DEVELOPING TOOLS AND FAIR PRINCIPLES FOR THE GEOROC AND METBASE DATABASES

Horst R. Marschall[1], Dominik C. Hezel[2], Gerhard Wörner[3], Matthias Willbold[3]

[1]Goethe-Universität Frankfurt, Institut für Geowissenschaften, Altenhöferalle 1, 60438 Frankfurt

[2]Universität zu Köln, Institut für Geologie & Mineralogie, Zülpicherstr. 49b, 50674 Köln

[3]Universität Göttingen, Geowissenschaftliches Zentrum, Goldschmidtstr. 1, 37077 Göttingen

Abstract

GeoROC and MetBase are the two largest geochemical and cosmochemical databases hosted in Germany. Both databases are currently migrated to a new home, with a now continued development and maintenance. A current DFG-LIS project aims to host and develop GeoROC at the University of Göttingen for re-building the technical infrastructure, data input, maintenance and interoperability. In this proposal, we aim to contribute to this endeavour by (i) adopting the recently built interactive visualisation and analysis tools for MetBase to GeoROC, and (ii) equipping the existing and future data with meta-data according to international FAIR standards, adopting the recently built interactive visualisation and analysis tools for MetBase to GeoROC. We further want to (iii) build community awareness for the need for FAIR-guided curation of geochemical databases. Community workshops will initiate an NFDI4Earth geochemical special interest group to deeper embed the databases within the German geo-/cosmochemical community, and promote FAIR and user-friendly GeoROC and MetBase with improved accessibility, appropriate meta-data, interoperability where data can be better visualised and analysed. GeoROC and MetBase have the potential of becoming a valued data source for geochemists, cosmochemists, petrologists, geologists and mineralogists for research and teaching as well as the ESS community and beyond.

Introduction

The number of published geo- and cosmochemical data have seen a substantial and steady increase over many years. In their entirety, these data provide important new and original insights into our understanding of the formation and evolution of terrestrial planets and early solar system evolution as well as understanding the formation and evolution of the early Earth. Further, the processes of chemical differentiation of Earth and their secular changes through its history, as well as growth and evolution of continents and complementary mantle require the analysis of large data sets from the geochemical as well as the cosmochemical domain. Building, curating, and effectively using large geochemical data sets is challenging and requires up-to-date data access (e.g., advanced filters, pre-defined expert databases), data visualisation (e.g., various types of typical geo-/cosmochemical plots, including various normalisation options, variable notations, units) and real-time analyses (modelling e.g., mixing, AFC-processes, isochrons) to unfold their full potential to research as well as a FAIR approach.

More than 20 years ago, two independent initiatives started collecting geochemical (GeoROC) and cosmochemical (MetBase) data in databases. GeoROC is the largest geochemical database worldwide and was established in 1999 by the Max-Planck Institute for Chemistry (MPIC) in Mainz and is now hosted by the GWDG in Göttingen. GeoROC currently holds analyses of >540,000 samples of predominantly volcanic rocks with presently a total of >20 million individual values of elemental or isotopic analyses. Today, the GeoROC database has been used and acknowledged in more than 15,000 peer-reviewed publications, many of which appeared in

high-ranking journals. GeoROC is, together with the US-based PetDB platform, a key-contributor and integral part of the EarthChem initiative. MetBase was also established about 20 years ago and hosted by a private collector in Bremen. Among others, the database consists of more than 500.000 individual data of, for instance, bulk and component chemical, isotopic and physical properties. Further, the database holds more than 90,000 references from 1492 until today. In 2006, the high value of the database was acknowledged by the Meteoritical Society with its Service Award. Thus, both databases are of high relevance for their respective communities, but now need to be modernised for efficient use and new scientific approaches.

Both databases contain chemical and isotopic data of bulk rocks and minerals from samples that cover the entire Earth's history. The GeoRoc database is excellently maintained at an up-to-date level. However, its metadata and FAIR implementation remain rudimentary and access to the database through the web interface is almost unchanged since its implementation. On the other side, the technical foundation of MetBase has recently been migrated from an outdated program to an SQL-based format, however, its metadata and FAIR implementation are equally rudimentary. MetBase, however, was recently updated with a new and innovative web-interface that provides various instant access, visualisation and analyses tools. It is therefore a logical step to implement the same tools to GeoROC. GeoROC is significantly larger than MetBase, which will make this implementation challenging.

GeoROC and MetBase will be moved to their new homes within the next ca. 2 years. Just recently, a substantial DFG-LIS research grant was awarded to the University Göttingen (Wörner et al.) that will ensure the future development, extension and modernisation of the GeoROC database, which is currently still managed at MPI Mainz. It is therefore the perfect time and opportunity to update the user interface of both databases in one step, equip them with a common metadata framework that is taken from international standards or affiliated databases such as EarthChem, iedadata or Astromaterials, and fully implement FAIR principles. Where necessary, new standards will be defined. We therefore aim for three goals, in the combination of the DFG-LIS grant and this pilot: (i) Facilitate ease-of-use of the databases by rolling out a web-based data handling and data analysis system (MetBase) to the GeoROC web-interface (this pilot) (ii) embed GeoROC and MetBase into a FAIR metadata framework, and (iii) work with international partners such as EarthChem or Astromaterials to harmonise the databases and tools for working with and input data into these. These goals will enable researchers to quickly and intuitively work with all data. Furthermore, students will be provided with an up-to-date user- and analysis interface to be trained in the analyses of large geochemical data sets.

Pilot description

GeoROC will be hosted in Göttingen by a companion DFG-LIS research grant ("DIGIS": Digital Geochemical Data Infrastructure), which will also provide new tools of semantic analysis for automatic data extraction from published literature, topical extension of the GeoROC database and its future curation. The major technical innovations of this proposed pilot will be the development of online-tools for data access, handling, presentation, and analyses for both databases. At the same time, we will be jointly working towards the connectivity of complementary initiatives, as well as linking up other international data infrastructures to GeoROC and MetBase to ensure full FAIR implementation.

Once re-building of the infrastructure (Göttingen) and web-presentation (Frankfurt/ Göttingen) and its link to data analysis tools is completed (Frankfurt), the implementation of FAIR principles will be achieved. In this context, contacts of the applicants to stakeholders of affiliate

international databases have been established over the years and will be used for a deep international integration and broad representation in the EES.

It is well known and demonstrated through numerous published work that geochemical data of rocks and minerals will serve as essential reference for further research in the entire field of Earth System Sciences. We believe that increasing the accessibility and ease-of-use of these two databases as well as ensuring traceability of data contained in them will facilitate additional and/or new research ideas within the EES community and encourage non-experts to implement geochemical data into their studies. For instance, global soil research depends on representative and global coverage of rock substrate analysis. Spectral NIR remote sensing of planetary surfaces requires mineral and rock reference compositions on a planetary scale. Archeological studies linking artefacts to their provenance are based on knowing the chemical and isotopic composition of potential sources. The data analysis tools that will be developed here and will make it easier for non-expert users to efficiently use the data bases even beyond the Earth sciences.

All scientists will benefit from novel approaches in data visualisation, analysis workflow, as well as the new ways to access and work with large sets of geochemical and cosmochemical data. MetBase already also includes an entire cosmochemistry online course for lecturers, students or even the interested layman. This course is highly modular and interactive and can be used by teachers and students independently or through a dedicated curriculum. In particular, the course links directly to MetBase where appropriate. This allows students to work with large compositional data sets.

It is one of the core incentives of this entire endeavour to implement FAIR principles to GeoROC and MetBase. Both databases are already easy to *find*. For example, GeoROC data are already directly accessible within selected online publications as well as through its current web page. In addition, the new online tools will allow for much better *access* to the data. The newly built infrastructure will be made *interoperable* from the beginning, using common standards and frameworks. All data are freely available and downloadable for *reuse*.

The major aspects in the research data life cycle addressed are: preservation, access, analyses, and reuse. Geochemical and cosmochemical data are typical examples of “long-tail” data. On the other hand, the life cycle of such data is undetermined but depends – among other aspects – on the quality of the analyses and the development of future (better) analytical methods.

The teaching and learning components as integral parts of a geochemical and cosmochemical database are entirely new, innovative and easily transferable, and can be an interesting educational element for other participants of the NFDI (cf. metbase.org -> Resources).

Deliverables

For this pilot, we plan the following deliverables:

- Developing interactive online-tools for data access (e.g., advanced filters, pre-defined expert databases), visualisation (e.g., various types of typical geo-/cosmochemical plots, including various normalisation options, variable notations, units) and analyses (modelling e.g., mixing, AFC-processes, isochrons) that enable direct scientific use of the databases.
- Embedding the databases in complementary, international initiatives (e.g., EarthChem, Astromaterials), using FAIR principles.

Roadmap

Adding scientific tools to previously multiple, separate databases and apply FAIR principles to these.

Requested funding

We apply for a one year full time equivalent PostDoc funding for a software developer to port the MetBase tools to GeoROC and create a landing page and single entry point for both databases. The software developer is already at hand (Dr. Premkumar Elangovan), who developed and maintains (so far) the MetBase infrastructure, access, visualisation and analysis tools and MetBase webpage.

Work Plan

Months 0-8:

Porting the MetBase access, visualisation, and analyses tools to GeoROC. This is in particular challenging, as GeoROC is a >1 order of magnitude larger database than MetBase.

Months 9-10:

Seeking collaborations with our colleagues from iedadata/earthchem/astromaterials for possible implementations of these tools and FAIR data with their initiatives.

Months 11-12:

Building the landing page and web-interface for both databases into a then single access point.