EGU22-13317
https://doi.org/10.5194/egusphere-egu22-13317
EGU General Assembly 2022

The critical role of unique identification of samples for the geoanalytical data pipeline

Kerstin Lehnert[1], Jens Klump[2], Sarah Ramdeen[1], Kirsten Elger[3], and Lesley Wyborn[4]
[1]Lamont Doherty Earth Observatory, Columbia University, New York, USA
[2]CSIRO, Mineral Resources, Perth, Australia
[3]GFZ German Research Centre for Geosciences, Potsdam, Germany
[4]Australian National University, Canberra, Australia

When researchers collect or create physical samples they usually assign a user-generated number to each sample. Subsequently, that sample can be submitted to a laboratory for analysis of a variety of analytes. However, as geoanalytical laboratories are generating ever increasing volumes of data, most laboratories have automated workflows and it is no longer feasible for laboratories to use researcher-supplied sample numbers, particularly as it is not guaranteed that user-supplied numbers will be unique in comparison to numbers submitted by other users to the same laboratory. To address this issue new, laboratory-generated numbers may be assigned to that sample.

Moreover, as a single laboratory rarely has the capability to offer all analytical techniques, individual samples tend to move from laboratory to laboratory to acquire the desired suite of analytes. Each laboratory may implement a different number to that sample. At the conclusion of their project, the researcher may submit the same sample to a museum or institutional repository, where the sample will be assigned yet another institution-generated number to ensure that all samples are uniquely identified in their repository.

Ultimately, by the time the researcher submits an article to a journal and wants to identify samples in the text or tables, they may have a multitude of locally-generated numbers to choose from. Not one of the locally assigned numbers to that sample can be guaranteed to be globally unique. It is also unlikely that any of these local numbers will be persistent over the longer term (decades), or be resolvable to enable the location of the identified resource or any information about it elsewhere on the web (metadata, landing page, services related to it, etc).

Globally unique, persistent, resolvable identifiers such as the IGSN play a critical role in the unique identification of geoanalytical samples that pass between systems and organisations: they cannot be duplicated by another researcher, laboratory or sample repository. They persistently link to information about the origin of the sample; to personas in the creation of the sample (collector, institution, funder); to the laboratory data and their creation (analyst, laboratory, institution, funder, data software); and to the sample curation phase (curator, repository, funder). They connect the phases of a sample's path from collection in the field to lab analysis to the

synthesis/research phase to the publication to the archive. Globally unique sample identifiers also enable cross linkages to any artefacts derived from that sample (images, analytical data, other articles). Further, identifiers like IGSN enable sub samples or sample splits to be linked back to their parent sample, creating a holistic picture of any information derived from the initial sample.

Hence, best practice is clearly to assign the globally unique resolvable identifier to the initial resource. Like a birth certificate, the identifier can be carried through the progressive stages of the research 'life-cycle' including laboratory analysis, generation of further data, images, publication, and ultimately curation and preservation. Where any subsamples are derived, they, and any data generated on them, can be linked back to the parent identifier.