BOF: Dark Data and the Long Tail of Science

Lesley Wyborn1, Bryan Heidorn2, Andrew Treloar3

1Geoscience Australia, Canberra, Australia, lesley.wyborn@ga.gov.au
2University of Arizona, Tucson, USA, heidorn@email.arizona.edu
3Australian National Data Service (ANDS), Melbourne, Australia, andrew.treloar@ands.org.au

DESCRIPTION
The primary output from most scientific endeavours is data. Increasingly data is gaining importance as an output in its own right, which when properly curated and stored, can be aggregated and repurposed for uses other than what the collector of the data originally intended. To date, most data aggregation initiatives focus on large volume data sets from a small number of custodians. The larger a data set is, the more likely it is to be stored in homogeneous formats and standardised – ie ‘head’ data. At the other end of the spectrum, ‘tail’ data mostly comes from small scale scientific research initiatives, is characteristically heterogeneous, and is seldom reused and rarely properly curated. This BoF seeks to bring together those that have small scale collections of important data to share experiences on how such data can be effectively accessed and curated as seamless, homogeneous national scale data sets.

OUTLINE
This BoF will have three short presentations followed by open discussion.

1. Why long tail data needs to be at the head of the queue in data capture and preservation initiatives. Lesley Wyborn, Geoscience Australia. 10 minutes
2. Tales of dark data capture initiatives in the US and/or internationally Bryan Heidorn. 15 minutes
3. Initiatives by ANDS to stop Oz data disappearing into the dark Andrew Treloar. 15 minutes.
4. General Discussion led by Bryan Heidorn

EXTENDED ABSTRACT
The increased use of instruments, including sensor networks, is enabling the collection of large volume and increasingly higher resolution scientific data sets. Many of these are collected by airborne or satellite instruments and the resultant data sets constitute proxies of real world phenomena (eg remote sensing satellite images that can proxy for vegetation types). New petascale computational infrastructures enable enhanced capabilities in modeling and simulation of these large volume data sets (Big Science). However, to be of value many of these large volume data sets need to be calibrated by precise measurements of point located sample data. Unfortunately these observational data are small in volume and can be collected by many individual researchers as part of a multitude of sampling campaigns (Small Science).

The collection of large volume data sets is usually done by a few specialized, but well funded research teams who have to undertake good data management practices in order to be able to manipulate, share and reuse their data. In contrast, the data from many small science projects is termed ‘dark data’ because it is rarely indexed, stored and described so it can be reused. Often, once the research paper has been written the scientist rarely has the resources or the incentives to ensure that the underpinning data are preserved so that it can either be reused by others and/or aggregated into more significant national scale data sets. Most initiatives to preserve and store data tend to focus more on the large volume; homogeneous file based data sets which can be Petabytes in size. Although expensive on hardware, these large volume data sets are relatively cheap to develop software storage infrastructures that facilitate reuse and repurposing. In contrast, although small science data sets are in the range of Gigabytes, they are expensive to develop effective software data storage infrastructures that enable reuse and repurposing.

This BoF will discuss why dark data is increasingly important in the era of the data deluge which is perceived to be dominated by large volume data sets. The BoF will provide a heads up on International and Australian initiatives to deal with the increasingly complex issue of aggregating small scale sample based data sets into homogenous national data assets that can be reused and repurposed for use cases that original collector rarely considered.

REFERENCES

ABOUT THE SPEAKERS
Dr Bryan Heidorn is Director of the School of Information Resources and Library Science at the University of Arizona and the president of the JRS Biodiversity Foundation. Dr Andrew Treloar is the Director of Technology for the...
Australian National Data Service (ANDS) (http://ands.org.au/), with particular responsibility for the Applications and Metadata Stores programs. Dr. Lesley Wyborn is Senior Geoscience Advisor at Geoscience Australia. She is leader of the GA/CSIRO Collaborative eResearch Project and the GA/NCI High Performance Computing Pilot.