



The IGSN Global Sample Number - A PID for Physical Objects

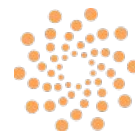
Jens Klump, Kerstin Lehnert, Sarah Ramdeen, Lesley Wyborn | 12 October 2021



Australia's National Science Agency



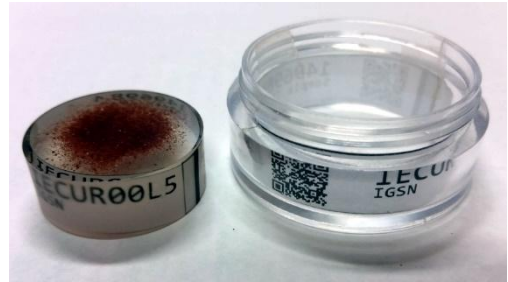
LAMONT-DOHERTY
EARTH OBSERVATORY
THE EARTH INSTITUTE AT COLUMBIA UNIVERSITY



NCI

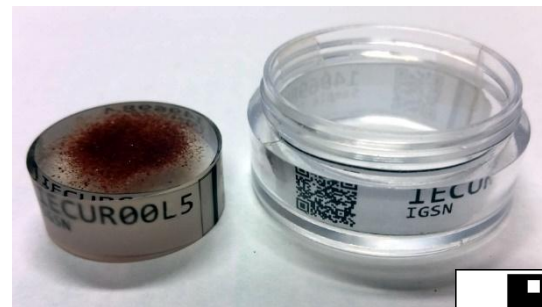


Samples are at the Heart of Research



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.



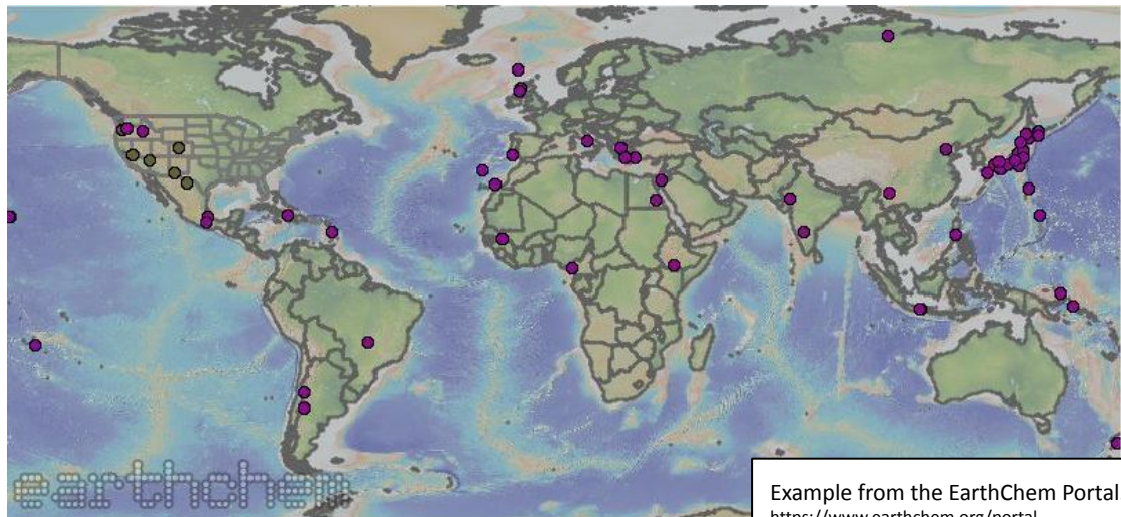
Globally Unique

<i>PetDB Identifier:</i>	ARGAMPH-003
<i>IGSN:</i>	N/A
	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)
<i>Other Names:</i>	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>



One sample, many names



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



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 or when URL will be superseded by a future technology.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

General Identifiers

Program:	ICDP
Expedition:	ICDP 5054
Type:	Hole
Name:	5054_1_A
IGSN:	ICDP5054EEW1001
Parent IGSN:	N/A
Release Date:	2017-4-1

Sampling Location

Latitude:	63.4063
Longitude:	13.203057
Coordinate System:	WGS84
Elevation:	522
Final Depth:	-1980.8
Location Type:	N/A
Location Name:	Are, Jaemtlands laen, Sweden
Location Description:	COSC-1 is located in the vicinity of the abandoned Froea mine
Country:	Sweden
Province:	Jaemtlands laen
County:	N/A
City:	Are

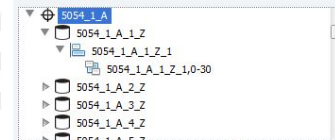
Geology

Material:	Rock
Rock Classification:	metamorphic rocks
From Corrected Depth:	102.7
To Corrected Depth:	2502.8
Depth Reference:	meter below ground level
Geological Age:	mid-paleozoic
Geological Unit:	N/A

Drilling

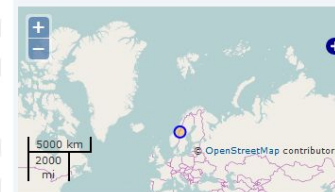
Drilling Method:	Coring>RockCorer wireline diamond coring, HQ and NQ bit size
Operator:	Lund University, Engineering Geology Larsson Drilling Consulting AB
Funding Agency:	Swedish Research Council (Vetenskapsrådet)
Total Length:	2400.1m

Sample Family



☑=Hole, ☐=Core, ☒=Core-Section, ☓=Core-Sample
The Sample Family shows a sub-sampling graph. Select entries to navigate samples. Core-Samples are issued to scientists on request. The naming convention for a Core-Sample is: *Expedition_Site_Hole_Core_Section_from-to(cm)*. Hole, Core, and Core-Section are following the same schema respectively.

Location Map



Drilling Start/End: 2013-9-5 / 2014-8-26 *
Latitude: 63.40630 * Longitude: 13.20306 *
Are, Jaemtlands laen, Sweden

Publications & Datasets

Lorenz, H., Rosberg, J.-E., Juhlin, C., Bjelm, L., Almqvist, B. S. G., Berthet, T., ... Tsang, C.-F. (2015). COSC-1: drilling of a subduction-related allochthon in the Caledonide orogen of Scandinavia. *Sci. Dr.*
doi:10.5194/sd-19-1-2015



Web Resolvable

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

SAMPLES



CSRWA275

<http://hdl.handle.net/10273/CSRWA275>

📍 [Lat : -21.438817477870284](#) [Lon : 116.82631124050286](#)

📄 Sample Type : Rock

📏 Datum : EPSG:28350

📅 Date Sampled : 9/May/2013

IGSN

URL used to resolve IGSN

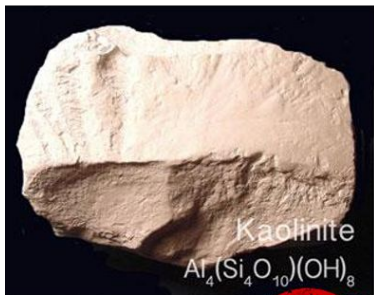
Sample name used in the field

Sample	
Collection Project	Hamersley basin distal footprints
Location	G2-L5
Sample IGSN	CSRWA275
CSIRO Sample Id	HM-96B
CSIRO Borehole Id	

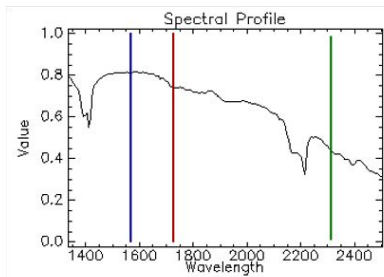


Linking Samples, Data, and Literature

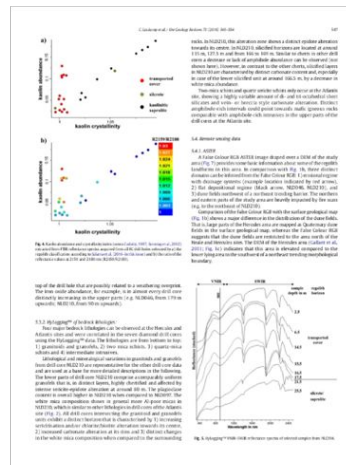
Specimen (IGSN)



Spectrum (DOI)



Publication (DOI)



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?



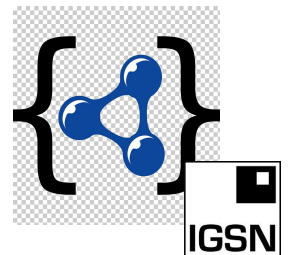
Is there a least common denominator?



- 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



How Search Engines Work



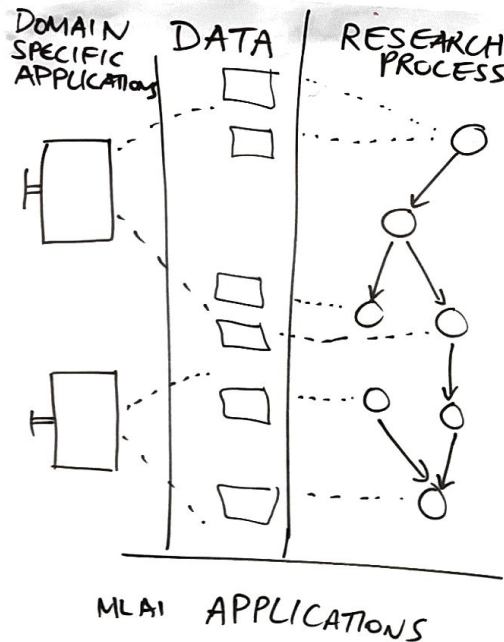


Harvesting IGSN metadata in JSON-LD

- 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.

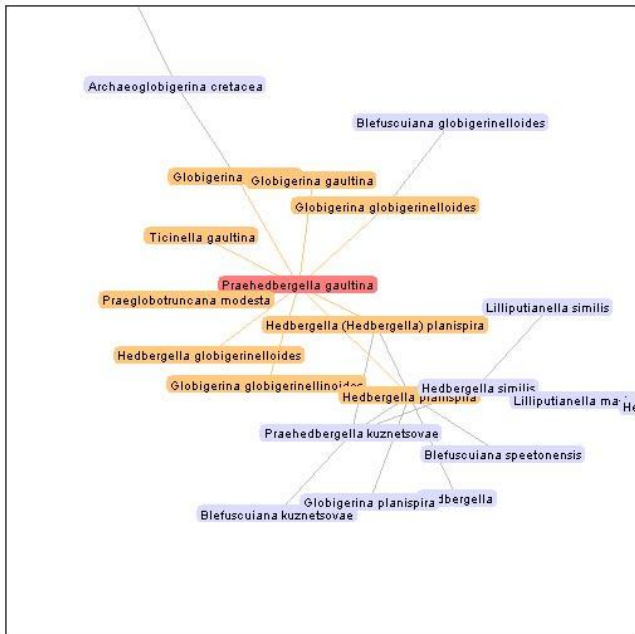


Linked data coming of age



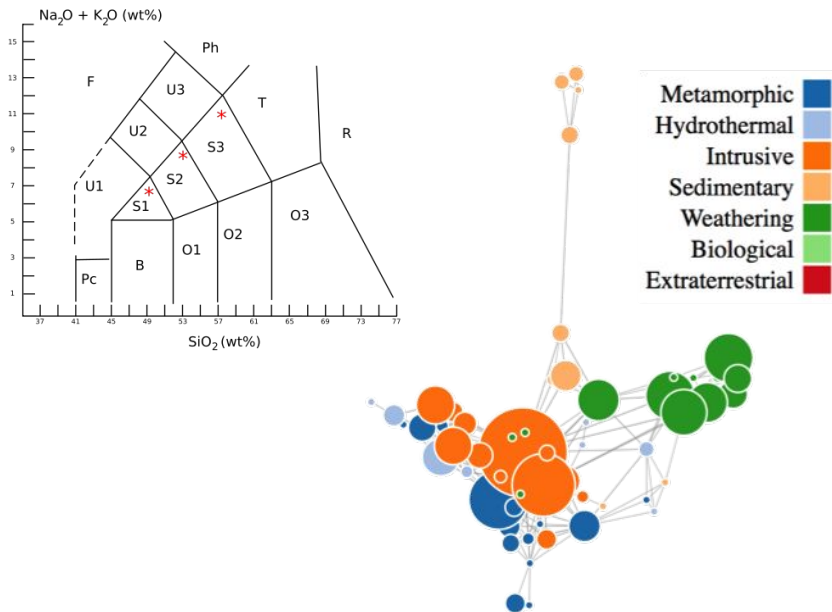
- JSON-LD provides a framework for encoding metadata in a semantic way.
- These metadata can be generated by machines, using PID as anchors in the real world.
- The relationships between objects and concepts can be portrayed as graphs.
- Machine learning can be applied to graph analysis.

Why is graph learning important?



- 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.

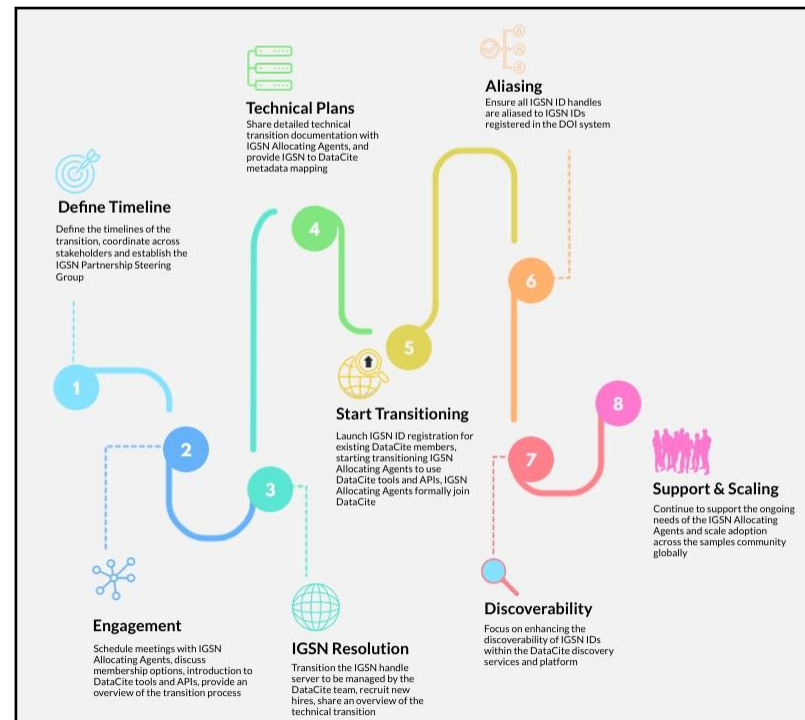
New research capabilities



- 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.





LAMONT-DOHERTY
EARTH OBSERVATORY
THE EARTH INSTITUTE AT COLUMBIA UNIVERSITY

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>

Australia's National Science Agency