EGU2020-12079, updated on 13 Aug 2022
https://doi.org/10.5194/egusphere-egu2020-12079
EGU General Assembly 2020

Enabling Science at Scale

Jeff de La Beaujardiere
National Center for Atmospheric Research, Computational and Information Systems Laboratory, United States of America (jeffdlb@ucar.edu)

The geosciences are facing a Big Data problem, particularly in the areas of data Volume (huge observational datasets and numerical model outputs), Variety (large numbers of disparate datasets from multiple sources with inconsistent standards), and Velocity (need for rapid processing of continuous data streams). These challenges make it difficult to perform scientific research and to make decisions about serious environmental issues facing our planet. We need to enable science at the scale of our large, disparate, and continuous data.

One part of the solution relates to infrastructure, such as by making large datasets available in a shared environment co-located with computational resources so that we can bring the analysis code to the data instead of copying data. The other part relies on improvements in metadata, data models, semantics, and collaboration. Individual datasets must have comprehensive, accurate, and machine-readable metadata to enable assessment of their relevance to a specific problem. Multiple datasets must be mapped into an overarching data model rooted in the geographical and temporal attributes to enable us to seamlessly find and access data for the appropriate location and time. Semantic mapping is necessary to enable data from different disciplines to be brought to bear on the same problem. Progress in all these areas will require collaboration on technical methods, interoperability standards, and analysis software that bridges information communities -- collaboration driven by a willingness to make data usable by those outside of the original scientific discipline.