AGU 2011
Abstract IN23C-1462

Maximizing data holdings and data documentation with a hierarchical system for sample-based geochemical data

Hsu, L., Lehnert, K.A., Walker, J.D., Chan, C., Ash, J., Johansson, A.K., Rivera, T.A.

Sample-based measurements in geochemistry are highly diverse, due to the large variety of sample types, measured properties, and idiosyncratic analytical procedures. In order to ensure the utility of sample-based data for re-use in research or education they must be associated with a high quality and quantity of descriptive, discipline-specific metadata. Without an adequate level of documentation, it is not possible to reproduce scientific results or have confidence in using the data for new research inquiries. The required detail in data documentation makes it challenging to aggregate large sets of data from different investigators and disciplines.

One solution to this challenge is to build data systems with several tiers of intricacy, where the less detailed tiers are geared toward discovery and interoperability, and the more detailed tiers have higher value for data analysis. The Geoinformatics for Geochemistry (GfG) group, which is part of the Integrated Earth Data Applications facility (http://www.iedadata.org), has taken this approach to provide services for the discovery, access, and analysis of sample-based geochemical data for a diverse user community, ranging from the highly informed geochemist to non-domain scientists and undergraduate students. GfG builds and maintains three tiers in the sample based data systems, from a simple data catalog (Geochemical Resource Library), to a substantially richer data model for the EarthChem Portal (EarthChem XML), and finally to detailed discipline-specific data models for petrologic (PetDB), sedimentary (SedDB), hydrothermal spring (VentDB), and geochronological (GeoChron) samples. The data catalog, the lowest level in the hierarchy, contains the sample data values plus metadata only about the dataset itself (Dublin Core metadata such as dataset titel and author), and therefore can accommodate the widest diversity of data holdings. The second level includes measured data values from the sample, basic information about the analytical method, and metadata about the samples such as geospatial information and sample type. The third and highest level includes detailed data quality documentation and more specific information about the scientific context of the sample. The three tiers are linked to allow users to quickly navigate to their desired level of metadata detail. Link are based on the use of unique identifiers: (a) DOI at the granularity of datasets, and (b) the International Geo Sample Number IGSN at the granularity of samples. Current developments in the GfG sample-based systems include new registry architecture of EarthChemXML to include geochemical data for new sample types such as soils and liquids, and the construction of a hydrothermal vent data system. This flexible, tiered, model provides a solution for offering varying levels of detail in order to aggregate a large quantity of data and serve the largest user group of both disciplinary novices and experts.