**Time to change the data culture in geochemistry**

Authors: Katy J. Chamberlain1† ORCiD: 0000-0002-2010-6182, Kerstin A. Lehnert2, Iona M. McIntosh3, Dan J. Morgan4, Gerhard Wörner5

1: Environmental Sustainability Research Centre, University of Derby, Derby, UK

2: Lamont-Doherty Earth Observatory, Columbia University, Palisades (New York), USA

3: Solid Earth Geochemistry Research Group, Japan Agency for Marine-Earth Science and Technology, Yokosuka, Japan

4: School of Earth & Environment, University of Leeds, Leeds, UK

5: Geowissenschaftliches Zentrum, Georg-August-Universität Göttingen, Göttingen, Germany

email†: [k.chamberlain@derby.ac.uk](mailto:k.chamberlain@derby.ac.uk)

*Geochemical data are vital for understanding Earth’s past, present and future. However, currently only a fraction of geochemical data are findable, accessible, interoperable and reusable, limiting their use in the broadest range of scientific studies. There is an urgent need for international co-ordination of geochemical data and methods to unlock their full research potential.*

Geochemical data are essential for understanding planetary and environmental processes, including the origin and evolution of Earth, the Solar System, and beyond. Increasingly complex analytical equipment has dramatically expanded the production and variety of geochemical data. However, only a fraction are FAIR (Findable, Accessible, Interoperable, Reusable1), owing to a lack of international community-wide standards and best practices. The lack of a shared data culture seriously hinders data quality assessments, incorporation of geochemical results into multidisciplinary studies, and the (re)use of geochemical data in modeling and big data studies. Output geochemical data needs to be corrected to defined standards, and detection limits are variable between laboratories and run times, introducing restrictions in comparisons between laboratories. Potential future discoveries are, therefore, hindered, necessitating an urgent community need to improve access to and reusability of geochemical data and unlock its full potential.

**Data value and longevity**

Every geochemical data point is valuable, each representing an investment of time, money and resources. Therefore, careful data description and curation are required to optimize return on investments, ensure long-term utility, and support transparent and reproducible scientific research. Data preservation is especially critical where samples are ephemeral and/or impossible to obtain again in the future, such as gas or water samples that can’t be preserved.

Geochemical data are often fragmented and diverse in their character, for example with respect to material analyzed (rocks, minerals, soils, water), bulk or in-situ measurements, concentrations and (isotope) ratios, compositional maps and more. These geochemical data are more than the final published values. They require comprehensive metadata (Table 12) to maximize their post-publication usefulness. At a minimum, metadata should include information about sample provenance and preparation, the laboratory, analysis date, instrument model, configuration and calibration, data quality (for example, accuracy and reproducibility of measurements including reference material data, blank values, and such), and data reduction procedures (for example normalization, fractionation correction). In doing so, geochemical data would be reproducible, can be compared between labs, and recalculated as appropriate as analytical techniques develop and standards are refined.

The value of geochemical data is further augmented if data and metadata are consistently structured (for example using XML or JSON schemes), so they are machine-readable and fully interoperable for computational and statistical approaches. Interoperability allows data in curated databases such as [GEOROC](http://georoc.mpch-mainz.gwdg.de/georoc/Start.asp)A and [PetDB](https://search.earthchem.org/)B and others to be integrated into large global datasets that that cover the entire Earth history. These can then be analyzed with respect to new research questions. GEOROC and PetDB have used the EarthChemXML to automatically encode and integrate their data and metadata, but most databases today need to be manually mapped to the EarthChemXML schemes.

**Table 1: Examples of geochemical metadata**

Caption: Suggested minimum required metadata that should be submitted alongside all datasets, irrespective of whether samples are rock, water or gas, adapted from2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Geographic info** | **Samples** | | | | |
| ***All samples*** | ***All samples*** | ***Rocks*** | ***Water & Gas*** | ***Fossils*** | ***Minerals*** |
| Sample number | Sampling time | Chemical characterization | Sampling container | Fossil group | Mineral name |
| Sample lat-long | Time zone | Sediment grain size | Sampling treatment | Fossil species | Occurrence type |
| Lat-Long accuracy | Collector / scientist | | Water / gas type | Fossil preservation | |
| Elevation | Archiver / curator | | Water / gas Classification | |  |
| Elevation accuracy | Sampling method | |  |  |  |
| **Laboratory** | **Analysis** | | | | |
| ***Interoperability*** | ***Reproducibility*** | ***Normalization*** | ***Fractionation*** | ***Detection Limit*** | ***Blank*** |
| Analyst | Reference sample name | Reference sample name | Reference sample name | Detection limit & unit | Blank type |
| Laboratory | Uncertainty value & unit | Normalization value & unit | Fractionation parameter | | Blank value & unit |
| Analytical code |  |  | Fractionation value & unit | |  |
| Instrument used |  |  |  |  |  |
| Measured parameter |  |  |  |  |  |
| Technique description |  |  |  |  |  |

**The problems**

Data quality assessment is undermined by the non-standardized approach to geochemical data publication and methods. For example, sharing reference material and rejected analyses is not common practice across the community. However, it is impossible to evaluate and test reproducibility, uncertainty estimation, normalization procedures and similarities between different research groups without publishing the complete data and metadata. Ultimately, this practice compromises the use of geochemical data. In particular, it impedes integration with other multidisciplinary datasets, as non-geochemists are more likely to be unfamiliar with uncertainties and different methods, affecting data comparison and interpretation.

Databases such as GEOROC, PetDB and [NAVDAT](https://www.navdat.org/)C have improved access to rock geochemical data and interoperability among themselves. However, maintaining common metadata standards, updating vocabularies, and providing harmonized data evaluation tools remain a challenge. Many other data curation systems and databases exist at national geological surveys and other institutions (for example, GeoKniga, USGS National Geochemical Database, National Geochemical Survey of Australia) but lack interoperability. Even large international repositories such as the [EarthChem Library](https://earthchem.org/ecl/)D have to follow their own ‘standards’ because a globally endorsed and governed standard of data does not yet exist. Ultimately, data publishing and metadata requirements are variably enforced between different funders and journals, and are not yet embedded into community culture3. It commonly falls to individual researchers to archive their new data in repositories.

**FAIR solutions**

Fortunately, there is growing recognition that a FAIR, robust approach to handling geochemical data is sorely needed. Since 20144 minimum metadata requirements for publishing geochemical data have been defined (Table 1). Funders and [publishers](https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards) have implemented policies that require researchers and authors to deposit data in public repositories, making data open**,** citable and available5. Further progress should be made to ensure that such policies and requirements are widely enforced by the various publishers and funding bodies, and that data is deposited in FAIR repositories such as the EarthChem Library.

New projects and initiatives are emerging at national and global levels to advance, integrate and share geochemical data standards and promote best practices. For example, Australia’s AuScope Geochemistry NetworkE, the [European Plate Observing System (EPOS](https://www.epos-eu.org/data-services/community-services-tcs/multi-scale-laboratories))F, and laboratories participating in NASA’s asteroid sample-return mission([Osiris-Rex](https://www.asteroidmission.org/)G), are all promoting the sustainable and universal use and re-use of geochemical data. In addition, the Digital Geochemistry Infrastructure project([DIGIS initiative](https://www.uni-goettingen.de/de/digital+geochemistry+infrastructure/643369.html)H) aims to modernize the widely used GEOROC database, (which hosts 1.8 million published analyses of igneous rocks) incorporating FAIR principles and improving metadata structure and user accessibility.  Streamlined connections with other geochemical databases, such as EarthChem and [Astromat](https://www.astromat.org/)I databases, will improve data interoperability.

In response to Open Access policies and science demands, ever-increasing numbers of geochemical databases are emerging at national, programmatic, and subdomain levels. As such, there is an urgent need for the geochemistry community to coordinate these efforts and collaboratively define the required data and metadata standards and best practices that will enable worldwide interoperability and re-use of geochemical data and methods. The international [OneGeochemistry](https://www.earthchem.org/communities/onegeochemistry/)J initiative is responding to this need, as it aims to create a global geochemical data network that facilitates and promotes discovery and access of geochemical data by unifying diverse data strategies developed across the globe6,7.

**Changing data culture in geochemistry**

Changing data culture requires broad adoption of new rules, workflows, and values. Open and comprehensive data sharing needs to be mandated and enforced by funders and publishers. All data should be shared via repositories that follow the FAIR principles, and every dataset should be assigned a globally unique identifier (such as a DOI), where every physical sample should have an IGSN to link the original sample material unambiguously to data and publications. Repositories should ensure compliance with domain-relevant community standards that describe the analytical procedure, sample provenance, and data quality. Adaptation to a FAIR data culture will require the potentially uncomfortable recognition that fully illuminating data and methods could identify mistakes or oversights. Researchers must establish the attitude that this identification is a positive outcome that will ultimately advance scientific progress.

Researchers will need guidance and resources to adopt and adapt to new practices. Innovative digital tools are needed to support automated data and metadata capture in the laboratory (for example using electronic laboratory notebooks [ELN]) and to facilitate data cleaning during database incorporation. Manufacturers of analytical instrumentation need to provide open interfaces to data and metadata generated by their software to be linked to ELNs. Institutions and funders need to allocate funds to support the effort of proper data management.

A fundamental requirement for geochemical data culture change is that researchers get credit for sharing their data. Universities and research institutes must recognize the value of open data curation and help grow awareness and appreciation of the FAIR data principles. Teaching of data handling and open science principles should be embedded alongside the earliest introduction to geochemical data during undergraduate study. Meanwhile, data and open science sessions at geoscience conferences must be promoted on centre stage instead of being scheduled in competition with discipline-specific sessions.

Lastly, a sustained global geochemical data infrastructure, like the OneGeochemistry initiative, is urgently needed to facilitate the use and maximize the impact of data to ensure scientific progress. International coordination is essential to ensure geochemical database interoperability and consistent analytical quality and metadata standards. Only then will the full potential of multidisciplinary progress be unlocked to address fundamental research questions regarding Earth’s uncertain future.

**Acknowledgements**

KJC & DJM acknowledge support from NERC (NE/S007520/1). KJC and DJM thank the participants of the Volcano-Tectonic Links workshop that seeded the idea for this comment. KL acknowledges support by NSF (EAR-1948806 and NASA (80NSSC19K1102). IM acknowledges support by JSPS KAKENHI Grant Number JP20K14586. GW and the DIGIS project are supported by the German Science Foundation (Wo362-53-1).

**Competing Interests**

The authors declare they have no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in institutional affiliations.

**Related Links**

A. GEOROC database:<http://georoc.mpch-mainz.gwdg.de/georoc/>

B. PetDB database:<https://search.earthchem.org>

C. North American Volcanic and Intrusive Rock Database NAVDAT: <https://www.navdat.org/>

D. EarthChem Library: <https://earthchem.org/ecl/>

E. AuScope Geochemistry Network AGN:<https://www.auscope.org.au/agn>

F. Multiscale Laboratories of the European Plate Observing System EPOS:<https://www.epos-eu.org/data-services/community-services-tcs/multi-scale-laboratories>

G. Osiris-Rex [KC5] :<https://www.asteroidmission.org>

H. DIGIS initiative:<https://www.uni-goettingen.de/de/643369.html>

I. Astromat database: <https://www.astromat.org/>

J. OneGeochemistry: https://www.earthchem.org/communities/onegeochemistry/

**References**

1. Wilkinson, M. D., et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018 (2016)

2. Staudigel, H. et al. Electronic data publication in geochemistry. *Geochemistry, Geophysics, Geosystems*, 4(3) (2003)

3. Goldstein, S. L., Lehnert, K., and Hofmann, A. W. Managing Open and FAIR Data in Geochemistry: Where are we a decade after the Editors Roundtable?, *EGU General Assembly 2021*, online, 19–30 Apr 2021, EGU21-16420, https://doi.org/10.5194/egusphere-egu21-16420 (2021).

4. Goldstein, S. L., Hofmann, A. W., Lehnert, K. A. Requirements for the Publication of Geochemical Data, Version 1.0. Interdisciplinary Earth Data Alliance (IEDA). [https://doi.org/10.1594/IEDA/100426 (2014).](https://doi.org/10.1594/IEDA/100426%20(2014).)

5. Stall, S. et al. Make scientific data FAIR. *Nature*, **570**, 27 (2019).

6. Lehnert, K., & Wyborn, L. A. OneGeochemistry: Toward a Global Network of Geochemical Data. In *AGU Fall Meeting 2019*, AGU (2019).

7. Wyborn, L., Lehnert, K.: OneGeochemistry: Creating a Global FAIR-Way to Access and Share Geochemical Data”. *Goldschmidt Conference 2020,* <https://doi.org/10.46427/gold2020.2910> (2020)