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?
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

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