Building an Internet of Samples: The Australian Contribution

Lesley Wyborn (1), Jens Klump (2), Irina Bastrakova (3), Anusuriya Devaraju (2), Brent McInnes (4), Simon Cox (5), Linda Karssies (6), Julia Martin (7), Shawn Ross (8), John Morrissey (9), and Ryan Fraser (2)

(1) National Computational Infrastructure, ANU, Canberra ACT, Australia (lesley.wyborn@anu.edu.au), (2) CSIRO, Mineral Resources, Kensington WA, Australia, (3) Geoscience Australia, Canberra ACT, Australia, (4) John de Laeter Centre, Curtin University, Bentley WA, Australia, (5) CSIRO, Land & Water, Clayton VIC, Australia, (6) CSIRO, Land & Water, Canberra ACT, Australia, (7) Australian National Data Service, Canberra ACT, Australia, (8) Department of Ancient History, Macquarie University, Sydney NSW, Australia, (9) CSIRO, Information Management & Technology, Canberra ACT, Australia

Physical samples are often the ground truth to research reported in the scientific literature across multiple domains. They are collected by many different entities (individual researchers, laboratories, government agencies, mining companies, citizens, museums, etc.). Samples must be curated over the long-term to ensure both that their existence is known, and to allow any data derived from them through laboratory and field tests to be linked to the physical samples. For example, having unique identifiers that link back ground truth data on the original sample helps calibrate large volumes of remotely sensed data. Access to catalogues of reliably identified samples from several collections promotes collaboration across all Earth Science disciplines. It also increases the cost effectiveness of research by reducing the need to re-collect samples in the field. The assignment of web identifiers to the digital representations of these physical objects allows us to link to data, literature, investigators and institutions, thus creating an "Internet of Samples".

An Australian implementation of the "Internet of Samples" is using the IGSN (International Geo Sample Number, http://igsn.github.io) to identify samples in a globally unique and persistent way. IGSN was developed in the solid earth science community and is recommended for sample identification by the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS). IGSN is interoperable with other persistent identifier systems such as DataCite. Furthermore, the basic IGSN description metadata schema is compatible with existing schemas such as OGC Observations and Measurements (O&M) and DataCite Metadata Schema which makes crosswalks to other metadata schemas easy. IGSN metadata is disseminated through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) allowing it to be aggregated in other applications such as portals (e.g. the Australian IGSN catalogue http://igsn2.csiro.au). The metadata is available in more than one format. The software for IGSN web services is based on components developed for DataCite and adapted to the specific requirements of IGSN. This cooperation in open source development ensures sustainable implementation and faster turnaround times for updates.

IGSN, in particular in its Australian implementation, is characterised by a federated approach to system architecture and organisational governance giving it the necessary flexibility to adapt to particular local practices within multiple domains, whilst maintaining an overarching international standard. The three current IGSN allocation agents in Australia: Geoscience Australia, CSIRO and Curtin University, represent different sectors. Through funding from the Australian Research Data Services Program they have combined to develop a common web portal that allows discovery of physical samples and sample collections at a national level.International governance then ensures we can link to an international community but at the same time act locally to ensure the services offered are relevant to the needs of Australian researchers. This flexibility aids the integration of new disciplines into a global community of a physical samples information network.