Geophysical Research Abstracts
Vol. 21, EGU2019-6255-1, 2019
EGU General Assembly 2019

Data Versioning Patterns and Principles

Jens Klump (1), Mingfang Wu (2), Gerry Ryder (3), Julia Martin (4), Lesley Wyborn (5), Robert Downs (6), and Ari Asmi (7)

(1) CSIRO, Mineral Resources, Perth, Australia (jens.klump@csiro.au), (2) Australian Research Data Commons, Melbourne, Australia, (3) Australian Research Data Commons, Adelaide, Australia, (4) Australian Research Data Commons, Canberra, Australia, (5) Australian National University, NCI, Canberra, Australia, (6) Columbia University, CIESIN, Palisades NY, USA, (7) University of Helsinki, Helsinki, Finland

To enable reproducibility of research results, it is important for a researcher to be able to cite the exact dataset that was used to underpin their research, especially when the dataset is large, dynamic and evolving over time. One aspect of reproducibility is the need for unambiguous references to a specific version of a dataset. However, such systematic data versioning practices are currently not available. This gap in research data curation was addressed by the RDA Working Group on Data Citation in their final report draft on systematic data versioning practices.

Versioning procedures and best practices are well established for scientific software and can be used to enable reproducibility of scientific results. The code base of large software projects does bear some resemblance to large dynamic datasets. Are versioning practices for code also suitable for datasets or do we need a separate suite of practices for data versioning? How can we apply our knowledge of versioning code to improve data versioning practices?

Over the past two years, the RDA Data Versioning Working Group has collected numerous use cases of data versioning practices and extracted data versioning patterns. A draft of the Group's report and recommendations for data versioning practices will be presented in this session.