Data At Risk

Elizabeth Griffin
Herzberg Institute of Astrophysics
5071 West Saanich Road, Victoria, BC, V9E 2E7, Canada
Elizabeth.Griffin@nrc.gc.ca

ABSTRACT

Important scientific research depends upon our ability to recover and re-use observations made years, decades, perhaps centuries, ago. Many "historic" data can be critical for studies of long-term trends, while some science can *only* be done if those observations, mostly of which pre-date the digital era, are accessible. Present-day archiving procedures are only now looking seriously at the recovery of relatively *recent* data, almost all of them "born digital". The correct recovery of scientific information from pre-digital ones presents a totally different set of challenges.

For those pre-digital measurements, the design of an information rescue depends on the kinds of data, their location, condition and potential, and – of course – on available resources. Few sciences have a commendable record in pre-digital data recovery, although some individual ground-breaking efforts set a good precedent. Data which are considered "at risk" in this context are mostly "hard copy", and include photographic observations, hand-written charts and records, magnetic tapes with no meta-data or format information, maps and specimens for which the concept of "information" is not easily quantifiable. Reliance on hard copies is misplaced: in contrast to electronic data they are unique, perishable, and cannot be entirely reproduced.

We discuss the urgent need to seek out, protect and ultimately access the information in "data at risk". Solutions are as varied as the data; even the blanket answer, "to digitize", has quite different interpretations, depending on the type of data. We describe the start being made on the problem by a focussed CODATA Task Group. We also outline solutions both current and in design for astronomical (photographic) data, and mention projects in a few fields not covered by other contributors to this session.

BACKGROUND

Much important and topical scientific research needs to incorporate observations made years, decades, perhaps centuries, ago. Studies of significantly long-term trends can *only* be carried out if those observations are accessible electronically, placing new emphases upon our ability to recover and re-use historic data regardless of their current forms and formats. The full and correct recovery of scientific information from observations which pre-date the digital era (and which are mostly the ones we loosely describe as "data at risk") present unique challenges; many scientific fields have not yet addressed at all, or not at all thoroughly, that aspect of their valuable legacies, although some individual ground-breaking efforts have set important precedents.

The tasks involved in rendering pre-digital data fully accessible in electronic form are fraught with unknowns and difficult decisions, and the work required can be thankless, sometimes unglamourous, often tedious and always varied, whether it involves deciphering hieroglyphics on decaying paper, keying-in hand-written information from inadequately completed log-books, or reconstructing procedures whose designers have passed beyond and taken their blueprints with them. It also requires a relatively small but non-zero amount of resources. The resources will surely become available eventually when the scientific needs shout loudly enough, but we

are not yet even half-ready to answer the calls because present-day archiving procedures are only looking seriously at the recovery and preservation of *recent* data, the "born digital" ones.

WHICH DATA ARE PARTICULARLY AT RISK?

The "data at risk" as defined in this context are scientific observations which are not in a format that permits full electronic access to the information which they contain. An "observation" is a record describing an object or event; it might be a direct image (e.g. of the sky) which also contains photometric and positional information, an image of a spectrum from an object, an actual 3D object (e.g. plant or seed), or some description – numerical, graphical or descriptive – of a measurement of an event or sighting (e.g. rainfall amount, temperature of water, percentage of cloud-cover, bird migration – we can all add other examples). Most of those observations pre-date what we call the digital era. They may be inherently non-digital observations (e.g. handwritten; photographic film or plate), or on near-obsolete digital media (such as magnetic tapes), or insufficiently described (lacking meta-data). For others, e.g. plant specimens, the concept of information in electronic form is meaningless except by imaging the originals, though their meta-data are absolutely essential. Some born-digital data can also be considered "at risk" if they cannot be ingested into managed databases because they lack adequate formatting or metadata. The risk common to them all is that data which are popularly regarded as *unuseable* tend to be regarded as *useless*, and may therefore be destroyed.

THE SCIENTIFIC CASE

The phrase "long-term trends" has crept into most of the natural sciences, but the implied time-scales vary enormously with discipline, from the æons connecting geological events to the changes of a few days in the evolution of a recurrently explosive star. While there is no question as to the scientific value that can be derived from a detailed study of a single object or event, the science becomes substantially more rich if we can study more than one image of the same object or event separated by a significant time-interval; the *differences* can teach as least as much, and usually a good deal more, than the study of a single event. However, the scope of such studies will inevitably be dictated by what observations are available. Modern astronomy, for instance, has well-managed archives of recent *electronic* observations, but few extend backwards in time for more than about 15 years. Astronomers are therefore able to develop a keen measure of phenomena that occur and re-occur within the span of a few months to a few years, but of the phenomena which happen on time-scales that exceed the span of their born-digital data archives they unfortunately know relatively little, and their science is consequently biassed.

Long-term trends are both significant and topical in natural phenomena like bio-system evolution, climate changes and environmental changes, whose time-scales may be commensurate with a human's life-span and can therefore only be studied through archived data of appropriate age. Few of those sciences (botany may be a noble exception) were making organized discipline-specific observations a century ago, so opportunities for *trans-disciplinary* applications of the data suddenly become important, making even more imperative the need to transform historic ones into electronic forms. However, when those transforms *can* be carried out correctly, the scientific gains are significant. The following represent just a small selection of success stories in different sciences, and were chosen because of the broad applications of information that were made possible through dedicated efforts in a few particular areas.

Ozone in the Earth's atmosphere: A good example of trans-disciplinary application of data was demonstrated by studies of the Earth's ozone layer [1,2] as determined from historic spectra – not of the Earth's atmosphere, but of stars whose light unavoidably passes through that

atmosphere and thus carries certain signatures of it. Astronomy's historic stellar spectra are presently only available in photographic formats and require specialist processing in order to yield electronic versions for modelling purposes, but they constitute an otherwise untapped source of observations which pre-date most of the equipment that first targetted ozone measurements *per se*. The newly-digitized stellar spectra are of course also available for modern astronomical research too.

Medical statistics: The scientific value of "lost" medical data was realised by the recent re-opening of a statistical study of heart disease and diet carried out 50 years ago and recorded on Hollerith punch-cards. Though the results were scorned at the time, digitizing and re-analysing those punch-cards has provided a goldmine of information for modern medicine [3].

Water flow-rates from a modified environment: Hand-written measurements of historic flow-rates in a stream descending from mountains near Cape Town (South Africa) were painstakingly transcribed into an electronic database [4]. The mountains had recently been clear-cut and replanted with non-native trees, and reference to the old records demonstrated that the water take-up of the newly-planted species was significantly greater than for native trees. A reduction in the supply to Cape Town's reservoirs was thereby explained; the economic pay-off from such knowledge dwarfed the salaries of the few who keyed in the flow-rates for analysis.

History's measurements of the ocean: The Global Oceanographic Data Archaeology and Rescue (GODAR) project, inaugurated in 1993, has retrieved and digitized millions of historic, non-electronic oceanographic data for use in research. The newly derived profiles of chlorophyll and plankton content and ocean temperatures that have resulted are proving vital in studies of climate change. The overall success of the project has commanded new funding for future phases.

Keeping better tabs on the weather: For the last 10 years NOAA, through its through its Climate Database Modernization Program (CDMP), has been improving climate and environmental information by including newly-digitized historic data alongside current ones. CDMP encompasses about 100 individual projects, and has seen through the processing and digitization of millions of records. NOAA's Climatic Data Center also performs research in Paleoclimate data from natural sources such as tree rings, ice cores, corals, and ocean and lake sediments. Together, the on-line archive of weather and climate information (the largest in the world) now extends back across centuries, even thousands of centuries.

Recovering "digital" data: The US National Archives and Records Administration (NARA) was tasked with the time-consuming challenge to recover the information about numerous scientific programmes and missions from 9-track magnetic tapes. Data from over 800 tapes were transferred to an electronic database. Most could be recovered with complete success, though in about 10% some data were lost through hardware faults that could be attributed to ageing in the magnetic-tape media.

Bird sightings, and the value of e-literacy: Bird-spotters can nowadays enter their sightings into an "e-Bird" database. Ornithologists are using those observations to map bird migrations and thence to fathom the causes for change, in a way that is quite unprecedented. "Citizen Science" may play a significant rôle in the recovery of data from all areas of natural science that are presently in individual holdings in the private sector. If all those historic observations of nature could be brought online, correlations with climate and environmental change would be on a much firmer footing than is presently the case.

INFORMATION RESCUE

Incorporating those non-electronic measurements into modern research requires appropriate "information rescue", probably at more than one level. The observations themselves, if on a limited-life physical medium such as film, paper or tape, need servicing to retard, even if not prevent, natural ageing processes; meta-data may need to be entered into a searchable catalogue, and the actual observations – whether handwritten numbers, maps, specimens or images such as photographed spectra – must be treated correctly so as to copy their scientific information faithfully into an electronic form which can then be managed like any other modern data-set. It is, of course, impossible to be specific; a bunker the size of a concert hall containing cases of plant specimens can have just as much scientific potential as a box of unique photographic spectra or a floppy disk containing unformatted binary characters.

The rescues that need to be designed are type-specific, application-specific, discipline-specific and probably location-specific, and often require expertise which may only exist in certain laboratories or groups and may not be widely shared. There can be no "one size fits all" design for information rescue because the concept of "information" is not easily quantifiable, though some of the individual challenges may share commonalities. Even the blanket solution, "to digitize", has a plethora of different interpretations, implying a correspondingly numerous set of tools and skills. The one feature which they all share is the (possibly) desperate need to achieve a rescue somehow, especially when a reliable scientific solution will *only* be reached when those legacy data are ingested electronically into a scientific model.

ASTRONOMY'S LEGACY IN THE PIPE-LINE

For about a century, up to the adoption of the CCD detector in the 1980s, astronomical observations were recorded on photographic plates. Because the cosmos is changing, always slowly and subtly, often periodically too and sometimes quite violently, each new observation adds to the history about an object or a region of sky. The older observatories preserved their photographic observations in a "plate store", and it was common for researchers to borrow plates, either from their own observatory or from some other one. The modern demands of digital methods and modelling mean that the photographic data cannot be included, even when they should be for scientific reasons; very few observatories have the necessary scanning equipment that can reproduce the grey-scales in the emulsions faithully enough, and fewer still have technicians who can upgrade and operate such scanners. As a result, plate stores have closed, some collections have been transferred elsewhere to release needed space, and the whole legacy has dwindled into a no-man's land of unfrequented history.

However, even those times are finally changing. Thanks to a few brave and successful initiatives, of which the most cited has to be DASCH[5] – the design and construction of a special instrument to digitize the world's largest collection (some 650,000) of astronomical plates at Harvard College Observatory – the immense potential which can be realised by creating a public database of historic observations is at last gaining recognition, its scientific rationale unquestionable. A younger counterpart at the Royal Observatory Brussels is also now in production (Pauwels & De Cuyper, poster paper), while at the Dominion Astrophysical Observatory in Victoria (Canada) the venerable but fully updated PDS scanner (*the* purpose-built instrument for digitizing astronomical plates, and once quite common observatory equipment) has been pressed into service to digitize and reduce selections of the heritage of 93,000 plates from the DAO 74-inch telescope (dating back to 1918) and 16,500 plates from the 48-inch telescope (inaugurated in 1962). Some observatories, particularly in eastern Europe, have been making scans of their photographic images using flat-bed scanners, though there are reservations about

the quality of the output for quantitative work.

It only takes a few chance events to make such projects mushroom – the unique opportunity to learn what a star was like before it exploded as a supernova, or the refinement of the orbit of a Near-Earth Object which will not, after all, collide with Earth; moreover, the expenditure of human resources is trivial compared to what is lost in the noise in other projects. Astronomy's observations can be classified very neatly into "image" or "spectrum", so the methodology, once developed, can be cloned elsewhere. However, a major bottleneck at present is knowing what the plate archives contain. Though observations were almost always recorded in official log-books, and card-index catalogues were created at least at the larger observatories, there are no electronic versions of those log-books, so there is (at present) no way that a researcher can search and decide what scans to request.

THE CODATA TASK GROUP

In order to design a comprehensive Information Rescue effort we need to know at the outset the measure of the challenge: roughly how many observations, what types, what age, where they are and how real are the threats of losing them for good. It is only when we know the true dimensions of the beast that we can design an efficient package of solutions. To that end, in late 2010 CODATA set up a Task Group on Data At Risk (DARTG)[1], its prime objective being to make an Inventory of the observations which are to be rescued. DARTG presently has 16 members, many of whom are practising scientists with experience in data acquisition and management. The intake of data for the Inventory is presently limited to the scientific disciplines which the DARTG members represent, but that is nevertheless quite broad: astronomy, space science, geology, climatology, meteorology, oceanography, biodiversity, geomatics and cartography. The Inventory will become the reference document for Phase II, in which the actual rescue efforts are to be designed with a view to optimising efficiency and effectiveness.

By working through the necessary steps to achieve its objective, DARTG will demonstrate an approach, process and practices for building an extensible Inventory of highly disparate, endangered scientific observations from a multitude of sources, and will contribute uniquely towards effecting the rescue of their scientific information content.

SPREADING THE WORD

In its quest to accentuate the need to be protective of the scientific content of fragile data, DARTG is compiling literature describing new science which has emanated from analyses of rescued historic observations. Contributions on any topic to that bibliography are welcome.

"Data At Risk" is the theme of Session 7 at PV2011, and the TG hopes that by advertising what it is attempting to do it may enlist the interest and support of the wider community. Indeed, some aspects of its mission cannot be completed *unless* that wider community support is tapped. The papers that follow, plus the poster papers, illustrate some of the points made above, by describing the actual recovery of nearly-lost data and by discussing the basics of meta-data descriptors which alone have the fundamental property of making a set of data intelligent to other researchers.

REFERENCES

[1] – Griffin, R.E.M.: The detection and measurement of telluric ozone from stellar spectra,

[1] http://ils.unc.edu/ janeg/dartg/

Publications of the Astronomical Society of the Pacific, 117, 885–894 (2005)

[2] – Griffin, R.E.M.: Measuring Terrestrial Ozone from Historic Astronomical Spectra. The Physics Teacher, 47, 22–27 (2009)

[3] – Krotz, D.: Newsletter, Lawrence Berkeley National Laboratory (January 2011)

[4] Goodall, V. Allsopp, N.: Jonkershoek – preserving 73 years of catchment monitoring data. Paper presented at CODATA annual conference (2010)

[5] See http://hea-www.harvard.edu/DASCH/