Identifying and Describing Billions of Objects

An Architecture to Tackle the Challenges of Volume, Variety, and Variability

Jens Klump, Doug Fils, Anusuriya Devaraju, Sarah Ramdeen, Jess Robertson, Lesley Wyborn, Kerstin Lehnert

26 April 2023 | EGU General Assembly 2023 | EGU23-10223

MINERAL RESOURCES
www.csiro.au

IGSN

NIWA
Taihoro Nukurangi

Australian Research Data Commons

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

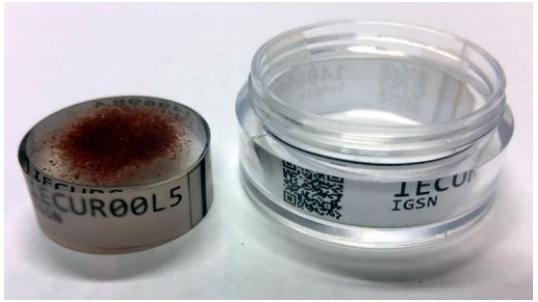
CSIRO

The Challenge of Volume

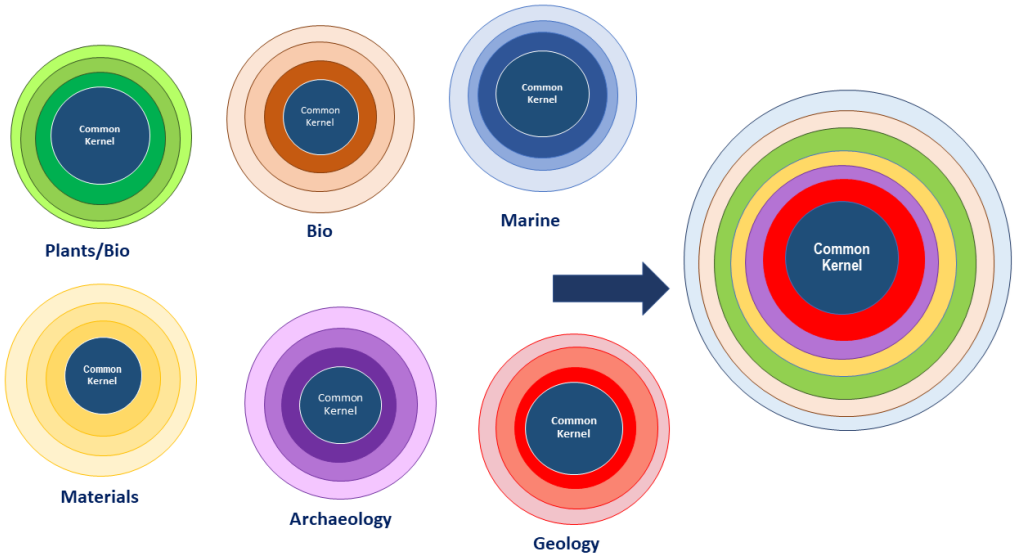
There are approx. 3 billion objects catalogued in natural history collections alone.

There are many more samples reported in the literature.

How do we make these vast holdings web-accessible and allow us to find the samples we are interested in?



The Challenge of Variety



Different communities have different needs for describing their assets.
The common kernel is only the least common denominator.
The system has to be able to handle multiple metadata schemas.

The Challenge of Variability

55637	Depth of reflector	Depth refl	m	100973
55638	Depth of Secchi Disk	z(SD)	m	6203
55639	Depth of stomatal pore	Depth stom pore	µm	143730
55640	Depth of the euphotic zone	z(eu)	m	84538
55641	Depth of thermocline	Depth therm	m	5339
55642	Depth, peat base	Depth peat base	cm	514745
55643	Depth, reconstructed	Reconstr depth	m	48383
55644	Depth, Redox discontinuity layer	RDL	cm	180538
55645	Depth, reference	Depth ref	m	1679
55646	Depth, relative	Depth rel	%	73790
55647	Depth, relative	Depth rel	m	102721
55648	Depth, sampling	Depth sampling	m	516531
55649	DEPTH, sediment, experiment	Depth sed exp	m	120637
55650	Depth, sediment, experiment, bottom/maximum	Depth sed bot	m	518762
55651	Depth, sediment, experiment, top/minimum	Depth sed top	m	518761
55652	Depth, sediment revised	Depth revised	m	18584
55653	DEPTH, sediment/rock	Depth sed	m	1
55654	Depth, sediment/rock, bottom/maximum	Depth sed bot	m	518756
55655	Depth, sediment/rock, top/minimum	Depth sed top	m	518755
55656	Depth, shift	Depth shift	m	146607
55657	DEPTH, soil	Depth soil	m	143506
55658	Depth, soil, maximum	Depth soil max	m	143512

pangaea_parameters

- Properties of an object can be expressed in many ways.
- Interoperability requires the use of controlled vocabularies with terms identified by URIs.
- Crosswalks between vocabularies facilitate semantic harmonisation and assist machine access and interoperability.

How do Search Engines Find Things?

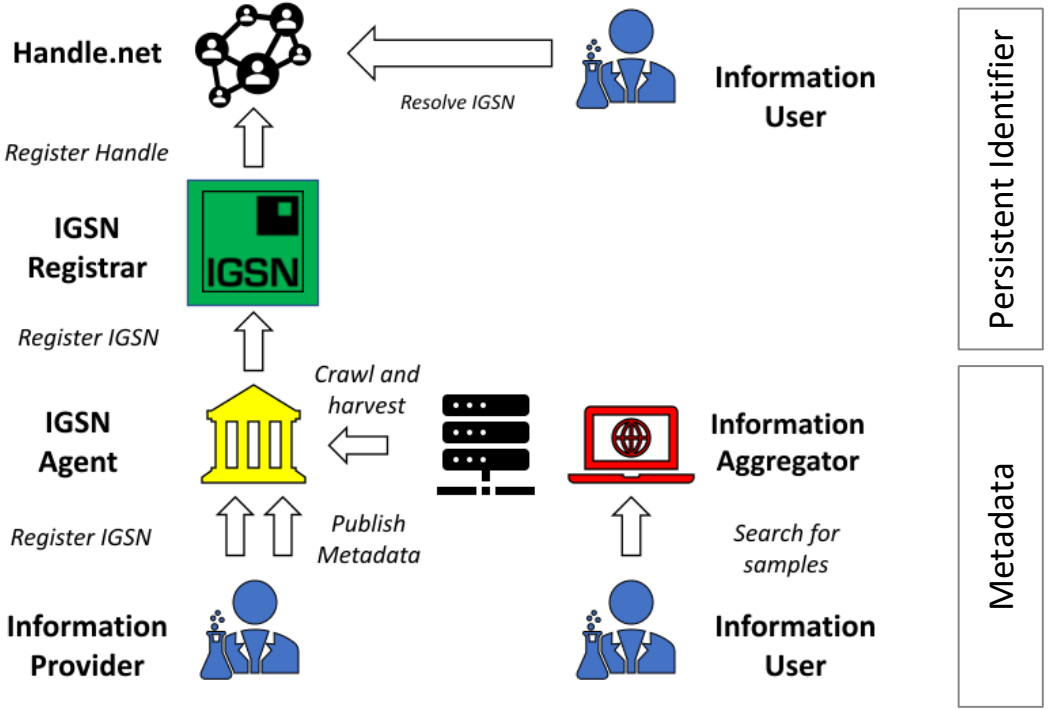
Schema.org

- Guidelines for using structured data mark-up on web pages.
- Formatted in JSON-LD (and other formats).
- Science on Schema.org (SOSO, ESIP cluster) provides guidance for schema.org markup in Dataset landing pages.

Sitemaps

- Sitemaps is a protocol in XML format meant for a webmaster to inform search engines about URLs on a website that are available for web crawling.
- Sitemaps can be registered with search engine operators.

The Architecture - IGSN as an Example



The Role of a Clearing House

Schema.org

- Publish recommendations on the use of metadata elements and controlled vocabularies.
- Provide a platform for communities for developing guidelines on the use of Schema.org in JSON-LD.

Sitemaps

- Publish recommendations on the use of sitemaps for assembling community specific catalogues.
- Collect and curate sitemaps sent by information providers and making them available to information aggregators.

Summary

- Compiling catalogues of vast numbers of objects on the internet faces the challenges of volume, variety, and variability.
- The challenges can be tackled by using the standard technology used by search engine operators.
- Community platforms such as the IGSN Organisation can act as clearinghouses and provide guidance.
  - Schema.org/JSON-LD metadata markup on landing pages and semantic harmonisation
  - Sitemaps to guide information aggregators to the landing pages relevant to their community and use case.
  - Provide a platform for developing community specific guidelines.

See also doi:10.5334/dsj-2023-005

Thank you!

This work was developed as part of the IGSN 2040 project and supported by the Alfred P. Sloan Foundation