Goldschmidt 2023 Abstract
https://doi.org/10.7185/gold2023.20833

What is your unit? Building integrated data experiences for community use and meta-analysis

LUCY PROFETA[1], KERSTIN A LEHNERT[1], MARTHE KLÖCKING[2], STEPHEN RICHARD[1], PENG JI[1], ANNIKA JOHANSSON[1], HANNAH A SWEETS[1], BÄRBEL SARBAS[3], DOMINIK C. HEZEL[4], ADRIAN STURM[5], STEFAN MÖLLER-MCNETT[6], MATTHIAS WILLBOLD[6], SEAN CAO[1], JUAN DAVID FIGUEROA[1], GERHARD WÖRNER[6] AND MR. LEANDER KALLAS, M.SC.[7]

[1]Lamont-Doherty Earth Observatory, Columbia University
[2]Göttingen University
[3]Max-Planck Institute for Chemistry
[4]Goethe-Universität Frankfurt
[5]Göttingen State and University Library
[6]Georg-August-Universität Göttingen
[7]University of Göttingen

Presenting Author: lprofeta@ldeo.columbia.edu

As the volume and diversity of geochemical and cosmochemical datasets continue to grow, integrating them into a comprehensive framework to benefit data analysis and machine learning becomes an increasingly pressing challenge. The successful integration of such large and diverse datasets requires addressing several critical issues, such as data quality, metadata completeness, and interoperability. This abstract describes a solution for integrating even bigger and more diverse datasets in geochemistry while maintaining and improving the highest data and metadata quality, with specific applications to data available through The Astromaterials Data Systems (AstroMat), IEDA2 and their partners.

Building domain-specific standardized metadata schemas in cosmochemistry and geochemistry is difficult due to the lack of uniformity in data and metadata formats. Different databases and research groups often use different data and metadata formats, making it difficult to exchange and integrate data. This lack of uniformity can result in data incompatibilities, inconsistencies, and errors, which can affect the accuracy and reliability of research results. Current data compilations are done primarily manually by researchers and data curators and can be very laborious processes to ensure high quality metadata is preserved.

To aid data producers, as well as end users, development of standardized data formats and metadata templates is being implemented concurrently between multiple large data systems: EarthChem, GEOROC, MetBase and Astromat. These systems provide access to a vast array of geochemical and cosmochemical data to researchers and the wider community. Shared vocabularies will ensure that data and metadata are consistently formatted, facilitating data exchange and interoperability. It will also provide researchers an easier path for planetary analogue studies.

Converging and expanding repository schemas (such as the one for the Astromaterials Data Archive) to match existing synthesis schemas will provide seamless ingestion from individual files to the integrated synthesis databases, producing large volumes of high quality, analysis-ready data.

These data will be retrievable both through user interfaces, as well as through public APIs to facilitate machine learning studies.

This large scale data integration has the potential to revolutionize the way geochemical data is collected, stored, and shared, enabling more comprehensive and accurate research.