An adaptable quality evaluation tool for geochemical data

Matthias Willbold[a,*], Axel D. Renno[b,§], Gerhard Wörner[a,*], Dieter Garbe-Schönberg[c,§]

[a] Fakultät für Geowissenschaften und Geographie, Universität Göttingen, Goldschmidtstr. 1, 37077 Göttingen
[b] Helmholtz-Zentrum Dresden-Rossendorf, Chemnitzer Str. 40, 09599 Freiberg
[c] Institut für Geowissenschaften, Christian-Albrechts-Universität Kiel, Ludewig-Meyn-Straße 10, 24118 Kiel
* Digital Geochemistry Infrastructure (DIGIS) - GEOROC §International Association of Geoanalysts (IAG) - GeoReM

Abstract

This project aims to develop a data quality evaluation tool for the GEOROC[1] geochemical database that can be adapted by users according to their specific application. To achieve this, the tool will utilise complementary primary- and metadata stored both, in GEOROC and the GeoReM[2] geochemical reference material database. Both geochemical databases contain published datasets that were produced using different analytical protocols at different levels of quality. This type of systematic offset between individual datasets represents a major obstacle when re-using quantitative compositional data for specialised applications. The outcome of this pilot will not only mitigate this obvious shortcoming in geochemistry by enabling the user to conveniently evaluate the extent of these offsets. The fully documented open source-based architecture of the software tool will also allow its adaptation to other Earth System Sciences databases that contain quantitative compositional data. This pilot contributes to the data curation track.

1. Introduction

Data quality control is a major challenge in Earth Sciences such as for geochemical data contained in the GEOROC database (**Geo**chemistry of **R**ocks of the **O**ceans and **C**ontinents). Hosted by the Georg-August-Universität Göttingen through the DFG-funded DIGIS[3] project, GEOROC provides free access to published isotopic and geochemical analyses primarily of igneous rocks and minerals. The database has been instrumental in advancing our understanding of the global distribution and composition of rocks and minerals throughout Earth's history. Yet, exploring the full capacity of analyses-based Earth Science data with respect to addressing prominent research questions critically depends on the future ability of users to assess, quantify and correct for the analytical bias of data as contained in these databases. In particular, making geochemical data from different sources comparable may well result in a fundamental re-assessment of commonly held paradigms in Earth Sciences along with seminal new findings. **No data evaluation tool is currently available for geochemical data stored in databases** and the aim of this pilot is to mitigate this situation**.**

Overcoming this obstacle by developing and providing an adaptable data evaluation tool for GEOROC is the main thrust of this pilot proposed here. Along this line, GEOROC contains highly valuable additional information in form of a large and curated dataset of geochemical reference materials, individually linked to the published datasets that could be used to critically evaluate data quality, specifically for tackling the problem of amending inter-laboratory or inter-method bias. However, this resource is not yet ready to be utilised for this purpose for users of the database. This is because there is no consensus within the geochemistry community on standardized selection criteria of reference materials to establish a robust inter-laboratory quality control system. In addition, the application of established protocols for testing the comparability of data obtained with different analytical methods or originating from different laboratories or

[1] https://georoc.eu/georoc/new-start.asp
[2] http://georem.mpch-mainz.gwdg.de
[3] https://www.uni-goettingen.de/de/digital+geochemistry+infrastructure/643369.html

measurement series is *de facto* an exception. We note that similar issues exist in other research fields in Earth Sciences where quantitative compositional data are re-used. In this context, we expect to contribute to a show case for best practice in the evaluation of quality of data contained in databases that **has applications to other research fields and their data curation challenges within the wider NFDI4Earth scheme**.

The tool to be developed will allow the method- and laboratory-dependent selection and re-use of GEOROC data for clearly described geochemical problems with a pre-defined tolerable range of measurement uncertainty of the data used for this purpose in the sense of a "fit for purpose" concept. Achieving this task will require (1) expert knowledge on the data infrastructure of the GEOROC database, (2) a strong background in the conception and implementation of standardised and adaptable quality control measures on published data sets, and (3) an expert knowledge of geochemical applications where data of highest quality are crucial. As such, this proposal will be conducted by team members of both DIGIS and the **I**nternational **A**ssociation of **G**eoanalysts[4] (IAG) who can provide this expertise in a joint effort. In the long term, facilitation of an adaptable data evaluation tool will set GEOROC apart from machine-based data harvesting protocols that cannot provide such a service.

2. Pilot description

The primary evaluation of geochemical data quality for unknown samples is done through the assessment of data for geological (and ideally certified) reference materials of known composition and analysed as unknowns alongside samples of interest. By comparing the measurement results for the reference materials with assumed 'true' values, the accuracy of measurement results for the samples of interest can then be gauged. This approach guarantees that measurement results for samples of interest can be corrected for inter-laboratory and inter-method bias and thus can be made comparable between different studies. As with all geochemical analyses, the analytical protocols used to produce measurement results have an impact on the measurement uncertainties of the reported values for analysed reference materials and samples of interest, which is, in turn, a vital information to assess their comparability. In many cases, only very incomplete analytical metadata exist for geochemical data in GEOROC, which makes it almost impossible to allow a robust evaluation of the measurement uncertainties. Only the expert-based compilation of published analytical data together with analytical metadata for reference materials allows a classification and evaluation of the data quality of literature geochemical data – if analyses of measurement results for samples of interest are linked to the analysis of reference materials.

Accordingly, evaluating the data quality for samples of interest collated in databases from different sources may be a daunting task for the uninitiated user of the GEOROC database. It not only requires cross-referencing data for different quality control reference materials but must also take into account different levels of uncertainty allocated to different types of analytical protocols used as well as dealing with information gaps resulting from incomplete reporting of analytical metadata. As such, the proposed solution to this issue and the main **Tasks** of this pilot are:

1) Create a workflow for evaluating geochemical data contained in GEOROC using expert background knowledge on quality assessment of geochemical data.
2) Provide users of the GEOROC database with an adaptable evaluation tool comprising a pre-defined (i.e. expert) set of parameters to filter data retrieved from the GEOROC database according to different levels of data quality and according to their specific application needs.

[4] https://www.geoanalyst.org

3) Develop an open-source software tool that will allow expert users to create their own purpose-built evaluation parameters. The latter will also facilitate the transfer of this tool and its principles to other geochemical and Earth System Sciences (ESS) databases.
4) Provide full documentation of the methods and workflows developed in this pilot for future use in other ESS fields with similar data quality challenges and analytical bias issues.

Notably, the same primary data for geochemical reference materials contained in GEOROC are also curated, to a large degree with additional analytical metadata, in the **Geo**chemical **Re**ference **M**aterial database GeoReM, formerly a sister databased of GEOROC at MPI Chemie in Mainz. GeoReM is the world's leading resource for chemical and isotopic data for geological and environmental reference materials with the aim of providing full traceability, quality control, and reproducibility of geochemical data. The IAG has taken on the responsibility to further curate and develop the GeoReM data base. Most relevant for the proposed pilot, the IAG and DIGIS have recently teamed up to transition the GeoReM database to the DIGIS-GEOROC infrastructure. This task will be supported through funding for a second round of the DIGIS-LIS proposal (under review). Still, data contained in the GeoReM and GEOROC databases are already linked through a common data structure and publication identifier. This circumstance provides an ideal starting point for implementing an adaptable data evaluation tool for curated geochemical data using the GEOROC – GeoReM database link as a test bed that can be applied to other compositional ESS databases.

While the development of features of such a tool will be part of Task 1 (above), some basic functionalities will include:

1) Ability to check the dataset for accuracy of data input (i.e. spurious errors that occurred when manually curating the data).
2) Implementation of a filter to access data according to methods and reference materials used, and that uses tested vocabularies, definition of equivalents and broader terms where needed, or even AI to allow for multiple category entries, strings, and spelling.
3) Implementation of an outlier test with relation to the respective analytical techniques used ('fit-for-purpose').
4) Pre-selection of data based on the use of matrix-matched reference materials.
5) Implementation of a basic 'expert quality control index' for individual published datasets based on data for reference materials published alongside samples of interest. The latter may be guided by applying principles similar to those of the 'z-score' approach implemented in IAG proficiency tests (Thompson 2022).

3. Relevance to the NFDI4Earth

Often, users of the GEOROC database (but also other geochemical or ESS databases) may not be *a priori* aware of the challenges involved in the identification and correction of inter-laboratory and inter-method bias. Therefore, the promotion of such an adaptable quality evaluation tool for compositional data in synthesis data bases is of general importance and can have applications in other research fields beyond geochemistry. It is also an excellent opportunity to educate all parties involved in re-using ESS data in non-scientific contexts (e.g., university teachers, public authorities, decision makers) about the advantages but also responsibilities when using compiled data products as a source of information. Such a tool necessarily requires curated, integrated and interoperable databases with state-of-the art data infrastructure such as those provided through GEOROC and GeoReM. We are therefore in an excellent position to showcase the advantages of fully curated database services adhering to FAIR principles as opposed to non-curated, non-integrated and non-interoperable stand-alone machine-based data harvesting protocols.

4. Deliverables

The expected outcomes of this pilot will have both knowledge-based as well as technical-based dimensions.

1) Cross-referencing and collectively evaluating meta data for geological reference materials as contained in both databases (GEOROC and GeoReM) will result in a work flow as well as an adaptable data evaluation scheme that we aim to publish in a peer-reviewed journal as a best practice work flow. This will enable expert users to adapt such a tool according to their own needs. It will also result in a software-based data evaluation tool, developed by analytical experts and to be used by both, experts in the field of geochemistry and other ESS researchers using geochemical data in their respective fields.
2) Development of a basic open-source data evaluation tool software will result in a fully documented application (published on the GEOROC and GeoReM websites) that can be continuously adapted and expanded by expert users.
3) Full documentation of the work flow and the open-source architecture of the application tool will also ensure that both can be adapted to other ESS databases.

We intend to use the impetus created by this pilot to increase awareness for the necessity of robust evaluation of geochemical data retrieved from databases as part of a promotional campaign using established links to the database community through DIGIS entitled **„Best-practice procedures for evaluation of geochemical data in ESS databases: Increasing the value of curated data-based geochemical research“**.

5. Work plan and requested funding

Months 1 to 3: Develop workflow (PI[5]). Prepare example user cases (PI, IT[6]). Define and design basic tool functionalities (IT)
Months 4 to 6: Develop basic adaptable open-source software tool, write documentation (IT). Feedback from users and stakeholders (PI).
Months 7 to 8: Refine example user cases, refine tool functionalities and workflow (PI). Adapt and continue development of basic adaptable open-source software tool, continue writing documentation (IT).
Months 9 to 12: Implementation of tool on GEOROC/GeoReM platforms, publish documentation (IT), prepare manuscript for publication in peer-reviewed science journal (e.g. *Geostandards and Geoanalytical Research*) (PI and IT)

This pilot will require support by a 12-month, full-time software developer position (TVL-E13, Stage 5) and a research student internship (Wissenschaftliche Hilfskraft ohne Abschluss), 80 hours per month over 12 months. The latter position will prepare (test) example datasets as well as run tests on beta-versions of the software tool.

References

Thompson M (2022) Assigned Values in the GeoPT Proficiency Testing Scheme. *Geostandards and Geoanalytical Research* 46(1):37-41 doi:10.1111/ggr.12408

[5] Principle investigators
[6] Software developer position