The IGSN–DataCite Partnership
Metadata Enrichment

Cody Ross & Rorie Edmunds
09 November 2022
DataCite Open Hours

@datacite
International Generic Sample Number

- Globally unique and persistent identifier for material samples
- Domain-agnostic; samples can be any material from anywhere in the universe

A new name, a new role for the IGSN Organization and Identifier

The IGSN e.V. Executive Committee announces the new name of the IGSN e.V.: *International Generic Sample Number Organization e.V.* The new name was approved by the IGSN e.V. General Assembly. The IGSN ID will now be called the International Generic Sample Number to reflect its expanding scope and utilization. Through the partnership between the IGSN e.V. and DataCite, IGSN IDs can now be used in any discipline to identify physical objects.

The mission of the International Generic Sample Number Organization e.V. is to provide guidance on best practices and standards for FAIR samples across disciplines. FAIR samples follow the principles of findable, accessible, interoperable and reusable research resources.
In October 2021, DataCite and IGSN e.V. agreed to form a partnership

- DataCite provides the IGSN ID registration services and supports ongoing sustainability of the IGSN ID infrastructure
- IGSN e.V. serves to implement and promote standard methods for identifying, citing, and locating material samples

**Phase 1 – Technical Transition**

- Engagement, technical planning, and initial transitioning subphases are completed
- IGSN ID registration under DataCite services was launched in September
- Re-registration & aliasing remain

**Phase 2 – Scaling Up**

- Support ongoing needs
- Strategy for discoverability of IGSN IDs
- Scale adoption and use of IGSN IDs
- Develop sample Communities of Practice in different domains
What is the Metadata Context?

**Metadata for Material Samples**
- For existing IGSN IDs: Map between IGSN metadata—registration and descriptive—and DataCite Metadata Schema
- For new IGSN IDs: What metadata is collected and how to populate DataCite Metadata Properties
- IGSN descriptive schema not universally adopted by IGSN Allocating Agents
- Work with domain sample communities towards community consensus for minimal descriptive metadata and extensions

**Cross-linking & Citation**
- Describe relations between material samples, and with datasets or publications

**Landing Pages**
- Displays a digital description of the sample identified by the IGSN ID
- Should include additional elements that improve discoverability
- Parts of a metadata record may be withheld to protect sensitive information
What Types of Objects?

All material sample types in all disciplines

PLUS

Collections/aggregates
- May not want to have a PID for every object, but rather the collection – consider granularity

Features-of-interest
- Collection sites
- A sample is a child of its feature-of-interest (i.e., a subsample) – mirror hierarchical relationships among IGSN IDs

Destroyed/discharded samples
- The material sample itself does not have to be persistent
- Metadata should provide information on the current status of the sample
How Is Metadata Captured?

Captured during any part of the samples workflow
● Often important to start capture in the field
  ○ Pre-assign and batch issue IGSN IDs in ‘Draft’ state—may have limited metadata or internet connection
  ○ Use field-based tools
  ○ Change to ‘registered’ or ‘findable’ once mandatory metadata properties are available
● At other stages of the process, more information is likely available

Integration into analytical systems
● Enables access and update to sample descriptions

By whom?
● Typically captured by the Principal Investigator of a project/campaign and their team who collect the samples, as well as analysts and curators/repository managers
How Is Metadata Enriched?

DataCite Metadata Schema

● IGSN IDs are registered with metadata encoded in the DataCite Metadata Schema following the ‘IGSN–DataCite Crosswalk Recommendation’ for material samples metadata
● To improve discoverability of IGSN IDs, as many properties as possible should be populated, and then enriched over time, using local sample metadata collected according to each institution
● Use APIs and Fabrica to update IGSN ID metadata as additional information about samples is gathered or contributed to local samples databases
● The Schema cannot capture the huge variety of information for different sample types from different fields
  ○ Changes to the Schema will be informed by the samples community to better support their needs

By whom?

● Again metadata enrichment is likely by the Principal Investigator and their team, and analysts and curators/repository managers
● If in a university environment, the library may also have a leading role in metadata enrichment
Metadata Capture in Sample Workflow

● Planning (what samples? from where?....)
  ○ Publisher: the organization registering the IGSN ID
  ○ Funding references: financial support for the collection or analysis of a sample
● Collection or creation (‘birth’)
  ○ Creators: sample collector/creators, chief scientists, and Principle Investigators
  ○ Alternate identifiers: including local sample identifiers
  ○ GeoLocation: where the sample was acquired
  ○ Publication year: the year the sample was made available to the research community
  ○ Date: when the sample was collected
● Transportation to facility
● Processing/preparation (clean-up, inspection, selection)
  ○ Subject: materials and other classifications describing the sample
Metadata Capture in Sample Workflow

- **Analysis (data generation)**
  - Related identifiers: part relationships to parent sample and subsample IGSN IDs
  - Date: if the sample is destroyed during analysis
  - Description: other information about the sample, including abstracts and methods

- **Repository ingest (long-term storage and protection)**
  - Creators: curator
  - Contributors: including hosting institutions, data managers, and researchers

- **Publication (data, results,...)**
  - Related identifiers: citation relationships to datasets, journal articles, and other published materials with PIDs
Mandatory Properties

1 **Identifier**
   - Automatically populated with a DOI upon the creation of an IGSN ID

2 **Creator**
   - Sample collector/creator, chief scientist, curator, or person who deposited the sample into a repository
   - If no appropriate name is available, populate with the IGSN ID Repository registrant or an appropriate standard value for unknown information

3 **Title**
   - Appropriate elements that help find and distinguish a sample. Exact syntax at the discretion of the IGSN ID registrant
     - Basic form of the object
     - Material(s) that compose to the sample
     - Local sample identifiers

4 **Publisher**
   - The organization registering the IGSN ID for the physical sample

5 **PublicationYear**
   - The year the sample was first made available to the research community; likely when the physical sample is registered

10 **ResourceType**
   - Use of ‘material sample’ or ‘feature-of-interest’ are recommended to distinguish between these concepts

10.a **resourceTypeGeneral**
   - ‘PhysicalObject’
6 Subject
● The materials that compose the sample. May be categorized under different schemata

7 Contributor
● All institutions and people contributing to the collection and management of the material sample
● Definitions in the contributorType controlled list may need to be revised

8 Date
● Log events relevant to the physical object
● dateType=Collected should contain the collection time of the material sample
● For destroyed samples, use dateType ‘Other’ with ‘Destroyed’ in the dateInformation field

11 alternateIdentifier
● Other identifiers for the material sample; e.g., local sample identifiers

12 relatedIdentifier
● New IGSN IDs are relatedIdentifierType ‘DOI’, thus are displayed in DataCite APIs and Commons and represented in the PID Graph
● Directly maps to relatedIdentifier and relatedIdentifierType in both IGSN registration and descriptive metadata

17 Description
● Use descriptionType=Abstract to add useful information about the sample at its ‘birth’
● Use descriptionType=Methods to outline the sample collection process

18 GeoLocation
● Location where the material sample was collected, noting that samples can be ‘non-geographic’
● Geographic concepts in the IGSN descriptive metadata map to the DataCite Metadata Schema
How Are Connections Relevant?

Child–Parent Relationships
• Vital to unambiguously link parents and derived children, including features-of-interest, and to capture the hierarchical nature of the links

Discovery and credit
• Also important to unambiguously describe relations between a material sample and associated datasets about the sample, publications in which the sample is referenced, and related researchers and institutions

Custom metadata
• Use the ‘HasMetadata’ relationType attribute to link to external sample metadata in a schema specific to an institution, discipline, or samples community