Preserving and Rescuing Heritage Information on Analogue Media

Elizabeth Griffin (Chair, "Data At Risk" Task Group, CODATA)
Dominion Astrophysical Observatory, Victoria, Canada

Organizations like the NDSA, the RDA and others, which focus on the preservation and stewardship of data, are concerned with *electronic* data. But there are many more kinds of data that are also in need of stewardship and preservation because of the unique *historic* information which they contain, be it of science, culture or the humanities. This session takes a special look at the data (meaning records, measurements, observations) that were obtained in pre-digital times, when the routine recording medium was paper, books or photographs; some may have been recorded onto, or transferred onto, early magnetic tapes but in unspecified formats and without any meta-data. All are inaccessible for research unless in electronic form. They are "at risk" because they are ageing and deteriorating, but also (and which is even more worrying) because what is inaccessible and is not used can get tossed.

For the past four years the ICSU committee CODATA (the Committee for Data in science) has supported a Task Group to look into this whole matter of at-risk data, and some of the TG members are here today to present this session. First and foremost, though, we must explain **WHY** we focus our attention on these types of data. It is because their scientific information is unique. Although modern data quality, modern instruments and electronic recording techniques are all substantially superior to what could be achieved in the pre-digital era, they do not extend far enough back in time to offer a significant time-base; most electronic archives did not commence until the 1980s, and by that time the changes in (for instance) our environment that are being brought about by human activity were already under way. We need to go back much earlier, to the pre-digital data, and establish a reference base-line and to document any natural changes that were also occurring. Models may be constructed from the shorter base-line of born-digital data, but the natural world is chaotic and no model can be validated unless tested against actual observations.

Many data, both historic and recent, may also be repurposed for research that is quite far removed from the original contexts. In astronomy, for instance, certain types of ground-based observations harbour unique information about the Earth's stratospheric ozone, many dating back to the early 1930s when routine monitoring of ozone had only recently started, at only one site, and the measurements were still prone to instrumental "noise". The astronomical spectra are images on glass photographic plates, intrinsically non-digital, stored in observatories at many different locations worldwide that are all unmanned and innocent of any on-line catalogue. The investigative project which I undertook required me to visit those observatories in person, select what I could find by manual rummaging, and carry the plates back to Canada's only special digitizing microphotometer (in Victoria) that can deal appropriately with those sorts of photographic plates.

Then there is the case of the hydrology data in Cape Town, South Africa. By digitizing 74 years' worth of past records of stream-flow rates, a group of researchers was able to demonstrate that the reason why Cape Town's reservoirs were getting lower than forecast was not so much through human activity, but because the mountains where the supply streams sprang had been reafforested with the wrong kinds of trees. Elsewhere there are attempts to digitize public health records; set those alongside data on local demography and you have a unique data base to study how and why epidemics of disease have occurred, and to recognize new pointers for the future.

Attracting the necessary funds to carry out digitization on any sizeable scale is itself challenging, often because the pursuit of “historic” data seems unglamorous and unnecessary to those whose focus is on the bigger, better and therefore newer, but success breeds success: GODAR (the Global Oceanographic Data Archaeology and Rescue) won support for a pilot project, and the resulting information was so impressive that the project has been able not only to continue but also to expand.

These are the sorts of new science that emerge from old data, and are what drive our “Data At Risk” efforts, but the new science will only be realized fully if the old data can be located, digitized (or re-digitized) and understood correctly. Those steps may need to involve understanding the instrumentation that was used and the context of the original observations, and may need to tap into cultural knowledge of the period, location or science. Some of the kinds of peripheral hoops that one has to negotiate in these endeavours are now illustrated by the speakers of this session.