Identifying and Describing Billions of Objects: an Architecture to Tackle the Challenges of Volume, Variety, and Variability

Jens Klump¹, Doug Fils², Anusuriya Devaraju¹, Sarah Ramdeen³, Jesse Robertson⁴, Lesley Wyborn⁵, and Kerstin Lehnert⁶

¹CSIRO, Mineral Resources, Kensington WA, Australia (jens.klump@csiro.au)
²Ocean Leadership, Washington D.C., USA
³Ronin Institute for Independent Scholarship
⁴Ministry of Business, Innovation & Employment, Wellington, New Zealand
⁵Australian National University, Canberra ACT, Australia
⁶Lamont-Doherty Earth Observatory, Columbia University, Palisades NY, USA

Persistent identifiers are applied to an ever-increasing diversity of research objects, including data, software, samples, models, people, instruments, grants, and projects. There is a growing need to apply identifiers at a finer and finer granularity. The systems developed over two decades ago to manage identifiers and the metadata describing the identified objects struggle with this increase in scale. Communities working with physical samples have grappled with these challenges of the increasing volume, variety, and variability of identified objects for many years. To address this dual challenge, the IGSN 2040 project explored how metadata and catalogues for physical samples could be shared at the scale of billions of samples across an ever-growing variety of users and disciplines. This presentation outlines how identifiers and their describing metadata can be scaled to billions of objects. In addition, it analyses who the actors involved with this system are and what their requirements are. This analysis resulted in the definition of a minimum viable product and the design of an architecture that addresses the challenges of increasing volume and variety. The system is also easy to implement because it reuses commonly used Web components. Our solution is based on a Web architectural model that utilises Schema.org, JSON-LD and sitemaps. Applying these commonly used architectural patterns on the internet allows us not only to handle increasing volume, variety and variability but also enable better compliance with the FAIR Guiding Principles.