The IGSN Global Sample Number - A PID for Physical Objects

Jens Klump, Kerstin Lehnert, Sarah Ramdeen, Lesley Wyborn | 12 October 2021

CSIRO
Australia's National Science Agency

LAMONT-DOHERTY EARTH OBSERVATORY
THE EARTH INSTITUTE AT COLUMBIA UNIVERSITY

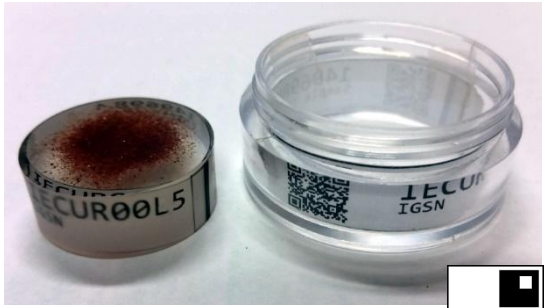
NCI

IGSN

Samples are at the Heart of Research

What are IGSN Global Sample Numbers?

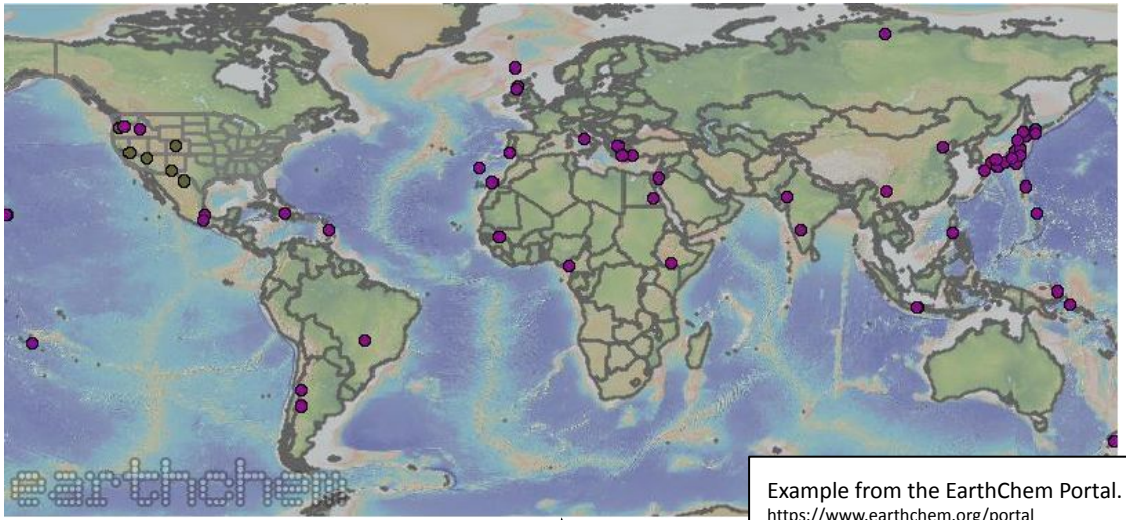
- IGSN are persistent, globally unique, web resolvable identifiers for physical samples & specimens.
- IGSN is domain-agnostic. Samples can be any material from anywhere.



Globally Unique

PetDB Identifier:	ARGAMPH-003
IGSN:	N/A
Other Names:	AMPH D-3(SUN, 1980) D3(ENGEL, 1964) PD3(TATSUMOTO, 1965) PD3(TATSUMOTO, 1966) AMPH-D3(MACDOUGALL, 1986) AMPH D-3(SCHILLING, 1975) S-10(SUBBARAO, 1972) PV D-3(ENGEL, 1965) AMPH-3D(PINEAU, 1983) AMPH 3-PD3(HART, 1971) PD-3(HEDGE, 1970) PD-3(MUEHLENBACH, 1972) AMPH3D(PINEAU, 1976) D-3(SCHILLING, 1971) D-3(SCHEIDEGGER, 1981)

Example from PetDB.
https://search.earthchem.org



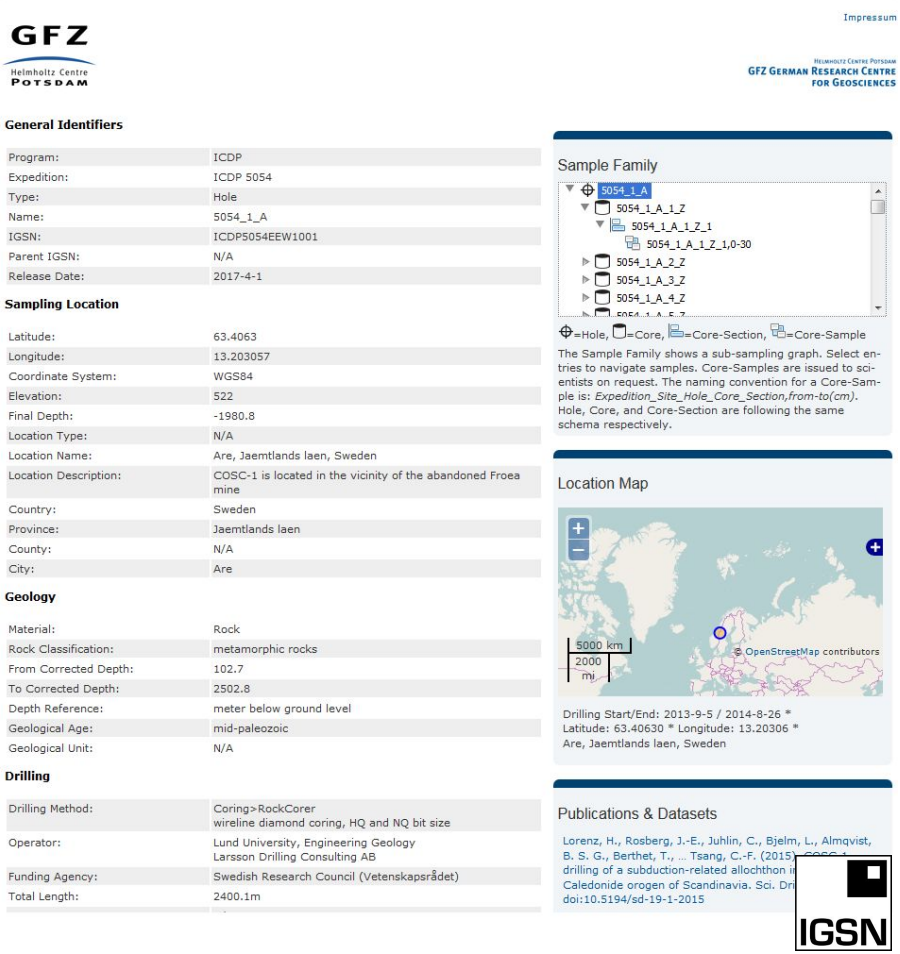
Example from the EarthChem Portal.
https://www.earthchem.org/portal

One sample, many names

One name, many samples

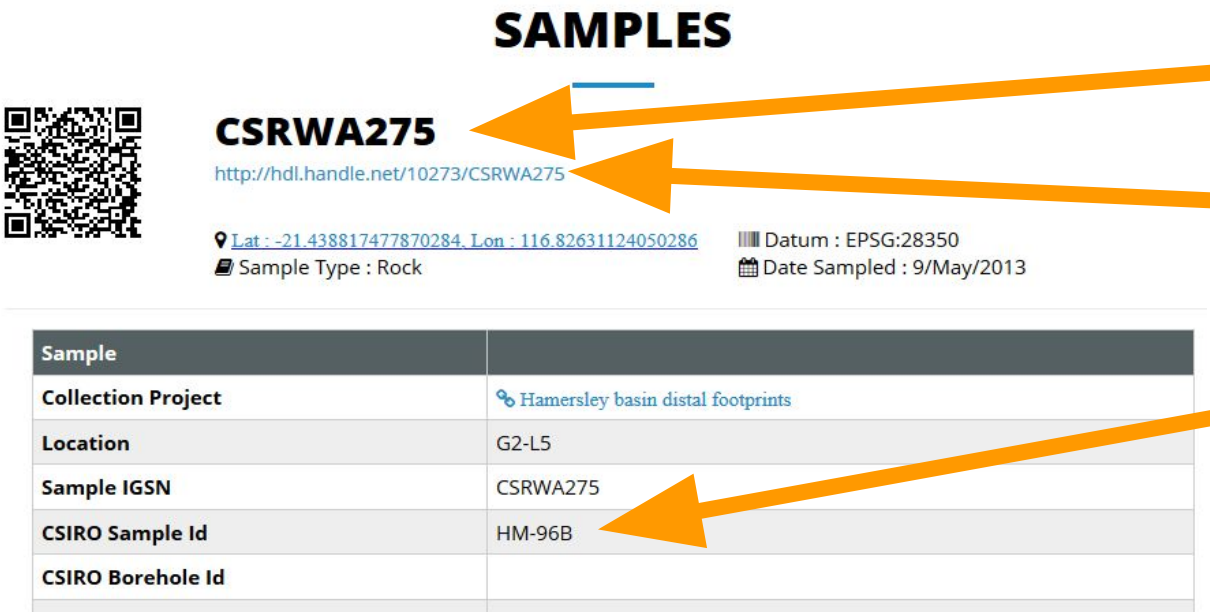
Persistent

- URLs change over time (“link rot”).
- A persistent identifier will always resolve to a landing page, even when the URL changes … or when URL will be superseded by a future technology.
- Examples: DOI, Handle, ORCID



Web Resolvable

URL of this page: https://rockstore.csiro.au/arrc/#/browsesamples/CSRWA275



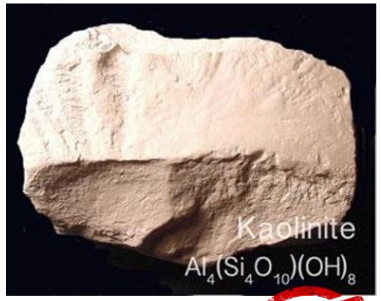
IGSN

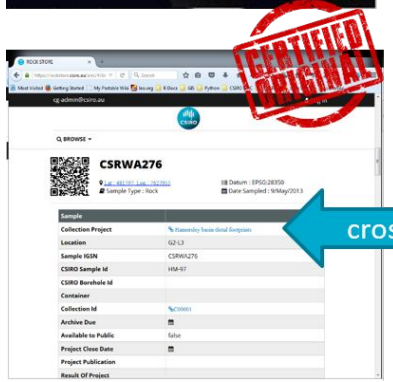
URL used to resolve IGSN

Sample name used in the field

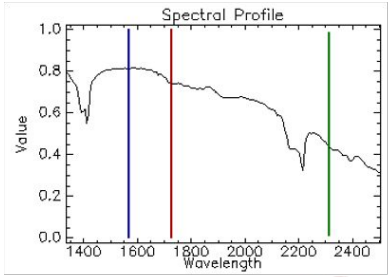
Linking Samples, Data, and Literature

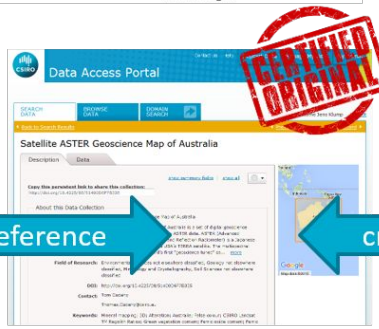
Specimen (IGSN)



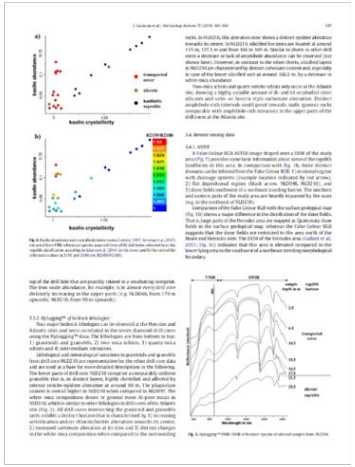


Spectrum (DOI)





Publication (DOI)



cross-reference

cross-reference

IGSN can link to other IGSN, to DOI (data, literature), or other PID. IGSN is a “related identifier” in DataCite metadata.

Tracking Samples

- CSIRO used IGSN as sample identifiers in a geochemical field campaign in the Nullarbor Desert.
- IGSN sample labels were pre-printed with QR codes.
- The QR codes were read into the FAIMS field data app.
- The streamlined process improved the efficiency of sample handling by 100%.

Scaling to Billions of Samples

- Historically IGSN grew out of the geological sciences but is now used for any kind of physical sample.
- Close to 10 million IGSN have been minted.
- IGSN might grow to 3 billion samples over the next ten years.
- How can we grow the system to accommodate this scale?







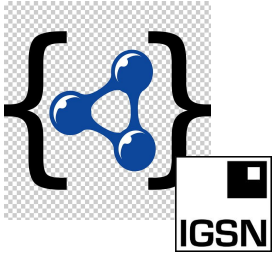


Is there a least common denominator?

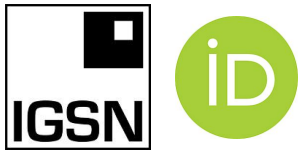
- What is the least common denominator to make the description of this object …
  - F – Findable
  - R – Reusable
- The description will depend on the intended use.
- Samples can come from anywhere in the universe, not only from Earth.

Technical Developments

- Let the machines do the heavy lifting.
- Persistent identifiers provide anchors to people, publications, data, code, samples, instruments, …
- Metadata become machine readable and can be harvested by standard web technologies.
- Linked data is finally coming of age (if it's done at scale and not done manually).



Linked Data for Physical Samples



IGSN

How Search Engines Work

Harvesting IGSN metadata in JSON-DL

- To be able to scale IGSN to billions of samples requires a shift to web technologies built to deal with such volumes.
- Search engine operators developed schema.org and sitemaps to enable web crawlers to efficiently harvest information from web resources.
- In future, metadata will be embedded in landing pages as JSON-LD.
- Crawlers find pages through sitemaps, and sitemaps of sitemaps.
- There are plenty of libraries around to deal with harvesting and parsing JSON-LD making implementation easier than OAI-PMH.
- JSON API will enable selective harvesting of metadata elements.

Linked data coming of age

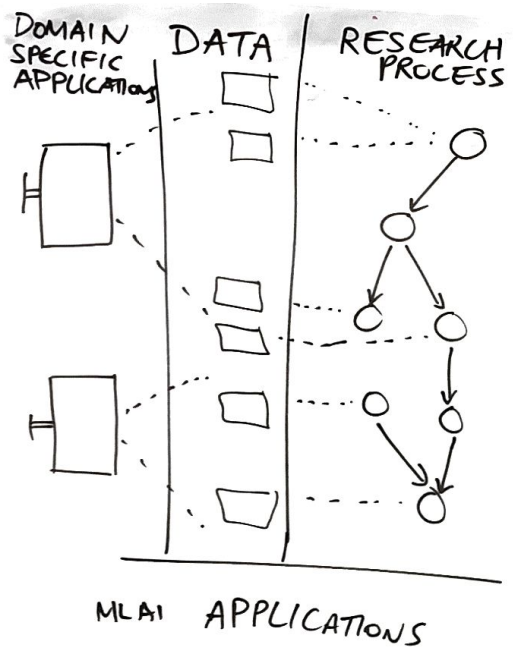
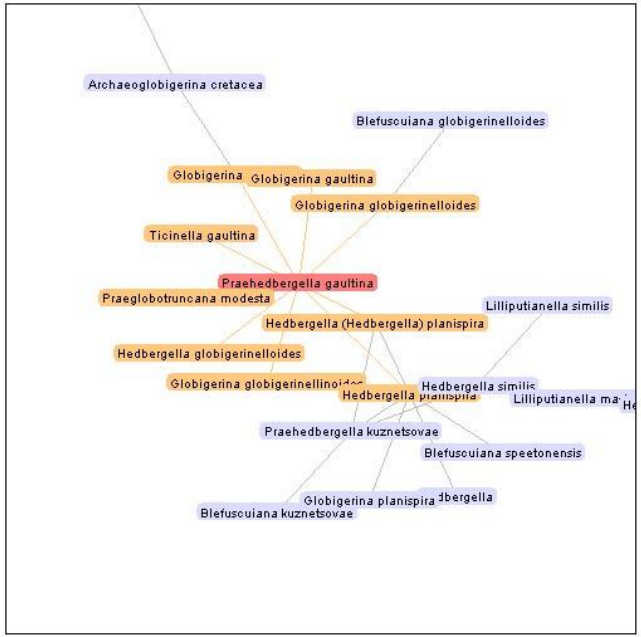


Figure courtesy of Alex Whan, CSIRO Agriculture & Food

- JSON-LD provides a framework for encoding metadata in a semantic way.
- These metadata can be generated by machines, using PID as anchors in the real world.
- The relationships between objects and concepts can be portrayed as graphs.
- Machine learning can be applied to graph analysis.

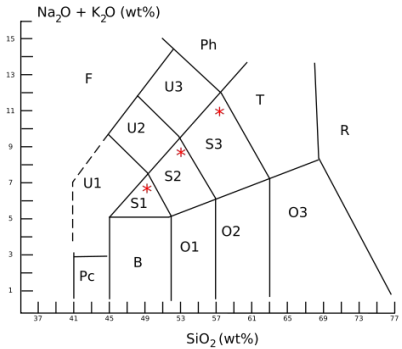
Why is graph learning important?

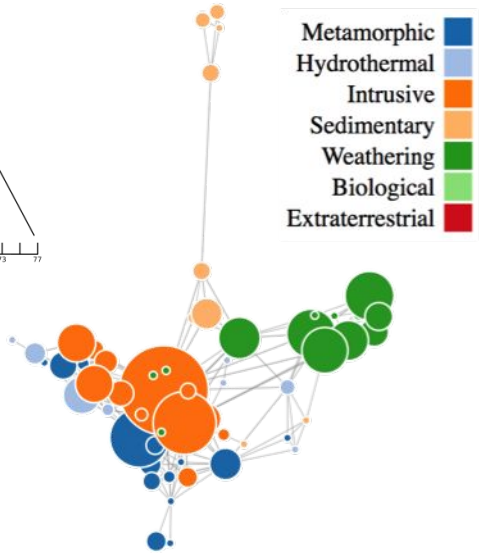


Huber & Klump, 2009

- Graphs can be analysed through algorithmic methods, e.g. pattern recognition, reasoning, etc.
- Availability of more training data led to better results in image analysis and natural language processing.
- Similar developments can be expected in graph-based learning.

New research capabilities



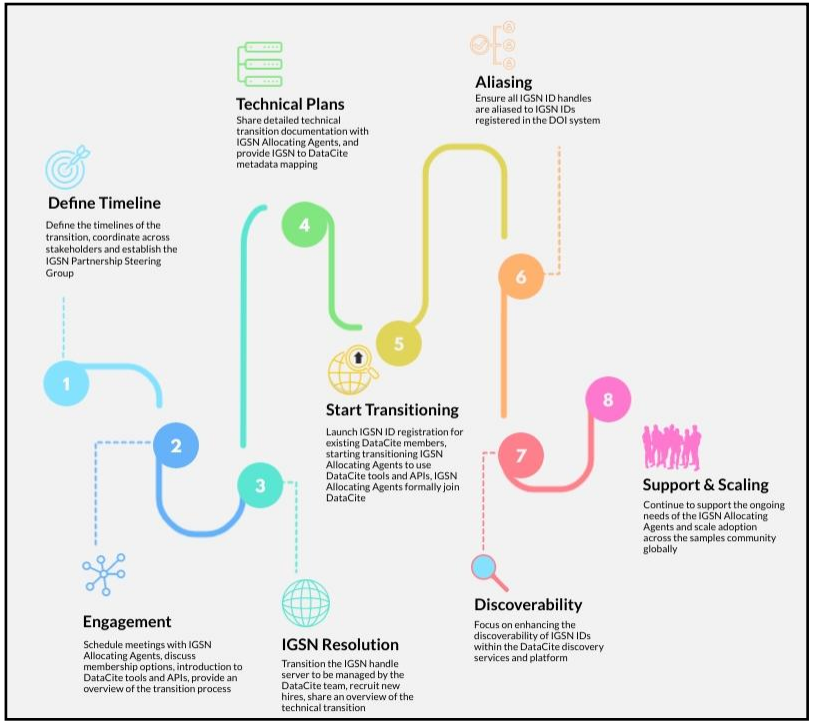


Hystad et al., 2019

- Exposing metadata as graphs through web architectures enables new research:
  - How are my samples related?
  - How are my mineral samples tied to groundwater chemistry?
  - What are their interactions with the soil microbiome?

IGSN / DataCite Partnership

- IGSN and DataCite entered a partnership agreement.
- DataCite will take on the minting of IGSN through DataCite services.
- IGSN PIDs will become DataCite DOIs.
- IGSN minting will be open to all DataCite members.



Summary

- Physical samples are at the heart of many research disciplines.
- IGSN allows persistent identification of physical samples across institutional boundaries.
- Through linking IGSN with DOI of data and literature, physical samples are linked to scholarly communication and become part of the record of science.
- Linking samples to the research knowledge graph enables new insights by analysing samples, data and literature in context.
- From 2022 onward, IGSNs will become DataCite DOIs.

Thank you

CSIRO Mineral Resources
Jens Klump
Jens.klump@csiro.au

Lamont-Doherty Earth Observatory, Columbia University
Kerstin Lehnert
lehnert@ldeo.columbia.edu

Lamont-Doherty Earth Observatory, Columbia University
Sarah Ramdeen
sramdeen@ldeo.columbia.edu

Australian National University / NCI
Lesley Wyborn
lesley.wyborn@anu.edu.au

IGSN e.V
http://www.igsn.org