarization of attitudes regarding the handling of observational data, and the activity of creating archives of astronomical data for use by posterity has consequently tended to fall in no-man's-land. To the technology team, a telescope that can successfully deliver a data-bank full of raw observations has achieved its specification, while to the scientist who is under pressure to publish papers on fresh science, the concept of voluntarily creating public archives out of data originating from personal ideas may seem more than a little alien. Nevertheless, the formation of useable and efficient archives of astronomical observations is an activity that has taken on new meaning with the advent and monopoly of digital detectors; it is a procedure that builds bridges as well as access routes and it opens new global perspectives for astronomical data, but it still relies too heavily on individual initiatives.

Historically, astronomical observations were obtained on photographic plates which the host institution indexed, guarding jealously the only tangible evidence of its observing activities. Yet as soon as the tangible medium gave place to an electronic and erasable one, enthusiasm for those commendable habits of a lifelong tradition seemed to evaporate. It was not merely the whims of fashion but hard practicalities that were responsible for undermining such a fundamental support activity; had today's comparatively sophisticated telecommunications, storage devices and access and retrieval capabilities been available twenty years ago, greater use would surely have been made at the outset of the new opportunities to develop electronic archiving in many areas and to exploit the time-dependent element that is present in all astronomical observations.

Certainly, the creation of catalogues was an important element of stellar astronomy long before the development of spectroscopy. The majority of the early catalogues concentrated on astrometric positions, possibly because of astronomy's commercial importance in aiding navigation at sea. Provided that

I. Appenzeller (ed.), Highlights of Astronomy, Vol. 10, 607–617.
© 1995 *IAU. Printed in the Netherlands.*

https://doi.org/10.1017/S1539299600012259 Published online by Cambridge University Press

their contents are homogeneous and complete, such catalogues could have considerable and widespread uses today especially for statistical work, but that potential will only be realized if such studies can be carried out cost-effectively, i.e. if the data are accessible by computer. Those now attempting to use the astrographic catalogues compiled during the earlier part of the twentieth century have come to recognize how important are *quality* and *accessibility* for good data archiving. The astrometric bias in those surveys also limits the application of many of the catalogues for studies of variable objects since many were excluded on account of their very variability. Modern sky surveys promise to be better in that respect, but cannot (yet) offer a similarly long time-base.

Approaching the Millenium

However patchily history has developed, it is certainly time that some of the most obvious deficiencies were now made good. Besides, fashions are on the move again, and the very fact that astronomical observations are being made at increasing rates across an ever widening range of frequencies means not only that an investigator needs to be able to invoke very many data very efficiently in order to maximize the potential of an investigation; (s)he also needs to be an expert in every single field – or to participate in consorted efforts. Consorted efforts, plus the increasing complexity of the style of modern data management, immediately invoke data-sharing and sorties into data-banks to retrieve the fruits of other programmes. Worldwide archives of astronomical data must therefore be only just round the corner. However, the correct timing to launch a worldwide enterprise on behalf of astronomers has proved extremely delicate. Traditions are strong and unequal, and the situation across the world and across different disciplines is far from level; visual-star observers, for example, have had their own managed data-bank for decades, and wide-field observers can already scan a list of index catalogues that are being prepared of many of the direct plates that lurk in various corners of the world, while the converse seems to prevail in radio astronomy, whose source identifications and designations appear as baffling as the objects themselves.

For spectroscopy, the battle for data archiving was fought and (it turns out) largely won three years ago at the Buenos Aires General Assembly, when the IAU agreed to take on board the *possibility* that a global movement to archive astronomical spectra could be made to work. To ease the project forward, the IAU published a Resolution (C 13) in which it placed its faith in the ability of a Working Group (size undefined) to investigate and extract local opinions, needs, aspirations and even condemnations. The 11-strong *Working Group for Spectroscopic Data*, which was inaugurated some months later during a follow-up at meeting the Vatican Observatory (Griffin 1992) has striven to merit that faith by advertising the initiative, presenting the arguments and entering into debate at a number of official and unofficial meetings across Europe and North America. Thus after its three probationary years the WG for Spectroscopic

https://doi.org/10.1017/S1539299600012259 Published online by Cambridge University Press

Data Archives felt sufficiently sure of its mission, and sufficiently encouraged by the subtle and sometimes not-so-subtle changes that had been taking place in professional attitudes (to quote one example, "The data belong to the Guest Investigators and are not our concern" has become "We believe that archiving of data is an important matter and we are glad the IAU is taking it seriously") that, together with the WG on Radio-astronomical Databases, this Joint Discussion was arranged in order to debate in public the next step ahead.

2 Focus on Specifics

Purposes and Policies

In view of the many recent meetings and workshops at which the archiving of astronomical data received thorough airing, JD 20 sought to concentrate on the principles for spectroscopic archives of the future. Individual initiatives in archiving activities were displayed in posters that sampled rather satisfactorily the whole gamut, from data-collecting activities such as those of KPNO, HST and radio astronomy, spectra-gathering from Haute-Provence Observatory, Ondřejov, Rozhen and IUE, databases of selected wide-field plates and images collected in Romania and Bulgaria, and a compilation of X-ray sources. Other posters discussing problems specific to radial-velocity compilations, data communications networks and the horrors of confusing nomenclature served to highlight the importance of factors external to the central activity of data-collecting.

In keeping with the intention to discuss the status of archiving astronomical *spectra*, the aspects of 'data' archiving which were raised during the meeting were applied to the specific activity of spectroscopy unless stated otherwise. That is not to say that only spectroscopic topics were addressed; far from it – the term 'spectra' did not even appear on the official menu. Nevertheless, there are many problems for astronomical archiving in general, and which the JD could discuss profitably in the context of spectroscopy; the tasks posed by that area alone are major enough. Any plan for comprehensive archives of *all* types of astronomical data at once would be much too broad; it would involve many resources and techniques, and it is doubtful whether a satisfactory solution could be created quickly enough before the whole was threatened by changing technology. In contrast, by starting near the beginning in one area only and developing strategies that could later be generalized, it will be possible to learn and win at the same time.

Objectives

The usefulness or "success" of an archive will depend upon how it is created and for what purpose. The two clear-cut design modes, object-based and observatory-based, are almost complementary. Object-based archives concern

https://doi.org/10.1017/S1539299600012259 Published online by Cambridge University Press

'hot' topics, and are usually created when it suddenly becomes scientifically imperative to gather together all observed spectra of a specific target (SN 1993J comes immediately to mind, but there are others) which is manifesting an event of such rarity that it is vital to collate all possible data, process them by optimal procedures and offer access to the world, who will milk the observations thoroughly for a short space of time. Those exercises actually strengthen the case for routine full-scale archiving ("If only your observatory produced ready-to-use versions of its observations anyway, how much easier ..."). The effort required to start up an object-based archive today is disproportionately large because none of the necessary search software, reduction routines or suitable communications networks are in place and a multitude of unexpected problems have to be solved during the creation of the archive in respect of just one target. Valuable skills are being acquired as by-products, but it goes without saying that such an approach is hardly the most efficient or attractive method of pursuing archiving as a long-term commitment, nor is it correct to estimate the effort required to create and maintain an instrument- or observatory-based archive simply by extrapolating the resources absorbed by a given object-based one.

An observatory-based archive, or *active archiving tool*, should contain everything *of archive quality* from one or more telescope or spectrograph for either a limited or an undefined length of time. It is thus controlled by instrument or technique and not by science, and has a greater potential than an object-based one. To select the contents of an archive on the basis of a scientific objective is to freeze a present-day concept into a tool of the future,and thus to build in a bias that limits its uses. The ultimate development target for an observatory-based archive is the *historical reference archive*, which will store many of mankind's accumulated astronomical observations as optimally reduced, adequately documented, historically traceable and globally accessible data for use by any astronomer in the world. It is imperative that it be simple to use and highly reliable. The greater the degree of built-in automation, the lower the manpower involved in its daily operation and the cheaper its maintenance in relation to other fundamentals of an observatory's operations.

Commitment

Many observatories have now either commenced or are planning some form of archiving tool, and policies to incorporate archiving capabilities in future telescopes are becoming more widespread. The 15-year project at CfA, achieving on-line retrieval for nearly 200,000 high and low dispersion spectra, and the catalogue of spectroscopic data for HD stars constructed in Toronto, are commendable examples of archiving activities that are being absorbed and carried out "in house". It is also generally accepted that the best stage at which to include the rudiments of an archiving tool in a new telescope is right at the start. These are encouraging signs, though just how far the designers and planners of new telescope facilities are both willing and able to commit themselves

https://doi.org/10.1017/S1539299600012259 Published online by Cambridge University Press

right now is not clear. Were there a fully-fledged historical reference archive in successful operation in the astronomical world, it is inconceivable that new facilities would not expect to contribute to it and the (relatively small) funds would be set aside without debate. Because the archiving movement is only in its infancy, there is inevitably an element of uncertainty which constitutes an inertial drag upon every proposed expenditure of effort in that direction, and it is only natural to prefer to play the angel and wait for others to rush in.

Each observatory archive is to some extent an individual creation, and whether it will be suitable for development into the ultimate historical reference archive will depend upon the level of importance, *as estimated at the design stage*, attached to the production of a suitable archiving tool for the astronomical public. A scheme to 'save all the bits', for instance, though an essential first step in sizing up the task it is little more, and if it is to become *the* working basis for an archive development it must learn to filter out those elements which are not, in the long run, going to add anything that is worth saving for future users. There is no advantage in diluting data of true archival quality with others of dubious calibration and uncertain value, and an archiving tool that is devoid of quality controls will also be too cumbersome for the selection tasks that lie ahead.

Initial Design

It is clear that the planning that will eventually shape a full-scale archiving capability must define and quantify its *selection* policies in a style that can be interpreted and understood world-wide. Homogeneity, compatibility and accessibility are all words that conjure up a commendable vision of cooperation and collaboration by consultation, and which will presumably become more feasible *globally* as more unified forms of computer technology spread. For instance, the basic raw observational data stored in a local data-bank must be recorded in well-defined formats (the obvious one at present being FITS, which the IAU has championed effectively), such that the vital information in the header that refers to the observation can be read automatically and entered in an index list. A variety of quality controls must be decided on and applied, and a uniform and transportable software must be selected or designed both for applying calibrations in automatic mode and for enabling requests – from humans or from computers – to be handled efficiently. It is equally clear that the same planners must define their nomenclature carefully from the start. The lack of a leading coordinator in archiving projects worldwide has not prevented many individual efforts from commencing in isolation, each with its own style of nomenclature and set of definitions, and the resulting duality of terms usage by either an observer or an information technologist is causing a confusion that could threaten to suffocate the project. At this stage of the game, an entity has to be defined by virtue of what it is to contain, not by its name; the latter is an arbitrary choice.

https://doi.org/10.1017/S1539299600012259 Published online by Cambridge University Press

The central ideal for an active archiving tool, translated elsewhere as "Get what you want, but no more than you want, when you want it" (Wamsteker 1989), conveys a simple but essential message to the designers. Calibration files, data processing software and observing proposals do not form part of those needs and are therefore unnecessary encumbrances; the researcher does *not* want to have to become a specialist in the data reduction of every technique that have at any time in the past yielded the spectra of interest, and an archiving tool that does not accept that obvious fact will only enjoy a very restricted usage. Associating the correct calibration observation(s) with given object files is a data handling problem for the observatory or its local data management team, not that of some astronomer the other side of the world.

In order that they be processed in the most effective manner, observational data need to remain where the expertise is. Any generalized archiving tool has to accommodate the very large number of variants of the common theme of data acquisition that currently occur. Observations made by space missions are buffered by some form of ground control, which ensures that recognized sequences of object and calibration exposures is followed and that a prescribed system of object naming is used. The choice of instrumentation is limited, and the changes that are made get recorded and logged automatically into the ground-control system. For spectra taken with most present-day ground-based instruments there is no such buffering (sometimes not even a staff astronomer is present to record changes); the more directly an observer interacts with the equipment, the more opportunities there are for making undocumented changes, however small, for taking liberties with any prescribed observing procedures (standards, calibration exposures, or the like) or for ignoring an advised object-naming scheme. Non-standard observing practices risk the introduction of non-uniformities into the final archive itself. Any resentment at any consequent restrictions in observing practices will vanish once an observer becomes a beneficiary too – such is human nature.

Cost

The major factor that dominates archiving policies today is of course resources, or cost. If each observatory were being required to develop and produce its own historical reference archive, the cost to each could indeed be appreciable, disproportionately so for smaller observatories. However, if it is possible to produce a suitable software package in a generalized and transportable form then the development of the historical reference archive comes much cheaper. It is the Working Group's new goal to carry out that development work, spreading the cost by seconding the expertise it needs from agencies or institutions, and building a product that can be interfaced with individual raw-data banks. Each observatory will then have only to ensure that the first stage – a sufficiently sophisticated form of 'saving the bits' – is installed and correctly managed. Many observatories already have the rudiments of such a scheme in place as a secu-

https://doi.org/10.1017/S1539299600012259 Published online by Cambridge University Press

rity measure for the observers; *those* running costs are small, even insignificant when compared with the actual cost of acquiring the observations in the first place. A manageable raw-data bank also offers a means for monitoring automatically the performance of an instrument (or observer) and the quality of the routines being followed at the telescope, thus providing another control on the cost-effectiveness of an observatory's budgeting.

However, the true cost of a successful archiving tool has to be viewed in the context of the overall cost of conducting research. It was estimated that most observatories would not be 'able' to spend more than 2 or 3% of their budgets on the general activity of archiving, though the willingness to do even that is clearly a question of choosing priorities. Organizations which do already offer a workable archiving system can testify that what they have purchased – the potential to increase the usefulness of their observational data – is cheap at the price, as indeed, by analogy, was the development of electronic mail or telefaxing. It has also been said that the historical reference archive is "too costly except for very expensive data". But the cost of a telescope does not determine the scientific merit of the observations made with it, nor is the cost of those observations limited to that of just the detecting equipment. Economic superiority is not the ideology which drives the science, and in any case astronomical data usually complement rather than compete in value. No-one would deny the importance of archiving observations of rare or sporadic events even though they may have been obtained by 'inexpensive' techniques.

It has also to be remembered that the true cost of obtaining useable spectra includes the effort of *reducing* the data, and if it is left to each researcher to perform piecemeal the necessary routines to extract all the observations in a suitable reduced format then their true costs will rise with their usage. If the observations are not reduced as fully or as frequently as could be the case, the price per unit of good-quality science derived from them will again increased. It is in any case incumbent on astronomers whose instruments have been publicly funded to assure the general public that the facilities are being used cost-effectively. If a particular experiment promises to be a major expense to the astronomical community, the project manager should ensure that adequate provision for data reductions and data sharing are established at the design stage, where they can be incorporated and manipulated most efficiently. Implementation of a full archiving scheme would both ease these difficulties in real time *and* contribute every benefit of the expenditures to the future.

3 Parallel Activities

As well as debating the philosophical goals of, and practical considerations pertinent to, the creation *per se* of archives of spectroscopic data, the meeting heard reports of recent activities of the three IAU Working Groups (for archives of spectroscopic data, radio-astronomy data and wide-field images) and of the

https://doi.org/10.1017/S1539299600012259 Published online by Cambridge University Press

archive of variable-star observations.

Photographic Plates

The Working Group for spectroscopic data had concentrated on sampling attitudes and discussing the possibilities of its mission in various corners of the earth, believing that a co-ordinated effort should not begin until a plan of action had been agreed and suitable expertise had been brought together. In the meantime it had discussed as a separate issue the question of safeguarding all the photographic spectra in the world (currently stored in about 35 different observatories); encouraged by its earlier proposal to create a World Plate Store, it had considered suitable sites, had selected Haute-Provence Observatory in France and had obtained the written consent of its Director. Almost all observatories when told of the plan had expressed cautious support, the caution probably stemming more from a fear of the expenses to be incurred in the shipment costs than in the future welfare of the glass. Of course, any ongoing research efforts that were making use of plates would not be allowed to suffer as a result of a blanket order to disband the plate archives.

Radio and Extra-Galactic Astronomy

The merging of published on-line information on stellar objects into a centre such as the *Centre de Données astronomiques de Strasbourg* began at a time when a great deal of catalogued information about stars (the Centre was originally called the *Centre de Données Stellaires*) was available (even though the bulk was originally only in printed form) and reasonably complete cross-identifications could be made. In contrast, most extra-galactic surveys post-date the commencement of computerized data-bases and therefore had a crisper start, but the relevant knowledge is much more confused and patchy. Since the need to facilitate the exchange of knowledge about extragalactic sources is currently at the nerve centre of so many astronomical projects and institutions it is puzzling to learn that much of the necessary work is still being left to the initiative of a few people. The design of an extragalactic data-bases should take advantage of the experience of their stellar precursors, avoiding duplication of effort but seeking complementarity in their products. Too much emphasis seems to be placed on the need to scan the literature for what may have been only a passing reference once to an object, and too little on bringing the actual observational data to the public's own working space. The need to distinguish between these fully complementary but totally separate activities cannot be emphasized strongly enough.

The WG on Radio-astronomical Databases reported that it had compiled many otherwise unavailable radio-source catalogues and prepared a bibliography of radio-source identifications (Andernach 1994a) but found only a lukewarm readiness by managers of existing object-oriented databases to incorporate its

https://doi.org/10.1017/S1539299600012259 Published online by Cambridge University Press

efforts. It also drew attention to the fact that only very few of the astronomical catalogues stored at the CDS have been collated into SIMBAD, and that currently only about 30% of published tabular material is being stored at the CDS. A more concerted effort must be applied *at an official level* to make better use of existing information (Andernach 1994b).

Wide-Field Plate Database

A master index of all wide-field plates, including Schmidt plates, taken during the past 100 years or so is currently being compiled by a small but dedicated team in Bulgaria, in association with Commission 9's Working Group on Wide-Field Imaging. The total is estimated at 1.7M plates, of which about one third are indexed in catalogues that are fully computer-readable, and an additional one fifth in catalogues that are partially computer-readable. The team is preparing a Wide-Field Plate Data-Base, whose objective is a merged list of basic information about all the 1.7M plates. Before inclusion in the data-base the contents of each source catalogue must of course be unified (2000 co-ordinates, UT, standard formats, etc.), a task which is possible straight away for 62 catalogues (involving about 0.37M plates) and has so far been completed for 40 of those catalogues (0.27M plates). The catalogues which are already, or will soon be, included in the data-base therefore represent 21% of the estimated total in existence. Attempts are being made to retrieve the remaining catalogues and to computerize the indexes of them all, and plans are also being discussed for phase two, which is the more daunting challenge of organizing and carrying out the digitization of some of those plates.

Archives of Variable-Star Observations

Photometric observations of variable stars have almost always been popular amongst amateur astronomers for a combination of reasons: the accessibility of the many bright objects to small telescopes, the possibility to make valuable observations without needing particularly sophisticated or expensive equipment, and the excitement of monitoring an event by plotting one's own measurements of a changing parameter. At the same time, the frustrations of missed data have heightened an awareness for widespread collaboration, and it is not surprising to find that the community of variable star observers have had *their* act together for some time – since 1911, in fact, when the American Association of Variable Star Observers (AAVSO) was formed in the USA – and are consolidating their position daily. By concentrating on just one field of observational research the AAVSO has developed and maintained a well-documented archive in computer-readable form, currently numbering some 7.5 million observations and growing at a rate of about 35% per year.

https://doi.org/10.1017/S1539299600012259 Published online by Cambridge University Press

4 Diverging themes

The meeting also considered the particular attractions of data archives for astronomers in a developing country, the advantages of archival material for education, and the protection of international archiving centres facing domestic problems of quite a different genre.

Developing Countries

In a developing nation such as China, archives of astronomical data are clearly playing an important rôle in enabling its astronomers to keep abreast of trends and technologies in the world and to participate, if only from armchairs in the first instance, in observations that would otherwise remain unreachable to them. Statistics indicated an encouraging awareness of, and willingness to use, archived data when and where available. Computer networks are still primitive in some part; not all countries are equally well endowed with national databanks, and their initial holdings tend to concentrate on specific local interests. There is always scope, it seems, for placing greater emphasis on education about the availability and possible uses of archived material, the more so since archives are an amazingly cheap method of providing real enrichment to a willing and able, if technologically impoverished, scientific community.

Education

The AAVSO has been active in developing tactics that use its archives to help promote an interest in the science of astronomy, especially for students where "hands-on" experiments are particularly popular. Education is an area in which archived data fulfil several important purposes – simulating events, demonstrating statistical properties through numerous examples, testing routines and models, and providing real data for teaching facts and for learning skills. Because of its straightforward concept, visual photometry is a powerful tool for stimulating an interest in astronomy among younger students, and the AAVSO is developing for schools a teaching programme that includes computer programmes, data access, video tapes, finding charts and manuals (for both pupils and teachers), and provides opportunity for the excitement of discovery.

The Future of Specialist Archiving Centres

Sonneberg Observatory, an organization of considerable repute and achievement in the field of sky surveys and data archiving, would be highly appropriate as an international centre or "Sky Library" for the wide-field plate endeavour. Unfortunately Sonneberg was currently facing untimely closure and the situation was aired somewhat passionately because the precedent of closing down an organization whose whole *raison d'être* had been in surveying and archiving could generate a very serious menace to similar enterprises elsewhere.

https://doi.org/10.1017/S1539299600012259 Published online by Cambridge University Press

5 Face to Face with the Future

A document had been circulated around the meeting by the WG for Spectroscopic Data Archives, proposing to raise the profile of its activities in order to undertake the actual creation of archives of observed spectra. It was felt that public opinion was now advocating this forward step, and references to the document which had been made during the debate sought only to clarify details, not argue the principle. When asked to endorse the proposal, the meeting did so with very little dissent.

Exactly what official form the new activity should adopt was not quite clear, and it was already too late, at that stage in the General Assembly, to submit a Recommendation for consideration by the Executive Committee. It was agreed that the status of *Working Group* should still be adequate for the task in hand, at least in the preliminary rounds, and the meeting seemed well content that something positive might soon commence.

References

Andernach, H.J. (1994a), *Handling and Archiving Data from Ground-based Telescopes*, eds. M. Albrecht & F. Pasian. ESO Conf. & Workshop Proc. **50**, pp. 117–124, ESO Garching

Andernach, H.J. (1994b), Bull. Inf. CDS, **45**, pp. 35–45.

Griffin, R.E.M. (1992), *Comments in Astrophysics*, **16**, pp. 167–185.

Wamsteker, W. (1991), *Databases & On-line Data in Astronomy*, pp. 35–46.

https://doi.org/10.1017/S1539299600012259 Published online by Cambridge University Press