Acta Geologica Sinica (English Edition), 2019, 93(supp.1): 32–33

FAIR Geoscientific Samples and Data Need International Collaboration

Kerstin LEHNERT[1, *], Lesley WYBORN[2] and Jens KLUMP[3]

[1] *Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, 10964, USA*
[2] *National Computational Infrastructure Facility and Research School of Earth Sciences, Australian National University, Canberra, 2600, ACT, Australia*
[3] *CSIRO ARRC, Kensington, WA 6151, Australia*

Citation: Lehnert et al., 2019. FAIR Geoscientific Samples and Data Need International Collaboration. *Acta Geologica Sinica* (English Edition), 93(supp. 1): 32-33

1 Introduction

The primary goal of the Deep-time Digital Earth project is to develop an open collaboration and data sharing platform that enables the transition of deep-time geoscientific research to a Big Data driven paradigm. Such an open platform will require the ability to effectively and efficiently access and integrate a wide variety of digital Earth data, from remotely sensed Earth observations and geophysical data to data generated in laboratories on individual physical specimens, all of which need to be temporally and spatially referenced to past paleogeology and paleogeography, rather than to today's geology and geography.

In the last decade or so, there has been considerable investment in developing data infrastructure to capture and preserve data from satellite and airborne campaigns, or major sets of data from sensor networks in specifically funded data centers. These are data that come in large volumes (as measured in terabytes or even petabytes), but as they are relatively homogenous and standardized they are comparatively easy to make findable and accessible. In contrast, data generated by laboratory analysis of samples are heterogeneous, and distributed across many disparate institutions, both nationally and internationally. These data belong to the 'Long Tail' of data as defined by Heidorn (2008), who noted that 'relatively little attention is given to the data that is being generated by the majority of scientists'. He also noted that 'The long tail is a breeding ground for new ideas and never before attempted science', emphasizing the scientific power that lies hidden in these long-tail data. But the infrastructure to support discovery, access, interoperability, and reusability of long tail data, i.e. their alignment with the FAIR Data Principles (Wilkinson et al. 2016), is relatively immature. International collaboration is essential to reach agreement on, and generate broad adoption of, best practices and standards for such data.

2 FAIR Principles for Open Science

Over the past few years, there has been growing recognition that the concept of 'Open' alone, be it applied to data, software, samples, or other research products, is not sufficient to dramatically impact the advancement of science and make new paradigms such as data-driven science/data intensive science a reality. Leading principles and practices have now been defined that aim to ensure that data and other products of research such as software and samples are not only openly accessible, but that they can be found easily on the internet (Findable), can be accessed persistently (Accessible), can be linked and integrated into a distributed research data infrastructure (Interoperable), and can be understood, interpreted, and used in a meaningful way with clear usage licenses (Reusable), not only for people, but also for machines. If these four requirements are met, data are FAIR. The same principles can be applied to samples that should be discoverable online, accessible as both digital and physical objects, linked to other entities such as persons, publications, datasets, funding awards, etc. Under these principles, as a minimum, data or samples must have unique and persistent identifiers and metadata appropriate to assist discover. They should be cited in a form equivalent to other scholarly outputs. They should be accessible through a standard, web-based protocol. They should be annotated with sufficient provenance information so that they can be reused with confidence and clarity. Both data and samples must be well curated, persistently accessible, and linked securely to associated publications and other resources.

Various initiatives, projects, and working groups are working to advance the implementation of the FAIR principles, among them the AGU project "Enabling FAIR Data", the European GOFAIR initiative, and Working Groups within the Research Data Alliance and the World Data System. These efforts are making it increasingly clear that many aspects of FAIRness, specifically reusability, are highly context and domain specific. Indeed, the original FAIR guiding principles explicitly point to "domain-relevant community standards", but there are many open questions regarding such domain-relevant community standards: Who develops and maintains them? Who determines which are the authoritative standards for any domain? Who has the authority to approve them and then govern them once they are developed? How granular do domain-specific standard definitions need to be?

3 IGSN and the Rise of FAIR Samples

Samples form the basis of a large portion of geoscientific

* Corresponding author. E-mail: fengzq.lehnert@ldeo.columbia.edu

http://www.geojournals.cn/dzxbcn/ch/index.aspx
https://onlinelibrary.wiley.com/journal/17556724

research – they are the raw material of much of geoscience, a basic element for reference, study, and experimentation. Collections of samples represent highly valuable, often irreplaceable records of nature (IWGSC, 2009) and carry enormous potential for future discoveries. Deep-time science relies heavily on the study of physical samples such as sediment cores, stratigraphic sections, and paleobiological specimens that hold many keys to knowledge about the Earth's history and evolution. Their chemical composition, mineral and fossil components, metamorphic grade, and other properties are used as proxies to decipher processes that shaped our Earth since its existence and the status of Earth's environments in the past. Samples need to be findable and reusable for a global research community to build on previous research. In the modern digital research ecosystem, transparency and reproducibility of research also demands access to samples (McNutt et al. 2016; Nosek et al., 2015).

Finding samples, particularly those that have been cited in scholarly publications, has been difficult because samples until recently have only been catalogued locally in institutional or personal databases and spreadsheets, and existing online catalogues have not been systematically federated. Federation of distributed online catalogs is necessary to make discovery and access of physical samples possible, but it can only be achieved when each sample has 1) a globally unique and persistent identifier, so it can be located and cited without ambiguity in publications and database, and 2) a virtual representation, i.e. a metadata record about the individual collection holdings that can be accessed online by both human and machines (Lehnert and Klump, 2012). The use of globally unique, persistent, and resolvable identifiers for samples ensures unambiguous and actionable links from publications to online metadata profiles (landing pages) and to other data generated by other studies of the same sample. It allows previously impossible linking and integration of sample-based observations across data systems and paves the road toward advanced data mining of sample-based data, because the full potential of sample-based data can only be realized when the vast numbers of individual observations are combined like pieces of a puzzle that can be explored to reveal large scale patterns in space, time, and property dimensions.

The IGSN is a persistent unique identifier for Geoscience samples that was developed at the Lamont-Doherty Earth Observatory starting in 2004 and which has evolved into an international PID system that is now adopted by a growing number and range of stakeholders worldwide, including researchers, collection curators, and data managers. More than 6.5 million samples have been registered so far. The role of the IGSN is underlined by its inclusion into the recommendations issued by the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) (Hanson et al. 2015) and successive implementation in author guidelines at publishers such as AGU and Copernicus to ensure proper linking of data, samples, and publications. International collaboration has been a fundamental aspect of the IGSN's success. The IGSN is now operated as a globally distributed system of Allocating Agents who provide local or national communities with IGSN registration services, while adhering to agreed-upon standards for sample metadata, metadata exchange protocols, and IGSN syntax. Allocating Agents are members of the IGSN e.V., an international non-profit organization, and thus directly participate in the governance of the IGSN.

4 Developing International Standards for Geochemical Data

For nearly 15 years, EarthChem (Lehnert et al., 2007) has operated a global geochemical data network that makes geochemical data FAIR (Findable, Accessible, Interoperable, Reusable). The power of FAIR geochemical data has been demonstrated by large-scale, global geochemical data syntheses like EarthChem that have, for nearly two decades, inspired and made possible a vast range of studies and new discoveries within the narrow domains they cover, facilitating the analysis and mining of geochemical data, and creating new paradigms in geochemical data analysis. Lack of consistent best practices and protocols for documenting, formatting, and encoding geochemical data that would allow more automated aggregation of data and networking of different data systems has hindered the growth of such data resources across all areas of geochemistry both within the US and internationally. As more data systems emerge at national, programmatic, and subdomain levels in response to Open Access policies and science needs, globally endorsed conventions for documenting geochemical data and standard protocols for exchanging these data among systems will need to be developed, implemented, and governed. Deep-time Digital Earth can help bring together relevant communities to explore and define possible solutions, and can be a powerful test bed for their adoption.

Key words: FAIR, findable, accessible, interoperable, reusable, data standards, samples, IGSN

Acknowledgments: We would like to thank the US National Science Foundation for their long-time support of the development of the IGSN (Grant Nos. NSF-0445178, NSF-0514551, NSF-0552123) and the Earth Chem system (Grant No. NSF-0522195) and operation of both systems as part of the IEDA Data Facility (Grant Nos. NSF-0950477, NSF-1636653). We thank the Alfred P. Sloan Foundation for a grant to Columbia University to support the development of a global, scalable, and sustainable technical and organizational infrastructure for persistent unique identifiers of physical scientific samples.

References

Heidorn, P.B., 2008. *Library Trends*, 57(2): 280–299.
Hanson, B., et al., 2015. *Eos*, 96.
Interagency Working Group on Scientific Collections (IWGSC), 2009. A green report of the IWGSC. https://usfsc.nal.usda.gov/sites/usfsc.nal.usda.gov/files/IWGSC_GreenReport_FINAL_2009. Pdf
Lehnert, K.A., et al., 2012. The Geoscience Internet of Things. *Geophysical Research Abstracts*,14, EGU2012-13370, 2012 EGU General Assembly 2012.
Lehnert, K., et al., 2007. *Geochimica Et Cosmochimica Acta*, 71 (15): A559–A559.
McNutt, M., et al., 2016. *Science*, 351(6277): 1024–1026.
Nosek, B.A, et al., 2015. *Science*, 348: 1422–1425.
Wilkinson, M.D., et al., 2016. *Scientific Data*, 3.

About the first author (also corresponding author)

Kerstin LEHNERT, female; PhD; graduated from the University of Freiburg in Germany; Doherty Senior Research Scientist at the Lamont-Doherty Earth Observatory of Columbia University and Director of the NSF-funded data facility IEDA (Interdisciplinary Earth Data Alliance). She is now interested in the development of community-driven data infrastructures to improve access and sharing of Earth and space science data and physical samples. Email: lehnert@ldeo.columbia.edu.