Goldschmidt 2023 Abstract
https://doi.org/10.7185/gold2023.19262

Ensuring consistent data quality and trustworthiness in global databases that synthesize legacy and modern geochemical data

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

[1]Göttingen University

[2]Lamont-Doherty Earth Observatory, Columbia University

[3]University of Göttingen

[4]Geo.X - Research Network for Geosciences in Berlin and Potsdam / GFZ German Research Centre for Geosciences

[5]Georg-August-Universität Göttingen

[6]Göttingen State and University Library

Presenting Author: marthe.kloecking@uni-goettingen.de

Synthesis databases, i.e. compilations of data published over decades by different sources, provide the foundation for many data analytics and machine learning techniques in modern geochemical data science. Yet how do we know that we can trust these databases? How can they ensure consistent data quality across such a broad range of data sources and analytical techniques?

The GEOROC and PetDB databases are leading, open-access sources of geochemical and isotopic datasets of terrestrial igneous and metamorphic rocks and minerals. Established simultaneously 24 years ago, they currently provide access to curated compilations of rock and mineral compositions from thousands of publications, ranging from the late 19th century to today, totalling >40 million single data values. The primary purpose of these geochemistry databases is to support and facilitate new research projects using previously published data. Nonetheless, there are a number of technical challenges behind providing such service to the community. The DIGIS initiative for GEOROC 2.0 and EarthChem for PetDB are now developing modern solutions to data submission, discovery and access to support the diverse demands of digital, data-driven geochemical research. One important, yet often overlooked, tool are the vocabularies needed as part of the rich metadata describing samples and analytical procedures linked to geochemical analyses, including, among others, a comprehensive list of mineral names or a description and categorisation of analytical methods and instruments. Whilst some of this information may not be critical for the specific research question the databases are used for, detailed sample and analysis descriptions, following a consistent schema, facilitate assessment of the quality of component datasets - in turn enabling trust in the synthesis databases.

These vocabularies should be implemented using trusted and established sources wherever possible. Both GEOROC and EarthChem are founding members of the OneGeochemistry initiative and closely collaborate with the Astromaterials Data System to develop a common set of vocabularies, following existing standards (e.g. the International Mineralogical Association's 'List of Minerals'). The aim of this collaboration is to allow easier data integration between the three data systems - enabling larger and more diverse databases for more ambitious data science applications in geochemistry research.