The role of unique identifiers in tracing the life cycle of a sample and any data derived from it

SARAH RAMDEEN[1], LESLEY ANNE I WYBORN[2], KERSTIN A LEHNERT[3] AND JENS KLUMP[4]

[1]Columbia University LDEO
[2]Australian National University
[3]Lamont-Doherty Earth Observatory, Columbia University
[4]Commonwealth Scientific and Industrial Research Organisation
Presenting Author: sramdeen@ldeo.columbia.edu

Data in many research domains, particularly in the natural sciences, are either derived from samples collected from the natural environment or generated in the lab. Properly curated collections of samples are a highly valuable resource to future research: may represent irreplaceable records of nature. Samples are key to open, reproducible, and transparent science in the modern research data ecosystem. lf a sample is FAIR it is possible to trace the entire lifecycle of a sample, from birth - collection in the field, to analysis in the lab, transfer between scientists and repositories, to publication, and to archive and reuse. To be FAIR, a sample should be uniquely identified, well described with sample appropriate metadata, findable in online catalogs, and connected to other relevant resources, including related observational and analytical data, images, publications, people, funding awards, and other digital information.

Ambiguity in sample names has generated significant confusion and has made it difficult to relate multiple analyses to a specific sample. The IGSN (International Generic Sample Number, rebranded in 2022 to reflect the expanding user community), provides a central registration system for assigning persistent, globally unique identifiers for samples and sampling features. Each IGSN is resolvable to a metadata profile for that sample. If a sample is subsampled, a new IGSN is issued for the child sample, and the relationship between the parent and child is maintained within the metadata profiles of both samples. The function of an IGSN goes beyond unambiguous identification to support connections between samples and related data, publications, people, organizations, instruments, and grants. The IGSN's data model is based on ISO 19156:2011 (Observations and Measurements), which allows it greater interoperability with other systems..

In the geochemistry community, data systems that support different stages in the sample and sample based geochemical data lifecycle are exploring standardized APIs and consistent data and metadata schemas to link their systems [1]. IGSN is a key part of this pipeline to support geochemical samples and any data generated on them from their path from creation to the archive.

[1] Lehnert, et al. (2021), EGU General Assembly, EGU21-13940