Recovering the Forgettery of the World

Elizabeth Griffin and the *CODATA* DARTG Team
Dominion Astrophysical Observatory, Victoria, Canada

Abstract
This Session of the conference examines an aspect of scientific data from a somewhat unusual perspective, by focusing on observations that are not, or not obviously, a formal part of the memory of the World. Almost exclusively those data are historic records; they complement modern ones in a way that can be crucial for accurate studies of long-term trends. A CODATA Task Group, "Data-At-Risk", is creating an Inventory of such data that catalogues their locations, types and volumes, and risk levels, as a prerequisite for designing efficient programmes to rescue the scientific information that they contain. The rationale for the CODATA effort is rooted in the knowledge that heritage data can offer unique evidence for solving some of the most pressing scientific unknowns that the world is currently facing. This Introduction examines that rationale, and the presentations that follow illustrate different types of data that are currently being rescued and brought into the public domain.

Author
Elizabeth Griffin is a volunteer researcher at the Dominion Astrophysical Observatory inVictoria (Canada). Since her PhD at Cambridge (UK) in 1966 she has researched astrophysics at Cmbridge, Oxford, Antwerp, Brussels, Toulouse, Boulder, Toronto and (now) Victoria. She Cairs an International Astronomical Union Task Force for preserving and digitizing astronomy's heritage data, and also Chairs the CODATA DARTG mentioned here. One recent focus of her research has been the recovery and re-use of historic data for trans-disciplinary studies.

1. Preamble

The Memory of the World is a superb resource for scientific research that explores changes in the biosystems and the physical conditions of our planet. Many natural sciences are needing to call upon that memory nowadays in order to understand and quantify, in particular, the sorts of *long-term* changes that not only biosystems but also our physical world - its oceans, air, glaciers and deserts - are experiencing. Many of those changes appear to be substantially more radical than anything recorded during the last century, but the baseline observations which are needed for assessing the reality of trends are unfortunately not readily available for inclusion in research. Modern analyses require data to be electronic, so records that are not in that format are effectively inaccessible. That immediately excludes almost all historic data, because they were recorded on non-digital media such as paper, film, photographic plate, books or undocumented magnetic tape, and have never been transformed from that

state. As a result, the very observations which are most needed to determine long-term trends cannot currently be incorporated into relevant research programmes. Modelling has become highly sophisticated, and increasingly efficient at representing the data that are supplied, but are not able to go beyond those limits with any confidence: extrapolation is unreliable in a chaotic world, and the only reliable solution is to capture in electronic form the information that is latent in so many of those pre-digital data. Though ambitious in many respects, as discussed below, such a programme will extend the publicly-accessible memory of the world backwards in time to encompass the decades when anthropogenic interference was far less vigorous than it is today.

But there are many challenges. Some historic data are deemed unuseable, some are physically degrading, and all risk being destroyed through ignorance, even though the time-spansnwhich they represent may be crucial for current research. This article introduces the topic and examines the general situation and outlook, while the following ones explore the matter from a number of different angles.

2. Two Hemispheres of Scientific Data

The power, capacity, speed and capability of modern computers has changed radically how today's researchers tackle large quantities of data and address major scientific problems. These developments are relatively new. A dozen or so years ago it was a tough challenge to import, let alone merge, data from disparate sources in order to produce a comprehensive model or extract a trend; today, current technologies are able to recognize patterns, correlations and correspondences between data that may have quite different fundamental properties.

These advances are timely indeed. The observations which are central to research into the natural world have become increasingly detailed as technology has evolved. Many modern data sets are large and complex, and their analyses involve interactive adjustments to complicated models that may require supercomputing power. The results that emerge can be used to predict how situations will change in the future. In principle, if the models are sufficiently reliable they can also be driven backwards to reproduce past conditions over a time-span that is significant compared to the recent changes that are irrefutable in an alarming number of situations. In practice, however, an extrapolation cannot have the degree of dependability that is needed, owing to the element of chaos in the natural world; the wide range of contributing factors and the subtle interplays between them can trigger the unexpected quite legitimately. A model may therefore only be valid for the time-span of the data from which it was generated.

Investigating the causes of long-term trends in the natural world has become urgent science. If recent changes are due to anthropogenic interference, the sooner each situation is understood and quantified the more likely it can be halted and even reversed. Determining the causes therefore needs observations from the decades when those anthropogenic interferences were much less widespread and vigorous than they are today. However, that immediately raises a problem, because few digital data-sets date back more than 25 years, and managed archives of most electronic data are even younger. The historical observations that are so badly needed are not in electronic form, so cannot be ingested into the modelling analyses without some specific transformation procedure. No matter what type of historic observations are called for, be it concentrations of stratospheric ozone, recordings of ocean temperatures, amounts and distribution of rainfall, counts and sizes of sunspots, bird migration patterns, frequencies of extreme weather events, behaviour of marine organisms, or a veritable host of many similar daily, seasonal or occasional events, if they were recorded on non-digital medium they will require transforming, and many may also need specialist knowledge to decode or interpret them correctly. It is a

serious concern that the very data which contain vital clues to a correct understanding of how and why our natural world is changing are inaccessible to analysts, and will remain so without some major effort. They are poles apart from modern, all-digital data and managed data-bases. They are almost "lost" to the scientific world.

Fashion also plays a troublesome role. Fashion has dangerously close links with consensus, and what tends to win research funding is what captures majority interests or promises prestige for the team, laboratory or country. The resultant polarizations of data have been seriously unhealthy for the sciences that need to access historic data, and if some of those observations get discarded for whatever reason, the science loses unrepeatable measurements. Nevertheless, the pendulum of fashion does swing back. The medical researcher, whose painstaking survey in the 1960s of patient cholesterol levels was scorned at the time, had fortunately stashed away the packs of Hollerith punched cards bearing the acquired measurements, and when they were recently re-discovered they were recognized as a gold mine of long-term data.

There is clearly significantly more to recovering heritage scientific observations than keeping artefacts from the past for their antiquarian interest. The physical form and format of the observations is of little direct relevance, except in understanding their quality and limitations. The medium is not the message; the scientific value lies primarily in the date-stamp associated with each measurement. Delving back into the past is a route to instant science, and while historic measurements may lack some of the detail of modern-day ones their uniqueness in time, and the possibility that they will supply actual records of the behaviour of a given property throughout an extended time-base, renders them of unrepeatable value.

While sterling attempts are been being made in some sciences to make data rescue a major feature of the research effort, in others (and for all its advantageous infrastructure of sorted and catalogued heritage, astronomy is perhaps one of the worst culprits), the rescue efforts are left to the passionate amateur who feels deep concern at the possibility of losing historic data and who is prepared to volunteer time and effort to doing what is required. Though the efforts of the amateurs are absolutely essential, an unfortunate side-effect is to exacerbate the existing polarization between the new and the old.

One basic lesson to promote is therefore the *complementarity* of old and new: they may add information of differing quality, but interpretation relies upon both. Education will certainly help, but real-life examples of new science that was *only* possible with recovered data can be vivid teachers. What needs to be done?

3. The Dangers of Doing Nothing

There is no doubt that scientific evidence which is crucial for understanding the impact of anthropogenic interference upon planet Earth and its countless biosystems will be found in historic data. Lack of such evidence has allowed arguments that are based on opinion rather than fact, and the confusion that currently surrounds efforts to ameliorate certain situations owes a lot to news sensationalism. The longer we wait for the true facts to emerge, the harder it will be to halt the situation, or reverse it if that is what is required. But not all the problems are on a global scale, nor are all of the solutions enormously difficult to reach. There is the tragedy of the African farming family, newly settled in a region and intending to set aside extra food the following season to cope with the droughts which - they heard from somewhere - occurred every 10 years, but faced severe starvation because the actual period of the drought was 6 years, and was about to happen; the weather records were not suitably accessible. In a quite different setting, the

environmentally-damaging effects of afforestation with non-native species in the mountains that are sources of Cape Town's reservoirs only came to light when a small group of researchers digitized 73 years' worth of paper recordings of stream flow and uncovered a strongly positive correlation. Both examples have much to teach. In both situations the observations which could furnish essential knowledge existed, but were not accessible. The information that was needed was not complex and not overwhelmingly large to handle, nor was complicated machinery involved at any stage, just a focussed human effort. Doing nothing had inflated expense, and ruined lives.

Not infrequently, scientific measurements find application in more than one field, even in different disciplines, though it may not be the same analysts who recognise alternative uses, and the analyses need not be simultaneous. or even close in time. Providing access to historic data in electronic form can yield unexpected side-benefits that are often far removed from the original purpose of the observations, so doing nothing denies opportunities for those initiatives. The astronomers who observed the spectra of hot stars in the near ultra-violet in the 1930s in order to study interstellar absorption features had no idea that their observations could be used to deduce the concentrations of the Earth's stratospheric ozone, and had they been able to bequeath fully digital spectra then, the data would have been seized upon at once by atmospheric scientists. Since digitization was not then an option, the spectra were left in their virgin state (but at least preserved), and it was three-quarters of a century before that ozone research could be carried out. Somewhat similarly, meteorologists are recognizing gems in the weather logs kept routinely by submarines or astronomical observatories, and atmospheric scientists see patterns in crop yields that tell as much about El Nino events as grain performance. We are now in a position to perform good-quality electronic transformations of all such records, but important background information will be harder to capture as time passes and less human memory can be tapped for the properties of the original data and their specificity. While it should not be assumed that *all* the historic information which might in principle be made accessible will contribute to a positive solution in some field, it is a sobering reflection that lives could be saved, deleterious situations avoided and damaging traits repaired by taking account of the huge wealth of heritage observational material that is presently in the world's "forgettery".

4. The Forgettery of the World

A forgettery is part and parcel of the human psyche, and a highly efficient archiving system for the brain to stash away images that may hurt, or surplus facts that are very infrequently wanted. It is also the convenient location for things unwanted; progress can only feel good if there is a forgettery to handle the casualties that got in the way. So with institutional science: technology and change are closely coupled, and new ideas should be seized while young even if the older technology and its output are not fully wound up. New technology engenders new expertise, but moving personnel from the older to the newer has the unfortunate repercussion that rather little expertise and resources remain for curating the data which were so central to yesterday's research. The world has an enormous "forgettery" of scientific records that it might once have kept for a purpose but has almost forgotten what that was. It has physical locations in attics, cupboards, closets, archive warehouses, private journals ... The challenge for today is to bring that portion of the world's forgettery into the world's growing memory, where it can offer irreplaceable contributions to today's scientific research.

Why is that still waiting to be tackled? A comprehensive programme to activate those forgotten pockets of the world's memory may seem dominated by practical difficulties, including dealing with unfamiliar data from unfamiliar equipment, and the necessary expertise may be diverse and difficult to

locate, but all that is surely less challenging than (say) launching satellites to make observations of the sky or the earth, and certainly less costly. What it *will* require is a united determination to overcome the inertia which has trapped so much valuable material doing nothing in a forgettery, so the first step is to publicise and educate. That was the basic rationale for this Special Session at a UNESCO conference.

Recognizing the seriousness of the situation, CODATA (the *ICSU* Committee for Data in Science & Technology) has mandated a Data-At-Risk Task Group (DARTG) to seek out sources of non-electronic data, and create an Inventory[1]. Time is not on our side; some of the materials to be recovered are already becoming unreadable - -deteriorating magnetic tapes, recorded data without sufficient meta-data (information about the information), photographic records that have been deprived of the essential equipment to measure them correctly, or hand-written sheets whose ink is fading. CODATA is clearly late in arriving on the scene, yet the two essentials for a constructive data-rescue programme - a recognition of the need to study long-term trends (the "why") and the capabilities of technology to manage the necessary tasks (the "how") - have only recently moved close enough together to promise truly worthwhile returns for the efforts to be invested. The world's forgettery can yet be rescued and put to very effective use.

5. Stages of Data Rescue

The requirements of a programme to rescue historic information are varied and demanding, but not unsurmountable. Each calls upon a number of different expertises, beginning with the history of the relevant experimentation, equipment and associated personnel, and can ultimately involve a broad selection of different groups and skills, so each needs to be sympathetic to the overall objectives. The challenges of data recovery on a scale to return significant science is beset with fascinating problems that are rather rarely encountered in modern programmes. Discovering what is out there - somewhere, coping with the condition of what one finds, and arguing for the costs as part of a modern programme demand skills that involve communication, archival techniques and human interactions, and all need to be handled successfully in order to overcome the prejudice that is sometimes encountered in modern research, *viz.,* that anything that is old must be inferior. Deciphering notes in observation log-books is often highly domain-specific, and the best minds may no longer be around. Understanding the purpose of an experiment, and through it the limitations imposed upon the observations and thereby of their interpretation, will entail recourse to old publications and reports, few of which may be openly available (either printed or electronically). Converting the necessary elements of an observation into a fully-transferable electronic record demands a finesse that cannot be over-stressed, while the required meta-data (e.g., as in FITS headers) for seamless ingestion into a modern database are *sine qua non.*

Each step requires a clear vision, unambiguous procedures and well-tailored programmes with milestones, benchmarks and templates. In reality the data that one may get to work with will be in assorted condition, and each will need specialized treatment at some level, implying devoted resources. Each stage can produce unexpected challenges; it is hard to be fully prepared, and setbacks should always be allowed for in the planning time-line.

[1] http://ibiblio.org/data-at-risk/items

5.1 Tracking down "lost" data

The biggest hurdle is undoubtedly the initial one of communicating the incommunicable: getting researchers to discover what was put in storage probably long before their own arrival at the laboratory, observatory, library, archive or bureau in question. One approach is to follow up leads through observations that were known to have been made; those can result in unexpected discoveries too. Another is to ask people to go systematically through their storage areas and list materials that have been there for more than a specified period. Activating the trans-disciplinary potential of scientific data can sometimes trigger discoveries too; enquiries about weather records (for instance) may jog an astronomical observatory into reviving what *it* regarded as routine records for internal use only, but which may in fact be significant for environmental studies. At the other end of the scale are records generated by projects in Government laboratories and which may be protected for a period from actual destruction, but are regarded as space-wasters and are sub-optimally stored. Such caches may be the archetypal dishevelled heaps of paper, getting more scruffed when shifted and in constant peril of an order to "clear out" the cupboards or rooms when a certain age limit is reached. We may never know what potentially important records have already been lost in that way, but can be sure that it has happened and will happen again unless their scientific potential is understood. Moving a department into new facilities may have unfortunate side-effects if old records are regarded as mere trash for disposal, though it is often during such relocations that real finds are made.

5.2 The state of discovered data

Some heritage materials are formally maintained in tolerably good condition, and include most of the necessary meta-data (possibly as a catalogue), a description of the purpose and source of the data, and pointers to publications that may have ensued. "Records" may be original observations, in a variety of forms and formats, or may be the information from those observations transcribed (most frequently on paper or some type of magnetic tape) as "measurements", either freestyle or in a pro-forma layout (e.g. a printed chart with blanks to be filled in regularly).

Most of astronomy's worldwide collections of 3 million or so photographic plates, some dating back over 100 years, are in passable condition. But to make the data publicly available electronically requires specialized digitizing procedures entailing purpose-built equipment and trained personnel, and the resources needed for that are hard to find. In a somewhat parallel situation, the US National Climatic Data Center's data modernization programme has a huge basement filled with files of hand-written weather records from worldwide sources. It will take a major effort to put the all information online, but the records themselves are in good condition. The Berlin botanical museum presents a different set of problems: a huge underground store houses *samples* of plant parts rather than observations of them; the archive is "live" inasmuch as new specimens are constantly being added, and the challenge is to capture electronically all the observations of the specimens, not only the meta-data but also the measurements have been carried out on them but never formally published. The very diversity of even these straightforward cases demands astute planning and design. Informing that planning is an immediate objective of DARTG's Inventory of data at risk.

An example of well-stored but almost abandoned data is the collection of forms recording ozone measurements made in Oxford (UK) from 1933-57. Neatly bundled and ordered, the set had been almost untouched for half a century; manual transcription of the information and expert analysis filled in a major gap in recorded patterns of ozone changes during the last century.

In contrast, the IEDRO[2] video, *Historic Weather Data Rescue and Digitization,* displays untidy heaps of papers containing weather records, lacking even basic sorting or classification. Such heaps may get moved on, and with each move comes a loss of specific information regarding why they were saved, by whom, what they should contain, and where they originated. If the heaps (or boxfulls, closed containers or locked store-rooms) have been left undisturbed for a long time they might be found in a comparatively virgin state but deteriorating physically - ink fades, paper turns brittle, crumbles, or gets attacked by mites or mildew, photographs discolour, emulsions become brittle and lift from their substrates, films crack, magnetic tapes oxidise and cannot be read - those are just some of the conditions that may await the investigator. A more unusual case of actual data loss is the sets of bolometric solar scans recorded on large glass plates from 1926-31 at a mountain site in Namibia; modern researchers could have mined them for information about variability in atmospheric constituents, but the plates were left abandoned on site and some have since found new life as window-glass in Namibian homes (after the offending wiggly lines were removed ...)

5.3 Influences of human attitudes

Natural ageing processes, even when speeded up by poor storage conditions and fungal infections and the like, are relatively slow compared to irreversible decisions by humans to jettison a collection because the space they occupy is wanted for some new activity, and it is sad to reflect how often the fate of such records is actually decided by ignorance. Collections have been destroyed "because no-one uses them", but would they not be used if the information they contain were available electronically and the significance of the date-stamps properly appreciated?

The way scientific research is funded tends to exacerbates the problem. The concept of pushing a boundary backwards in time by digitizing heritage observations does not command the same prestige as building new equipment to carry out observations of world-breaking class, even though the latter projects are almost always the more expensive by a wide margin.

6. Pushing Forward

In order for its endeavour to succeed, DARTG needs to demonstrate and publicise the unique role which historic data throughout the natural sciences can very often play once they are made widely accessible in machine-readable formats. The Inventory which has been started will demonstrate the extent and types of *known* data that need to be the foci of rescue projects, whether for preserving, cataloguing or actual digitizing, and the more complete it is the better it can demonstrate the relative urgency and fragility of the various entries. DARTG's ultimate objective is to extend the memory of the world backwards in time through a period that is scientifically significant, and conferences such as this can offer valuable platforms for broadcasting DARTG's message and engaging broad participation by the community at large.

The excitement of scientific research is always enhanced by the discovery of the unexpected, and historic data have already provided results that were never anticipated at the time of their recording, either by those who made them or by those who now recover them. Keeping an open mind for opportunities of a trans-disciplinary nature is therefore key, both for maximizing the return on the recovery effort and for extracting new science that can stretch the imagination far beyond expectation. It is not ours to judge whether the wisdom thereby derived is more valuable because it teaches how to reafforest hills, or keeps

[2]International Environmental Data Rescue Organization; www.iedro.org

isolated farmers in touch, or measures the earth's atmosphere through which we star-gaze, or clinches questions of climate change on a global scale. Every new turn in the zig-zag path of scientific progress opens new vistas that are essential aids in an international quest to harmonize with our parent planet.

This contribution has provided some background for the Session on "data at risk". The following communications describe selected examples of work in progress in a variety of disciplines.