GeoResJ 6 (2015) 106–107

Contents lists available at ScienceDirect

GeoResJ

journal homepage: www.elsevier.com/locate/GRJ

Guest Editorial: Special issue Rescuing Legacy data for Future Science

Research and discovery in the natural sciences, particularly for documenting changes in our planet, is empowered by gathering, mining, and reusing observational data. However much of the data required, particularly data from the pre-digital era, are no longer accessible to science. The data are hidden away in investigators' desks on printed paper records, or are no longer readable as they are on deteriorating or outdated media, and are not documented in a way that makes them re-usable. Special initiatives are required to rescue them and preserve such data so that they can contribute to the scientific debates of today and those of the future. Data rescue efforts are key to making data resources accessible that are at risk of being lost forever when researchers retire or die, or when data formats or storage media are obsolete and unreadable.

Interest in the topic of data rescue is growing rapidly as researchers are realizing the potential loss of so much data from the scientific record. In the past few years, activity in, and acknowledgment of data rescue has steadily increased in the Earth Sciences, with events such as the International Data Rescue Award (co-sponsored by Elsevier and the Interdisciplinary Earth Data Alliance (IEDA) in 2013 and in 2015), scientific sessions and town halls at the American Geophysical Union and European Geophysical Union annual meetings, and the CODATA Task Group activity on 'Data at Risk'. In particular, the data rescue awards and this special issue are a result of the commitment of Elsevier, the data facilities and an international team of volunteers dedicated to promoting and ensuring the preservation of legacy scientific data so that they can be available for reuse and repurposing, often for use cases never considered by the person who originally recorded the data.

This special issue on "Rescuing Legacy Data for Future Science" was established to showcase and recognize efforts within the Earth Sciences that advance preservation, access, and usability of valuable research data. We encouraged contributions that describe data rescue efforts that included the product of a rescue mission; the accessibility, usability, and sustainability of the rescued data; the information value of the data set; and the workflow of the rescue process. Topics from all areas of the Earth and Space Sciences were considered.

Although the importance of data rescue is becoming recognized in the Earth and Space Science Informatics community, even we were surprised at both the number and the breadth of the submissions to this special issue. We believe that this response shows how critical the topic of data rescue is, both to the individual researcher who desires to preserve their personal contribution to scientific legacy, and the present-day researcher who requires adequately documented, curated, and accessible data for use in current research projects.

As this special issue shows, data rescue can have many facets, as described in the paper by Griffin on 'When are old data new?'. Traditionally, many think of data rescue as involving the conversion of old analog data (e.g., paper records, microfiche, books, charts, maps, photographic plates) into modern digital formats. The majority of papers in this special issue are about traditional data rescue and include papers on a diverse range of topics such as the Belfast Harbour tidal gauges (Murdy et al.); sea level data archaeology (Bradshaw et al.); flood disturbance data (Moody et al.); the preservation of geological survey data (Ramdeen, and Riganti et al.); data about zooplankton collected on research cruises (Wiebe & Allison); early satellite data records (Gallaher et al.); historic bedrock information (Fallas et al.); geomorphologic maps and information (M. Smith et al.), seismic shothole drillers' lithostratigraphic logs (R. Smith); and historic seismograms (Okal).

Today, data rescue is progressively involving the restoration of valuable early digital data on old and obsolete media. The term 'remastering' is used for this type of data rescue and is widely used in the petroleum industry, having been borrowed from the audio industry where it refers to quality enhancement of digital sound and film archives. Early digital data can be far more at risk than data that are only available on paper media, as the paper copies can be more persistent (but not always). For instance, data in scientific journals from the 17th and 18th centuries are still readable, but due to technological obsolescence, it can be difficult to find machines that can still read data stored on early tapes, disks, etc. from the 1950's to 1980's (e.g., there are very few computers around that are able to read data on an 8 inch floppy disk from the early 1970's). In many cases, the media used to record early digital data are physically degrading: some media are now so fragile that they simply disintegrate when attempts are made to read them by machine. Even if the media are still stable, and the data can be retrieved, the data rescue effort often needs to find the people who did the recording, as they are the only ones who actually understand these early formats and can make the data intelligible. The paper by An et al. describes an heroic effort to remaster an archive of nuclear explosion seismograms recorded in northern Kazakhstan during the period 1966–1996 whilst Diviacco et al. document a major initiative in Italy to rescue large volumes of vintage seismic data: both of these papers also involved translating old paper records to modern digital formats.

Increasingly, as the capacity of our computational systems continues to grow, the size of the data sets that can be analyzed is concomitantly increasing. Early digital data sets were broken up into very small file sizes, because of technology limitations; for example, satellite data recorded as individual scenes on tapes. To effectively mine these data sets, the data need to be not just

http://dx.doi.org/10.1016/j.grj.2015.02.017
2214-2428/

remastered onto more modern media, but the data have to be transformed into new seamless collections to make them more accessible and more useable to current and future communities: the paper by Purss et al. on 'Unlocking the Australian Landsat archive' is an example of this form of data rescue.

Another activity that can be considered a type of "remastering" from older to newer "media" is the migration of data and metadata from older out-of-date systems and websites, to those that are more current. This activity is happening countless times around the world as databases and systems must be upgraded, and Klump et al. describe a good example.

Long tail data, defined as data produced by individuals and small teams for specific purposes, is particularly at high risk of being lost to future science. Such data are often stored on local hard drives or in personal note books and not properly managed, documented or curated, and it is difficult for others to utilize these results in their research, particularly when this involves aggregating multiple small data sets, collected over many years, into cohesive collections. It is not just a matter of simply transferring the results onto modern digital media: many long tail data sets collected in laboratories are at risk because they have less established documentation standards and formats, whilst supporting information on analytical attributes such as calibrations and blanks, were traditionally recorded in analog laboratory notebooks. To enable their aggregation with other similar collections, not only do the digital analytical results have to be remastered and standardized, but the analogue laboratory notebooks have to be located, matched to the results, and then digitized. The papers by Hsu et al. and Wehmiller & Pellerito document how some old 'dark' laboratory data sets can be rescued so that they become ready to being integrated with equivalent, more modern, analytical data sets.

This special issue illustrates yet another form of data rescue whereby data collected for science campaigns of the past are carefully reinterpreted and/or repurposed for different use cases to improve reusability. This is shown by the work of Corbel & Wellmann and Grealish et al. Another example of improving reusability is synthesizing similar data into searchable and accessible data systems, as described by Collier et al. A study documenting the workflow for physical sample metadata rescue (Hills) also contributes toward reusability of data and metadata.

We believe that this special issue of GeoResJ will provide valuable reference materials for those embarking on data rescue projects for similar data types, and it may lead to communities of practice forming around the development and sharing of the best tools and techniques for rescuing specific data types.

We hope that the many ideas and techniques described in this special issue will inspire many researchers and data facilities to take on more data rescue initiatives in support of future science and realize that there are others who have successfully achieved what to so many is seemingly very difficult. Yet the valuable contribution to modern science that these data rescue efforts are now making cannot be underestimated, particularly for those studies that aim to document and quantify changes in our planet through time.

We would like to express our sincere gratitude to all the authors for contributing to this special issue, as well as the many reviewers for helping us in assessing the papers and giving constructive comments to the authors. We would also like to thank the Editor-in-Chief and the entire staff of GeoResJ for guiding us in a very organized and supportive manner through the editorial process.

Lesley Wyborn
National Computational Infrastructure, Australian National University, 56 Mills Road, Acton, ACT 2601, Australia
E-mail address: lesley.wyborn@anu.edu.au

Leslie Hsu
Kerstin Lehnert
Geoinformatics Center, Lamont-Doherty Earth Observatory, Columbia University, 61 Route 9W, Palisades, NY 10964, USA
E-mail addresses: lhsu@ldeo.columbia.edu (L. Hsu), lehnert@ldeo.columbia.edu (K. Lehnert)

Mark A. Parsons
Rensselaer Polytechnic Institute (RPI), 110 Eighth Street, Troy, NY 12180, USA
E-mail address: parsom3@rpi.edu