RESCUING AND RECOVERING LOST OR ENDANGERED DATA

Elizabeth Griffin[1]

[1]**Herzberg Institute for Astrophysics, Dominion Astrophysical Observatory, 5071 West Saanich Road, Victoria, BC, Canada, V8E 2E7**
Email: Elizabeth.Griffin@nrc.gc.ca

ABSTRACT

This paper summarizes the vital importance to the scientific community of rescuing historic scientific data, presently in various informal, non-digital formats, from likely oblivion and making them accessible digitally for trend analyses. It proposes models whereby historic biodiversity and astronomical data can be recovered as Canadian initiatives, in the hope of stimulating further discussion of such simple yet essential rescue missions in the context of Canadian scientific research.

Keywords: Historic data, archival data, non-digital data.

DIGITAL DATA IN RESEARCH

Most modern scientific data are collected and arranged in electronic (i.e. digital) databases, a format that can only be interpreted by highly specific computer software. To the man in the street, such data are highly user-*un*friendly: unreadable by the eye, cannot be photocopied, mailed or studied visually, nor copied by hand into another medium. Nevertheless those electronic data are the supreme core of modern scientific research; no matter how numerous, complex or widely distributed the data-sets may be physically, modern networks can sort, classify, compare or combine subsets or whole groups in ways that far outclass anything that could have been feasible even in principle 30 years ago when most research data were non-electronic, let alone dreamed of in much earlier decades when a "computer" meant a person with pencil, paper and some mathematical training. And all in a space of time that is mind-bogglingly short.

But does this mean that scientific research can *only* handle digital data? Surely scientists have been recording observations for more than 30 years? How can we study century-long trends if our data-handling methods are so rigidly selective as to the type of data that can be ingested by a given analysis?

THE VALUE OF HISTORIC DATA IN MODERN RESEARCH

A number of disciplines (astronomy, aerial photography, architecture, history ...) can point to hoards of irreplaceable records and observations on photographic plate or film; oceanography, botany, zoology and meteorology (and many, many more) are aware of invaluable hand-written records in private diaries or notebooks: records of the weather, bird migration, wild-flower sightings and frequencies, times and heights of ocean tides, and so on. Do those not contribute to – and maybe provide crucial evidence for – studies of behavioural patterns, long-term trends, and the like? Of course they do, but are we as a community of scientists making the necessary efforts to rescue those non-digital observations? By and large – No.

Some scientific projects are producing so many data today that it takes exorbitant computer power to cope with the volume, even at today's speed and sophistication. Some scientists also

argue that it would be better to learn to cope with today's data first, to the total exclusion of historic ones if necessary, because the new ones are of much better quality. Better quality they

may indeed be, but however wonderful their resolution, detail or signal-to-noise ratio, today's data can never recapture events of the past. And it is so often the latter, *in tandem with* more recent material, which help solve problems or enrich understanding by providing evidence of trends. An ability to model a snapshot or a cross-section can tell us certain things about the subject, and at least as much about the choice of software used, whereas an ability to model and reproduce *changes*, in both directions of time, tells much more about the reliability of the basic model.

HISTORIC DATA AND TREND ANALYSES

What characteristic time-scales are of greatest significance? The answer will depend on the scientific discipline and on the question(s) being asked. In the physical world characteristic time-scales of trends may be microscopic – or huge; in cosmology they may need to encompass the beginning of time itself. In broad terms, if a suspected variability is strictly periodic, the length of data sequence should be commensurate with at least a few cycles. There are many instances of periodic variability in the natural world, especially within the physical, inanimate universe: planetary orbits, flashing quasars, the rotation of the Earth, its annual orbit round the Sun. But valuable as it surely is to refine the parameters of such periodic changes to a further decimal place, much (I want to say much *more*) can be learned about the physical body in question if modulations to the period are uncovered. Period modulations, which can only be identified from series of observations that are very much longer than a few cycles, may indicate the influence of another gravitational source (e.g. that a supposedly binary star is in fact a triple system), or a physical change in the body in question that is affecting the cause of the observed periodicity. In the terrestrial world where so many climatic factors are governed by chaos, and in the biological world where local factors or genetic influences are paramount and periodicity itself may have only subordinate significance, it is the modulations, or trends, that are the noteworthy observables: the intensities of different El Niño events, alterations to migratory patterns of birds, changes in locations of flora – to name just a few.

THE COST OF LOSING – OR IGNORING – HISTORIC DATA

While the foregoing may appear commonplace, even trite, I have stated it thus in order to emphasize that precise knowledge (i.e. based upon observation) is perforce *circumscribed by the time-base of the data which scientists can access today.* Any further claims are merely extrapolations of models which have been proven only in relation to available observations. And given that all the analyzed data nowadays need to be computer-readable, the upshot is that many sciences effectively commenced only as recently as their oldest digital data. On that criterion even astronomy, widely regarded as particularly long established, generally commenced in the 1990s when astronomers not only installed digital detectors but were also able to introduce computers of sufficient capacity and speed to store what was being collected.

The rider, however, is amazingly encouraging: every single step taken to rescue and preserve older ("historic") observations increases the scope for all researchers to extend, broaden and enrich knowledge in the discipline in question. But it comes with a dire caveat: *many historic observations are in danger of permanent loss, so action is urgent.*

SCIENTIFIC DATA ARCHIVES IN CANADA

Young though Canada may be as a civilized nation, its collections of scientific observations rival and may well exceed those of other nations in both breadth and time. For example, for 6

months in 1919 Canada was home to the world's largest astronomical telescope – the 72-inch of the Dominion Astrophysical Observatory (DAO) near Victoria – before the Mount Wilson 100-inch reflector in California was completed; the DAO's archive of photographic plates (mostly of stellar spectra) contain observations made right from the beginning of routine operations. and plates from its more recent 48-inch as well (http://www.hia-iha.nrc-cnrc.gc.ca/dao/). The DAO is also home to the Canadian Astronomy Data Centre (CADC) (http://cadc-ccda.hia-iha.nrc-cnrc.gc.ca/), which offers access to a range of dynamical databases of both Canadian and international astronomical data. Another source of scientific data near Victoria is the Institute for Ocean Studies, specifically created to monitor and research seismology and associated events. The David Dunlap Observatory (DDO) (http://www.astro.utoronto.ca/DDO/) in North Toronto, younger than the DAO by 16 years, has its own extensive photographic archive – partly of images and partly of spectra. The World Ozone and Ultra-Violet Radiation Data Centre (WOUDC) (http://www.mrc-smc.ec.gc.ca/woudc/) is also located in North Toronto.

HISTORIC DATA IN PERSONAL COLLECTIONS

However, "observations" can refer not only to formal measurements made with purpose-built equipment in accredited institutions, but also to unpublished records in private keeping. With climates spanning arctic and tundra to near tropical (or so some have us believe it is in Victoria) and with native cultures encompassing Indian, Eskimo, fish farmer, coast, lake, prairie, valley and mountain dweller, not to mention all those introduced from overseas, Canada has an environmental and cultural richness that is hard to equal in any other country. Surrounded by such biodiversity, many Canadians collect or have collected natural and environmental observations out of personal curiosity – measurements of the weather (temperature, rainfall, snow depth, cloud cover, etc.), records of bird sightings, dates when certain flowers bloomed, the depths of lakes or tides – the list is surely much longer. Those snippets of information, though presently inaccessible digitally (and not even discussed in conventional literature), are nonetheless potentially of vital importance. Most of those records will be hand-written in personal notebooks or diaries, probably well out of sight; many will have outlived their authors. Does anyone really want them? The answer is, emphatically, Yes. Those records bear invaluable and irreplaceable witness to local and global conditions that perhaps no other modern source can reveal. Moreover, Canada now has a fascinating opportunity to develop leadership in displaying and sharing its many resources of historic biological and zoological data.

RESCUING HISTORIC BIODIVERSITY DATA

Canada is one of the 32 partners of GBIF (http://www.gbif.net/portal/index.jsp), the Global Biodiversity Information Facility. Created in 2000, GBIF is a unique endeavour to bring data on biodiversity to the desktop of anyone with access to the Internet, and is designed as an interconnected set of databases containing information about all 1.8 million recorded species of organisms, from bacteria to whales, that have received official scientific names. Paradoxically, while biodiversity is greatest in the world's tropical regions, i.e. in developing lands, the superior collections of scientific observations are unquestionably located in developed countries. But while GBIF may therefore want to concentrate on field-trips to collect and bring home *current* biodiversity information, much of which will be, or can quickly become, digital, let it not overlook the rich sources of historic information in personal keeping, *whatever their present*

format, since those are the ones which are likely to harbour the seeds for trend analyses.

But while world data centres are always open to receiving new data, especially of historic

provenance, there is usually little positive drive to acquire such from private possessions, or to recruit the assistance of the general public. Although it takes a project of the complexity and power of GBIF to make full use of accessed information by collating and comparing, filling in gaps and assessing the errors where data overlap, it requires only a small additional organization to contact many of those "hidden" data-sets in private ownership. To involve the general public (i.e. anyone not within biodiversity research) requires thoughtful, repeated advertising through the most likely channels (local media, etc.) summarizing the GBIF project and explaining, with examples, how access to observations presently in private collections can prove immensely useful. Citizens with useable records in their possession are then invited to visit their local library or equivalent cultural centre with their record-books or lists, where a librarian will photocopy and return the records – unless the holder wishes to donate them. The librarian will key-in the information, either personally or under supervision, in a nationally-agreed format that includes all necessary meta-data, and will transmit the results to whichever appropriate organization (maybe Canada's newly restructured Library and Archives Canada) is designated as the central nerve.

This simple model for a united effort to locate and digitize "lost" scientific information can energize a pilot project representing Canada's particular contribution to GBIF. Based on the premise that co-operation will be forthcoming provided that involvement is fully explained and doesn't cost the individual anything, it should prove both extremely cost-effective and very productive. Despite the size and natural diversity in continental Canada, the project should work as well here as anywhere; the difficult terrain to cross is not distance – telephone and radio take care of that – but enlisting the co-operation of Canadians who may be unsuspectingly hoarding those handwritten observations.

AN OCEANOGRAPHIC PRECEDENT

There is already an impressive precedent for an effort of the kind proposed above. In late 1992 the international oceanographic community established GODAR (Global Oceanographic Data Archaeology and Rescue) (Global Oceanographic Data Archaeology and Rescue) to locate and digitize ocean profile and plankton data-sets not then in digital format and to ensure their submission to world data centres. There were also moves to migrate to new media any electronic data at risk through medium degradation. Seven years later an impressive amount of new information had been acquired by the GODAR project (some 2.3 million temperature profiles, 600,000 plankton observations ...), which have since been made accessible world wide through the World Ocean Database. The objective of GODAR was "to increase the volume of historical oceanographic data available to climate change and other researchers", and so encouraging have the results proved that the scope of GODAR has been increased to include other pertinent data.

RESCUING HISTORIC ASTRONOMICAL DATA

Another scientific discipline in which Canada can take the lead is Astronomy, an area in which Canada has long been prominent in instrumentation, data acquisition and research. Trend analysis is particularly germane in astronomy, since all celestial objects are evolving and new observations frequently complement existing series of data; rarely do they simply supersede them. But despite the huge potential of Canada's archived plates to complement relatively

modern series by extending them backwards in time, they can be neither interpreted nor usefully re-used and shared without a focussed effort and specialized equipment to translate the photographic information into digital, calibrated, scientifically-meaningful units.

Almost no astronomer today has the requisite high-quality, high-fidelity instrumentation for converting photographic observations into digital ones. The traditional digitizing instrument purpose-built in the 1980s by Photometric Data Systems (their scanners are still known by the company's initials: PDS), has been a victim of the electronic revolution, and any scanners that were not discarded at the time now need major hardware and software upgrades. Just a few observatories in the world have upgraded and maintained a PDS scanner, and the DAO is one such place. Flat-bed scanners, which are both cheap to purchase and rapid in action, seem satisfactory for some types of direct images but are not the ideal answers as they introduce problems of their own; the best approach seems either to upgrade a PDS (probably mandatory for spectra) or to design a much faster replacement by employing new-technology hardware.

A Working Group of the International Astronomical Union has for some time planned a scanning laboratory at the DAO, the principal objective being to produce ready-to-use stellar spectra from selected plates for world-wide dissemination (via the CADC). This "Spectroscopic Virtual Observatory" will commence with Canadian plates, but will then progress to plates from other archives worldwide and will thus establish a dominant lead in this aspect of data rescue too. While the basic problem is not solely a Canadian one, its solution by and within the Canadian scientific community will establish a fantastic precedent, to the enormous benefit of scientific research not only in astrophysics and space sciences but in related disciplines such as atmospheric science too (see http://www.lizardhollow.net/PDPP/htm).

A GROWING URGENCY

Responsibility for undertaking the rescue and accessibility of historic data for the benefit of scientific communities tends to fall upon the shoulders of the few individuals who initiate the ideas, but those individuals are unlikely to command the necessary resources to accomplish their task. Meanwhile, plates are slowly but surely deteriorating with age; observatory Directors with no training in photographic techniques would prefer the storage area of their plate archives to be given over to more pressing current needs; a frequent complaint is that "no-one asks to use the plates nowadays", which is hardly surprising since almost no observatory has yet generated a digital inventory of what its plate archive contains (Griffin 2002), and the only way of finding out is by personal visit and manual search. Although plates lacked the sensitivity of the CCD, their potential, even for research on bright stars, seems to be limitless – once the data are accessible, ideas grow that may never have occurred to present-day scientists, and were certainly very far removed from the thinking of whoever made the original observations.

REALISING NEW POTENTIALS

Quite apart from the unquestionable usefulness of extending backwards in time unique evidence of change or of period modulation in celestial objects, there are also possibilities for cross-disciplinary research. One such project currently being pursued at the DAO involves determining concentrations of ozone in the Earth's atmosphere in the years that long pre-dated widespread ground or space ozone measurements with purpose-built equipment. One problem haunting the quantitative interpretation of the recent decline in ozone abundances is the lack of solid evidence regarding natural trends and levels. Certain historic stellar spectra showing the far ultra-violet can be used for that purpose, and the project being pursued at the DAO

(Griffin 2005) selects stellar spectra from mid-latitude observatories dating from 1930–1960. Because one cannot search an inventory digitally it has been necessary to make personal visits to numerous observatories and to carry plates back to Victoria for scanning. Nevertheless,

every new piece of information thus gleaned is another valuable data-point in the all-important search for trend information; it then remains for the high-speed computer to incorporate those historic observations into analyses based on more modern data-sets.

WHO PAYS?

Identifying the resources to carry out these rescue missions is of crucial importance. With Canada now looking very seriously at ways in which it can federate observational data from whole disciplines of scientific research, and placing some emphasis on securing a position in the forefront of scientific endeavours, this might be a good moment to plead for the necessary funding – which, incidentally, is small fry compared to the cost of modern-day scientific equipment. Even preparing a digital inventory of Canadian plate archives still awaits an award of the necessary funds.

THE STATUS OF A DATA ARCHIVE

A final word takes a critical look at the attitudes that we, as observational scientists, apply to the preservation and maintenance of data, both historic and recent, for use and re-use by ourselves, our peers and our successors. In a recent discussion with prominent astronomers, I mentioned that even new telescopes, which cost such large sums that they have to be funded through international consortia, still put a data archive very low down on the list of priorities and that when (as seems inevitable) a cost-overrun occurs the data archive is the first item to get pushed off the bottom. The rather surprising response upheld that "if you have to choose between a data archive and a new instrument, of course you'd choose the new instrument". The point, surely, is that a data archive IS an instrument, of the broadest and most common-user design imaginable, and the sooner we can come round to regarding it in that light the more successful our scientific endeavours are likely to be in the long term.

REFERENCES

Griffin, R.E.M., (2002) The need for contents catalogues, *SCAN-IT Newsletter*, 1, 24
Griffin, R.E.M., (2005) The detection and measurement of telluric ozone from stellar spectra. *Publ. Astron. Soc. Pacific* (in press).