eResearch AUSTRALASIA2019

Physical Sample Identifiers at Scale: Bringing Samples into the Linked Research Ecosystem

Jens Klump[1]

Jens Klump[1], Kerstin Lehnert[2], Doug Fils[3], Sarah Ramdeen[4], Lesley Wyborn[5]

[1]CSIRO Mineral Resources, Kensington WA, Australia, jens.klump@csiro.au
[2]Lamont-Doherty Earth Observatory of Columbia University, Palisades NY, USA, lehnert@ldeo.columbia.edu
[3]Consortium for Ocean Leadership, Slater IA, USA, dfils@oceanleadership.org
[4]Ronin Institute, Huntsville AL, USA, sarah.ramdeen@gmail.com
[5]NCI/ANU, Canberra ACT, Australia, lesley.wyborn@anu.edu.au

INTRODUCTION

Samples have always been at the heart of scientific research [1]: samples were either taken from nature or produced in laboratory experiments. and Over the past two centuries, we have collected hundreds of millions of samples, and we are still collecting more. While infrastructures for scientific literature and data have evolved into a networked and searchable research information ecosystem, online access to sample information has lagged way behind and often we cannot even unambiguously identify which samples were the basis of which dataset and publication.

THE CURRENT STATE OF PERSISTENT IDENTIFICATION OF SAMPLES: THERE IS A GROWING NEED

Samples are often named ad hoc in the course of a research project with names that are most likely not globally unique. A query of the EarthChem database gives more than one hundred samples named “M1” of diverse rock types from almost any location. The concept of IGSN was developed to uniquely identify samples to enable large synoptic geochemical studies, but even though IGSN originated from applications in the geosciences [2], it is now finding broader interest in other disciplines as diverse as life sciences, materials research, agriculture, or archaeology [3]. In addition, there is an increasing trend to apply identifiers at finer and finer granularities. For example, a drill hole can be the parent identifier, a core sample a child, each mineral separate is another derivative and then an analysis spot on a mineral another derivative. Identifying samples at this finer granularity creates a need for a very robust and sustainable identifier system.

Persistent unique identifiers (PID) are a critical element in digital research data infrastructure to unambiguously identify, locate, and cite digital representations of a growing range of entities - publications, data, instruments, organisations, funding awards, field programs, and others. Globally unique and web resolvable persistent identifiers allow us to link all these elements together.

Most will be familiar with Digital Object Identifiers (DOI) and IGSN was developed in analogy to DOI as a globally unique and web resolvable persistent identifier for physical samples [4]. However, a publication may refer to only a handful of datasets: in contrast, there can be hundreds if not thousands of samples referred to in a paper and increasingly publishers want samples to be identified [5].

In Summary, uptake in more communities, application at a finer granularity and more demanding guidelines from publishers now require that Sample identifier systems have to scale, and it is predicted that there will be a need for billions of identifiers

SCALING THE SYSTEM

To be able to scale the IGSN infrastructure to billions of samples, interconnected with a comparable number of datasets and their related publications, requires a redesign of both the organisational model and technical architecture of current persistent identifier infrastructures. Growing the scale of persistent identifier systems also needs coordination across the key identifier systems such as ORCID, DataCite, Crossref, etc.

Scaling persistent identifier system to hundreds of millions of objects, or even billions of objects is a challenge for the governance of such a large system and its technical implementation. In 2018, the IGSN Implementation Organization (IGSN e.V.) received a grant from the A.P. Sloan Foundation to review the current organisational and technical implementation of IGSN with the aim of developing it into a sustainable and scalable structure that will be able to serve very large numbers of sample identifiers to a diverse research community working with physical samples.

Preliminary work has shown that the scale of identifying physical samples and linking them to all other elements that constitute the record of science, will severely challenge the scalability of existing persistent identifier systems and current web architectures. The move away from XML-based schemas and services will mean a paradigm shift for persistent identifier systems and greater use of linked data in the record of science.

REFERENCES

1. McNutt, M., Lehnert, K. A., Hanson, B., Nosek, B. A., Ellison, A. M., & King, J. L. (2016). Liberating field science samples and data. Science, 351(6277), 1024–1026. https://doi.org/10.1126/science.aad7048

2. Lehnert, K. A., Goldstein, S. L., Lenhardt, W. C., & Vinayagamoorthy, S. (2004). SESAR: Addressing the need for unique sample identification in the Solid Earth Sciences. In AGU Fall Meeting 2004 (SF32A-06). San Francisco, CA: American Geophysical Union. Retrieved from http://adsabs.harvard.edu/abs/2004AGUFMSF32A..06L

3. Hobern, D., Hahn, A., & Robertson, T. (2018). Options to Apply the IGSN Model to Biodiversity Data. Biodiversity Information Science and Standards, 2, e27087. https://doi.org/10.3897/biss.2.27087

4. Lehnert, K. A., Klump, J., Arko, R. A., Bristol, S., Buczkowsky, B., Chan, S.-L., et al. (2011). IGSN e.V.: Registration and Identification Services for Physical Samples in the Digital Universe (IN13B-1324). Presented at the American Geophysical Union Fall Meeting, San Francisco, CA: American Geophysical Union. Retrieved from http://abstractsearch.agu.org/meetings/2011/FM/sections/IN/sessions/IN13B/abstracts/IN13B-1324.html

5. Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K. A., Nosek, B., Parsons, M., Robinson, E., and Wyborn, L. A. I. (2019). Make scientific data FAIR. Nature, 570, 27-29. https://10.1038/d41586-019-01720-7