THE NEW PARADIGM OF DATA PUBLICATION

Kerstin Lehnert and Leslie Hsu[1]

Open and persistent access to scientific data has become a popular topic. Accessing past, present, and future scientific data is fundamental to making scientific research transparent and reproducible, and it ensures that the products of past and current research can be re-used to empower future science and to benefit society. Governments, funders, academic institutions, professional societies, and publishers alike have issued new data policies, statements, and directives that endorse or demand open access to data. As a consequence, scientific workflows, research communication, and scholarly values are changing to recognize and encourage data sharing. And new ways for making data open and persistently accessible have emerged: Data Publication.

Scholarly publication has always been the preferred way of making data available, especially in disciplines such as mineralogy, petrology, and geochemistry where data volumes are small enough that they can be included in articles as data tables and/or electronic supplements. Unfortunately, data published in this way becomes highly dispersed across the literature; finding, accessing, and mining this data is difficult to impossible. The review process primarily focuses on the scientific relevance of the presented results and not on aspects of data re-usability; compliance with data standards is not enforced and critical metadata are often missing. And then there are all those data that did not give rise to a publication. What happens to them? They usually stay hidden on local hard drives and are eventually lost.

Data should be submitted to a domain repository where they will be properly curated and preserved, and where their value will grow as they become discoverable, citable, re-usable, and integrated with similar data into comprehensive, large-scale data collections (Fig. 1). Such data collections, or syntheses, are the foundation for the type of data-driven, abductive discovery that Hazen (2014) envisions for mineralogy. In igneous geochemistry and petrology there are the databases of PetDB, GEOROC, and NAVDAT, which have already made this type of data-driven science a reality. These databases integrate thousands of statistically significant and dense sampling measurements into large-scale searchable syntheses that have, for more than a decade, driven important global scientific discoveries (Lehnert and Langmuir 2007) and changed the way "geochemists do geochemistry" (Hofmann 2008). Examples of the impact of these databases include studies on the diversity in mid-ocean ridge basalt (MORB) composition (Gale et al. 2013), on the global distribution of elements in Earth's outermost layers (Rauch 2011), and on global patterns of intraplate volcanism (Conrad et al. 2011). Developing and maintaining these databases isn't easy. Significant effort is required to compile data from individual articles and there are struggles with incomplete, inconsistent, and ambiguous metadata. That sort of effort is neither scalable nor sustainable, but it could be if all the relevant original data were saved directly to a domain repository.

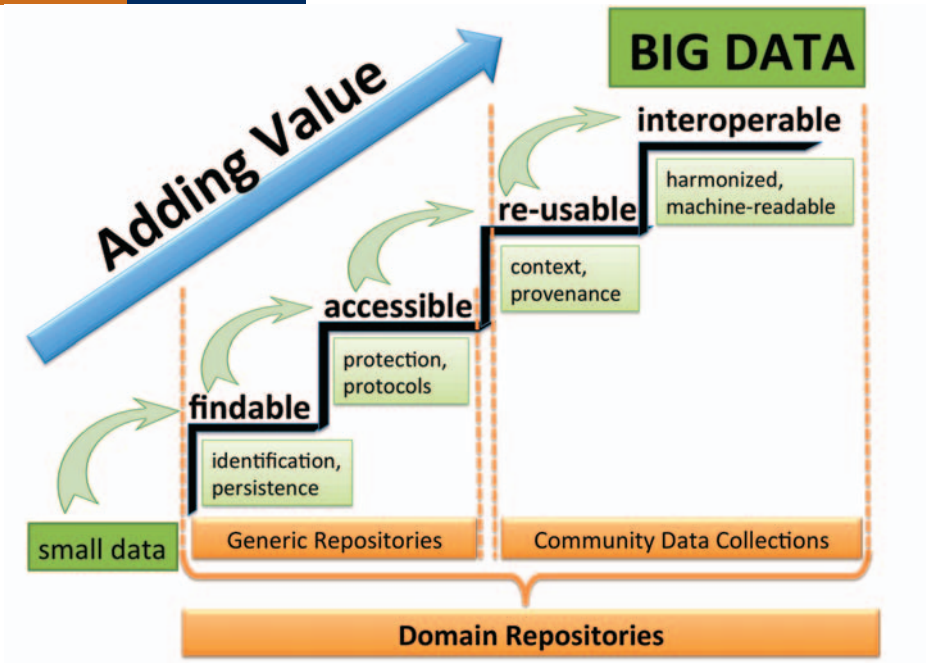


Figure 1 A summary of how data repositories augment the value of small data to eventually grow into BIG data resources for advanced data-driven research. The figure illustrates the difference between generic and domain-specific repositories and what community data collections can offer. Community data collections are smaller, thematically focused, well-curated, and offer data systems that are highly valuable resources for specific science communities, but they lack the sustainability, infrastructure, and data curation practices that constitute a domain repository

Publishers now acknowledge that domain data repositories are best poised to ensure access and maximize impact of data. In October 2014, publishers and Earth science data facilities formed a new partnership: the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS, http://www.copdess.org; Hanson et al. 2015) (Fig. 2). They signed a joint Statement of Commitment that, among other recommendations, advocates that "Earth and space science data should, to the greatest extent possible, be stored in appropriate domain repositories that are widely recognized and used by the community, follow leading practices, and can provide additional data services."

Figure 2 Publishers, leaders of Earth science data facilities, and funders met at the AGU Headquarters in Washington, DC, in October 2014 and established the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS, http://www.copdess.org).

1 Lamont-Doherty Earth Observatory
Columbia University, Palisades, NY, USA

This recommendation applies to traditional journals as well as to the new type of data journals, e.g. *Geoscience Data Journal* (Wiley), *Scientific Data* (Nature Publishing Group), *Earth System Science Data* (Copernicus), and *Earth & Space Science* (AGU/Wiley). Data journals offer a way to publish (and get credit for!) data without the requirement to present novel discoveries or groundbreaking insights. Articles in data journals describe the data and methods of collecting and processing them. The actual data are submitted to a digital repository. As an example, see the article "RU_CAGeochem, a database and sample repository for Central American volcanic rocks at Rutgers University," which was published in the *Geoscience Data Journal* by Carr et al. (2014; doi: 10.1002/gdj3.10) with the data deposited in the EarthChem Library (Carr et al. 2015; doi: 10.1594/IEDA/100534). This type of data publication offers a great opportunity for late-career scientists, in particular, to share and preserve any unpublished data that are in danger of being lost when they retire. The IEDA Data Rescue initiative (Hsu et al. 2015) helped several investigators to compile and publish valuable geochemical data of Apollo samples that had not been in digital form or accessible in any publication (Delano 2014).

So, what constitutes an "appropriate" domain repository, and how do you find the right one for your data? A repository needs to have the expertise, operational infrastructure, and sustainability to comply with leading practices for data stewardship; an editorial process to assess the quality of submitted data and metadata; use of persistent and globally unique identifiers such as the Digital Object Identifier (DOI) and international Geo Sample Number (IGSN) so data can be properly cited and linked; provisions for the data's secure and long-term preservation; protection of deposited data until released for public access; and clear policies for the use and citation of data holdings. Domain repositories must maintain standards for relevant context and provenance information, which is usually very specific for a given data type (e.g. fractionation correction for isotope ratio measurements) (Deines et al. 2003; Goldstein et al. 2014). Appropriate repositories enrich and organize data to facilitate new discoveries. Generic data repositories, such as FigShare, Dryad, or institutional repositories, simply lack the expertise to do this. Databases such as the Library of Experimental Phase Relations (LEPR; http://lepr.ofm-research.org/; Hirschmann et al. 2008), MetPetDB (database for metamorphic petrology data, http://metpetdb.rpi.edu; Spear et al. 2009), and RRUFF (database of Raman spectra, X-ray diffraction, and chemical data for minerals; http://rruff.info) offer important domain-specific features, but currently lack the ability to guarantee long-term preservation of data or citability of contributed data. Trusted domain repositories use widely accessible open formats and work with institutions to guarantee access for decades into the future, migrating data to new formats and media as they evolve.

Recognized domain repositories for the mineralogy/petrology/geochemistry community are sparse. EarthChem (www.earthchem.org) provides services for long-term accessibility, persistent identification, and quality assurance of geochemical and petrological data following community-vetted guidelines and using the data curation infrastructure of the Interdisciplinary Earth Data Alliance (IEDA, www.iedadata.org), an accredited member of the World Data System. EarthChem provides data templates to help investigators organize and format different types of data and the appropriate metadata. EarthChem data managers assist users and perform quality assessment. Users have control over the date that their data become available for public download.

Interdisciplinary Earth Data Alliance's data curation services are currently serving EarthChem and the Marine Geoscience Data System, and it is expanding. Interdisciplinary Earth Data Alliance is now also partnering with LEPR and MetPetDB to allow these latter data collections to use its services and so make these latter data collections function as proper repositories for their specific communities. By becoming partners in the Interdisciplinary Earth Data Alliance, LEPR and MetPetDB will be able to provide long-term preservation of data and DOI registration of submitted datasets while maintaining control over their systems and user communications. This scalable approach will allow far more researchers within the Earth sciences community access to the data and should facilitate greater publication options.

By promoting and applying best practices of data publication and citation, investigators can help sustain the cyberinfrastructure resources that enable data to be accessed by the wider research community and thereby create a positive feedback to more accessible data.

We encourage you to take part by publishing your data in the EarthChem Library. If you cannot find a template for your specific data type, we invite you to contact EarthChem at info@earthchem.org and help us develop it. The EarthChem Library files are linked to published manuscripts, contributing to the causes of reproducible science and reusable data.

REFERENCES

Carr MJ and 5 coauthors (2014) RU_CAGeochem, a database and sample repository for Central American volcanic rocks at Rutgers University. Geoscience Data Journal 1: 43-48, doi: 10.1002/gdj3.10

Carr MJ and 5 coauthors (2015) RU_CAGeochem v.4, a database and sample repository for Central American volcanic rocks at Rutgers University. EarthChem Library, doi: 10.1594/IEDA/100534

Conrad CP, Bianco TA, Smith EI, Wessel P (2011) Patterns of intraplate volcanism controlled by asthenospheric shear. Nature Geoscience 4: 317-321, doi: 10.1038/ngeo1111

Deines P, Goldstein SL, Oelkers EH, Rudnick RL, Walter LM (2003) Standards for publication of isotope ratio and chemical data in Chemical Geology (editorial). Chemical Geology 202: 1-4

Goldstein SL, Hofamnn AW, Lehnert KA (2014) Requirements for the publication of geochemical data. Integrated Earth Data Applications (IEDA). doi: 10.1594/IEDA/100426

Hanson B, Lehnert KA, Cutcher-Gershenfeld J (2015) Committing to publishing data in the Earth and space sciences. Eos 96, doi: 10.1029/2015EO022207

Hazen RM (2014) Data-driven abductive discovery in mineralogy. American Mineralogist 99: 2165-2170, doi: 10.2138/am-2014-4895

Hirschmann MM and 8 coauthors(2008) Library of Experimental Phase Relations (LEPR): A database and Web portal for experimental magmatic phase equilibria data. Geochemistry, Geophysics, Geosystems 9: Q03011, doi: 10.1029/2007GC001894

Hofmann AW (2008) Mantle myths, mantle reservoirs, and databases. Geochimica et Cosmochimica Acta, (Goldschmidt 2008, abstract volume) 72 (12S): pA384

Hsu L and 8 coauthors (2015) Rescue of long-tail data from the ocean bottom to the Moon: IEDA Data Rescue Mini-Awards. GeoResJ 6: 108-114, doi: 10.1016/j.grj.2015.02.012

Lehnert K, Langmuir CH (2007) The PetDB Data Collection: Impact on Science. Geological Society of America Abstracts with Programs 39: 153

National Science Board (2005) Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. Report NSB-05-40

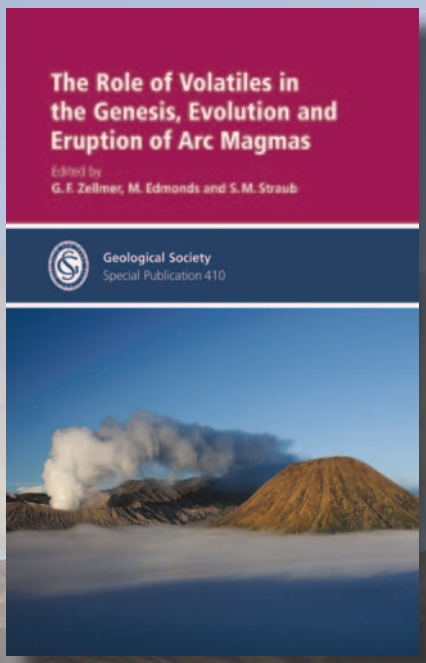
Rauch JN (2011) Global distributions of Fe, Al, Cu, and Zn contained in Earth's derma layers. Journal of Geochemical Exploration 110: 193-201, doi: 10.1016/j.gexplo.2011.05.008

Spear FS and 11 coauthors (2009) MetPetDB: a database for metamorphic geochemistry. Geochemistry Geophysics Geosystems 10: Q12005, doi: 10.1029/2009GC002766

The Geological Society

serving science & profession

PUBLICATIONS



The Role of Volatiles in the Genesis, Evolution and Eruption of Arc Magmas

Edited by G.F. Zellmer, M. Edmonds and S.M. Straub

The volatile cycle at subduction zones exerts strong controls on the petrogenesis, transport, storage, evolution and eruption of arc magmas. This volume highlights recent progress in understanding aspects of this cycle, including its bearing on eruption triggering and volatile release into the atmosphere, through case studies and comprehensive reviews.

ONLINE BOOKSHOP
ORDER CODE
SP410

Published 17 March 2015
GSL Special Publications
Hardback | 292 Pages
ISBN: 978-1-86239-689-0

List price: £100.00
Fellow's price: £50.00
Other societies price: £60.00

The Use of Palaeomagnetism and Rock Magnetism to Understand Volcanic Processes

Edited by M.H. Ort, M. Porreca and J.W. Geissman

Published 14 April 2015
GSL Special Publications
Hardback | 281 pages
ISBN: 978-1-86239-629-6

ONLINE BOOKSHOP
ORDER CODE
SP396

List price: £100.00
Fellow's price: £50.00
Other societies price: £60.00

Ore Deposits in an Evolving Earth

Edited by G.R.T. Jenkin, P.A.J. Lusty, I. McDonald, M.P. Smith, A.J. Boyce and J.J. Wilkinson

Published 02 January 2015
GSL Special Publications
Hardback | 333 pages
ISBN: 978-1-86239- 626-5

ONLINE BOOKSHOP
ORDER CODE
SP393

List price: £110.00
Fellow's price: £55.00
Other societies price: £66.00

ORDER ONLINE: WWW.GEOLSOC.ORG.UK/BOOKSHOP

Lyell Collection
Geological Society publications online

The Geological Society's Lyell Collection: journals, Special Publications and books online. **www.lyellcollection.org**