Geophysical Research Abstracts
Vol. 18, EGU2016-11686-1, 2016
EGU General Assembly 2016

The Big Challenge in Big Earth Science Data: Maturing to Transdisciplinary Data Platforms that are Relevant to Government, Research and Industry

Lesley Wyborn (1) and Ben Evans (2)

(1) National Computational Infrastructure, Australian National University, ACT Australia (lesley.wyborn@anu.edu.au), (2) National Computational Infrastructure, Australian National University, ACT Australia (ben.evans@anu.edu.au)

Collecting data for the Earth Sciences has a particularly long history going back centuries. Initially scientific data came only from simple human observations recorded by pen on paper. Scientific instruments soon supplemented data capture, and as these instruments became more capable (e.g, automation, more information captured, generation of digitally-born outputs), Earth Scientists entered the 'Big Data' era where progressively data became too big to store and process locally in the old style vaults.

To date, most funding initiatives for collection and storage of large volume data sets in the Earth Sciences have been specialised within a single discipline (e.g., climate, geophysics, and Earth Observation) or specific to an individual institution. To undertake interdisciplinary research, it is hard for users to integrate data from these individual repositories mainly due to limitations on physical access to/movement of the data, and/or data being organised without enough information to make sense of it without discipline specialised knowledge. Smaller repositories have also gradually been seen as inefficient in terms of the cost to manage and access (including scarce skills) and effective implementation of new technology and techniques.

Within the last decade, the trend is towards fewer and larger data repositories that increasingly are collocated with HPC/cloud resources. There has also been a growing recognition that digital data can be a valuable resource that can be reused and repurposed - publicly funded data from either the academic of government sector is seen as a shared resource, and that efficiencies can be gained by co-location.

These new, highly capable, 'transdisciplinary' data repositories are emerging as a fundamental 'infrastructure' both for research and other innovation. The sharing of academic and government data resources on the same infrastructures is enabling new research programmes that will enable integration beyond the traditional physical scientific domain silos, including into the humanities and social sciences. Furthermore there is increasing desire for these 'Big Data' data infrastructures to prove their value not only as platforms for scientific discovery, but to also support the development of evidence-based government policies, economic growth, and private-sector opportunities.

The capacity of these transdisciplinary data repositories leads to many new exciting opportunities for the next generation of large-scale data integration, but there is an emerging suite of data challenges that now need to be tackled. Many large volume data sets have historically been developed within traditional domain silos and issues such as difference of standards (informal and formal), the data conventions, the lack of controlled or even uniform vocabularies, the non-existent/not machine-accessible semantic information, and bespoke or unclear copyrights and licensing are becoming apparent. The different perspectives and approaches of the various communities have also started to come to the fore; particularly the dominant file based approach of the big data generating science communities versus the database approach of the point observational communities; and the multidimensional approach of the climate and oceans community versus the traditional 2D approach of the GIS/spatial community. Addressing such challenges is essential to fully unlock online access to all relevant data to enable the maturing of research to the transdisciplinary paradigm.