IGSN ID Implementation Exemplars: GFZ Data Services

Rorie Edmunds, **Kirsten Elger**

Blog post published July 10, 2024 in *DataCite Blog - DataCite*
https://doi.org/10.5438/cmcs-a589

Abstract

Through the 2021 partnership between DataCite and IGSN e.V. , DataCite services can be used to register International Generic Sample Numbers (IGSN IDs) for material samples. Over the coming months, the blog series 'IGSN ID Implementation Exemplars' will showcase sample management workflows developed by the community that incorporate IGSN ID registration.

Keywords

Information Systems (Subfield), IGSN, Members

Recommended citation

Edmunds, R., & Elger, K. (2024, July 10). IGSN ID Implementation Exemplars: GFZ Data Services. *DataCite Blog - DataCite*. https://doi.org/10.5438/cmcs-a589

Copyright

Copyright © Rorie Edmunds et al. 2024. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Through the 2021 partnership between DataCite and IGSN e.V., DataCite services can be used to register International Generic Sample Numbers (IGSN IDs) for material samples. Over the coming months, the blog series 'IGSN ID Implementation Exemplars' will showcase sample management workflows developed by the community that incorporate IGSN ID registration. In each post, we offer practical guidance on how to work alongside disciplinary sample experts to register IGSN IDs within DataCite services.

GFZ Data Services is hosted at the GFZ German Research Centre for Geosciences. It is a research data repository for DOI-referenced data and scientific software from the Geosciences, and provides IGSN ID registration for geosamples. GFZ Data Services was a founding member of IGSN e.V. in 2011, and until 31 December 2022, when the IGSN PID infrastructure transitioned under DataCite, GFZ Data Services was responsible for managing the central IGSN handle server that registered more than 10.5 Million IGSN IDs. As IGSN Allocating Agent, GFZ Data Services has assigned close to 39,000 material samples with IGSN IDs for its community. Almost 3,000 IGSN IDs have been directly registered within DataCite services since the transition.

GFZ Data Services webpage showing a research data repository with a search bar and various scientific images arranged in hexagonal shapes. The page includes data categories such as agriculture, archaeobotany, and rocks/minerals. There is also a section explaining "

The IGSN ID services of GFZ (right) can be reached via the GFZ Data Services website (left). It provides an overview of IGSN IDs, the FAIR WISH Project (see below), as well as metadata tools and guidance.

Samples Communities and Domain-specific Metadata

For IGSN ID registration, GFZ Data Services works with many internal and external groups across all of the Earth Sciences, including a large number of long-tail communities. GFZ therefore does not have a central sample database for the whole institution. Instead, some groups are using digital sample management systems that are bespoke to their needs and typically not used beyond their group, while others are collecting sample descriptions in spreadsheets saved on their own computers. While GFZ Data Services is operating a central IGSN ID catalogue and metadata database, it has to be highly flexible in how it obtains the required sample metadata and has therefore developed several strategies to achieve this.

For example, the International Continental Scientific Drilling Program (ICDP) records sample metadata in the mobile Drilling Information System (mDIS), and thus already has a database. To register IGSN ID metadata, GFZ Data Services and the ICDP Operational Support Group at GFZ have developed semi-automated metadata export routines that paste information from mDIS directly into IGSN Extensible Markup Language (XML) from where it is then mapped to the DataCite Schema. So far, GFZ Data Services has registered almost 16,000 samples for different scientific drilling projects.

Another example is EarthShape, a critical zone observatory for which over 3,000 samples have been registered with IGSN IDs. In this case, GFZ Data Services has added additional functionality to a former electronic lab notebook such that it is now a digital sample management system with

integrated IGSN ID registration. To keep the integrity of a sample tree (see also below), it is ensured that if a sample from the n-th hierarchy is registered, every parent is registered until the very first sample.

FAIR WISH Project

The project FAIR WISH—FAIR Workflows to Establish IGSN IDs for Samples in the Helmholtz Association was funded by the Initiative and Networking Fund of the Helmholtz Association within the HMC Project Cohorte 2020 of the Helmholtz Metadata Collaboration Platform HMC. The aim of FAIR WISH was to foster the wider application of IGSN IDs in the Geoscience domain and to enable the use of IGSN IDs by researchers who want to assign them to their samples but don't have access to a database. The objective of the project has been to develop:

1. Workflows to generate machine-readable IGSN ID metadata and automatically register IGSN IDs for use cases representing different states of digitization: from very basic and analogue sample description tables written on paper to a very modern biogeochemical sample database that is filled using a field app.
2. Discipline-specific metadata profiles for different sample types, including the identification of controlled vocabularies for use at the Linked Open Data level.

The main product of the project is a FAIR Samples Template, a customizable Excel template to provide standardized sample descriptions, and is designed to be used by researchers who have no clue about XML metadata. The researcher first selects the metadata properties that they would like included in the description, then this generates a personal metadata table into which they can fill their sample information. GFZ Data Services has developed the SAMIRA software to read and map the metadata descriptions from the templates into the IGSN ID and DataCite metadata. Documentation and a video tutorial explain how to use the template.

> Screenshot of an Excel spreadsheet detailing a sample database structure with variable names, definitions, and mandatory status, alongside an XML code snippet and an XML file icon. The arrow points from the spreadsheet to the XML, indicating data transformation.
> *A sample metadata description generated in the FAIR Samples Template and converted to XML.*

This template, which is now in its second version, is suitable for individual and hierarchical samples. GFZ Data Services has already identified some improvements for the next version, and is currently waiting to see if a new project proposal will be approved that enables it to further automate the template by incorporating all the controlled vocabularies in nominal lists and not requiring them to be copied and pasted.

Landing Pages

GFZ Data Services currently uses PHP to generate landing pages for material samples registered with an IGSN ID. Following on from above, IGSN ID landing pages contain online sample descriptions according to a metadata schema/profile that likely depends on the sample type. A

sample is connected with its landing page through the IGSN ID, which is encoded in a QR code that resolves to the landing page. It is important to say that the IGSN ID is not replacing any local sample number or name, but it is complementary. So a sample tag contains both human-readable versions of the sample number and the IGSN ID, and machine-readable QR code that leads to the IGSN ID landing page.

As an example of this, the below figure shows one of GFZ Data Services' landing pages for a sample from ICDP and that includes citation information for both related paper and datasets. Whenever it exists, a 'parent sample' property is mandatory for IGSN ID sample metadata, as it is key to making connections between different samples and subsamples. This connection is used to visualize a sample family tree that can be browsed to identify not only a sample's parents, but also its siblings. On the example page, not all of the 4,000 samples from this drilling project are shown, but users can further expand the sample tree to reach each related sample. Since they are web-resolvable identifiers, IGSN IDs can act as anchors for the relationships between a material sample and its associated data and literature. To enable this connection, IGSN IDs must be both cited in data publications and included in the machine-readable metadata sent to DataCite DOI registration services. For IGSN IDs, this has been possible since the release of DataCite Metadata Schema 4.0 in 2016. Examples of sample citations can be found in the data catalogue of GFZ Data Services (e.g., Lorenz et al., 2019).

> Data sheet for a rock sample collected from Äre, Jämtlands län, Sweden, listed by the GFZ German Research Centre for Geosciences. Includes general identifiers, sampling location, geology, methods, drilling details, repositories, and related
> *A landing page for a sample from ICDP. The page includes the sample's hierarchical relationships, its location, and citation information of associated papers and datasets.*

When possible, the location of the sample is shown on a map, providing information on the actual sample location, as well as information about the core repository or laboratory where potential users might contact a person to ask whether they can also take a subsample for further investigation. The IGSN ID landing pages also provide links to related data publications and papers.

Sitting on top of all of everything, GFZ Data Services has a dedicated sample catalogue that has different search filters according to sample-type material classification and that links to the sample landing pages.

SAMIRA – Sample IGSN Registration Automation

To connect everything, GFZ Data Services has developed software called SAMIRA (Sample IGSN Registration Automation) that reads sample descriptions from the FAIR Samples Template or from databases, writes XML following the IGSN Descriptive or DataCite Metadata Schema, and communicates with DataCite for IGSN ID registration and metadata exchange via a REST Application Programming Interface.

The SAMIRA software is meant to be highly generic and flexible, and bridges the IGSN ID legacy system with a new one that is being developed, including the IGSN ID registration via DataCite

services. The outputs are in XML currently, but it is planned that outputs will be database entries in the next version, and eventually in JSON. The ability to generate IGSN ID landing pages is also close to being ready, and the catalogue will be connected through OAI-PMH.

Moving Forward

In the future, GFZ Data Services wants to further automatize the sample registration workflows and develop sample-type specific metadata profiles. It wants to have a full and user-friendly implementation of the controlled vocabularies in the FAIR Samples template that can be used by people and in the metadata schema. GFZ Data Services will be ready to connect to additional sample management systems and also look further into the development of sample-type specific services beyond the Earth Sciences.