Geophysical Research Abstracts
Vol. 14, EGU2012-7451, 2012
EGU General Assembly 2012

Massive Meta-Data: A New Data Mining Resource

W. Hugo
South African Environmental Observation Network

Worldwide standardisation, and interoperability initiatives such as GBIF, Open Access and GEOSS (to name but three of many) have led to the emergence of interlinked and overlapping meta-data repositories containing, potentially, tens of millions of entries collectively. This forms the backbone of an emerging global scientific data infrastructure that is both driven by changes in the way we work, and opens up new possibilities in management, research, and collaboration.

Several initiatives are concentrated on building a generalised, shared, easily available, scalable, and indefinitely preserved scientific data infrastructure to aid future scientific work.

This paper deals with the parallel aspect of the meta-data that will be used to support the global scientific data infrastructure. There are obvious practical issues (semantic interoperability and speed of discovery being the most important), but we are here more concerned with some of the less obvious conceptual questions and opportunities:
1. Can we use meta-data to assess, pinpoint, and reduce duplication of meta-data?
2. Can we use it to reduce overlaps of mandates in data portals, research collaborations, and research networks?
3. What possibilities exist for mining the relationships that exist implicitly in very large meta-data collections?
4. Is it possible to define an explicit ‘scientific data infrastructure’ as a complex, multi-relational network database, that can become self-maintaining and self-organising in true Web 2.0 and ‘social networking’ fashion?

The paper provides a blueprint for a new approach to massive meta-data collections, and how this can be processed using established analysis techniques to answer the questions posed. It assesses the practical implications of working with standard meta-data definitions (such as ISO 19115, Dublin Core, and EML) in a meta-data mining context, and makes recommendations in respect of extension to support self-organising, semantically oriented ‘networks of networks’.