EGU21-13940, updated on 01 Dec 2022
https://doi.org/10.5194/egusphere-egu21-13940
EGU General Assembly 2021

Linking data systems into a collaborative pipeline for geochemical data from field to archive

Kerstin Lehnert[1], Daven Quinn[2], Basil Tikoff[2], Douglas Walker[3], Sarah Ramdeen[1], Lucia Profeta[1], Shanan Peters[2], and Jonathan Pauli[2]

[1]Columbia University, Lamont-Doherty Earth Observatory, Geoinformatics, Palisades, United States of America (lehnert@ldeo.columbia.edu)

[2]University of Wisconsin, Madison, WI, United States of America

[3]University of Kansas, Lawrence, KS, United States of America

Management of geochemical data needs to consider the sequence of phases in the lifecycle of these data from field to lab to publication to archive. It also needs to address the large variety of chemical properties measured; the wide range of materials that are analyzed; the different ways, in which these materials may be prepared for analysis; the diversity of analytical techniques and instrumentation used to obtain analytical results; and the many ways used to calibrate and correct raw data, normalize them to standard reference materials, and otherwise treat them to obtain meaningful and comparable results. In order to extract knowledge from the data, they are then integrated and compared with other measurements, formatted for visualization, statistical analysis, or model generation, and finally cleaned and organized for publication and deposition in a data repository. Each phase in the geochemical data lifecycle has its specific workflows and metadata that need to be recorded to fully document the provenance of the data so that others can reproduce the results.

An increasing number of software tools are developed to support the different phases of the geochemical data lifecycle. These include electronic field notebooks, digital lab books, and Jupyter notebooks for data analysis, as well as data submission forms and templates. These tools are mostly disconnected and often require manual transcription or copying and pasting of data and metadata from one tool to the other. In an ideal world, these tools would be connected so that field observations gathered in a digital field notebook, such as sample locations and sampling dates, can be seamlessly send to an IGSN Allocating Agent to obtain a unique sample identifier with a QR code with a single click. The sample metadata would be readily accessible for the lab data management system that allows the researchers to capture information about the sample preparation, and that connects to the instrumentation to capture instrument settings and the raw data. The data would then be seamlessly accessed by data reduction software, visualized, and further compared to data from global databases that can be directly accessed. Ultimately, a few clicks will allow the user to format the data for publication and archiving.

Several data systems that support different stages in the lifecycle of samples and sample-based geochemical data have now come together to explore the development of standardized interfaces and APIs and consistent data and metadata schemas to link their systems into an efficient pipeline for geochemical data from the field to the archive. These systems include StraboSpot (www.strabospot.org; data system for digital collection, storage, and sharing of both field and lab data), SESAR (www.geosamples.org; sample registry and allocating agent for IGSN), EarthChem (www.earthchem.org; publishers and repository for geochemical data), Sparrow (sparrow-data.org; data system to organize analytical data and track project- and sample-level metadata), IsoBank (isobank.org; repository for stable isotope data), and MacroStrat (macrostrat.org; collaborative platform for geological data exploration and integration).