Connecting Geology with the Internet of Things

Jens Klump1, Simon Cox2, Lesley Wyborn3

1 CSIRO Earth Science and Resource Engineering, Perth, Australia, jens.klump@csiro.au
2 CSIRO Land and Water, Melbourne, Australia, simon.cox@csiro.au
3 Geoscience Australia, Canberra, Australia, lesley.wyborn@ga.gov.au

ABSTRACT

Internet of Things refers to "uniquely identifiable objects (things) and their virtual representations in an Internet-like structure". Research in geosciences has created a tremendous wealth of data, material collections, and literature. Existing and emerging capabilities of geoinformatics can maximize the utility and impact of these resources. Major problems for achieving this include incomplete documentation of samples in publications, use of ambiguous sample names, and the lack of a central catalogue that allows finding a sample's archiving location. The International Geo Sample Number provides solutions by offering a persistent identifier for geological specimens that is interoperable with the DOI, a system well established for the identification of scientific literature and data. In linking physical specimens with data, the IGSN implements a component of the Internet of Things for the geosciences.

INTRODUCTION

Internet of Things is a term that refers to "uniquely identifiable objects (things) and their virtual representations in an Internet-like structure" (Wikipedia). We here use the term to describe new and innovative ways to integrate physical samples in the Earth Sciences into the emerging digital infrastructures that are developed to support research and education in the Geosciences.

Research in geosciences has created a tremendous wealth of data, material collections, and literature, and can be expected to produce even more in the future. Existing and emerging capabilities of Geoinformatics –in the sense of cyberinfrastructure for the Geosciences– can maximize the utility and impact of these resources to the benefit of science and education. Geoinformatics resources include (a) information systems that provide fast and easy access to data, (b) linkages between literature, data and samples that create the potential for new interpretations of the data and materials beyond interpretations already published, and (c) data analysis, visualization, and modelling tools that seamlessly integrate with the data collections.

Digital data such as the ones in EarthChem, PANGAEA, or other data repositories refer to physical samples. Fundamental data sets have been generated from collections of natural history museums or state geological surveys. The application and long-term utility of sample-based data is critically dependent on (a) availability of information (metadata) about the samples such as geographical location and time of sampling, (b) links to other data sets derived from individual samples that are dispersed in the literature and in digital data repositories, and (c) access to the samples themselves. Major problems for achieving this include incomplete documentation of samples in publications, use of ambiguous sample names, and the lack of a central catalogue that allows finding a sample's archiving location.

LINKING GEOSCIENCE DATA AND SPECIMENS INTO AN INTERNET OF THINGS

The International Geo Sample Number (IGSN, www.igsn.org) provides solutions. The IGSN is a unique persistent identifier for samples from geosciences research that can be obtained by submitting sample metadata to geosamples.org [1]. If data in publications are referenced to the IGSN, the geosamples.org database allows access to all sample metadata including the owner and current location. Use of the IGSN will, for the first time, allow to establish links between samples (or the digital representation of them), data acquired on these samples, and the publications that report these data. The IGSN is based on the Handle system, its technical and governance structure are modelled after DataCite. The IGSN is already used by some collections and drill core repositories and it has also been proposed that geochemical laboratories include it into their laboratory information management systems (LIMS).

Samples can be linked to a dataset by including IGSNs in the metadata record of a dataset's DOI® when the dataset is registered with the DOI® system for unique identification. Links between datasets and publications already have been implemented based on dataset DOIs® between some geoscience journals and data centres that are Publication Agents in the DataCite consortium (www.datacite.org) (see [2] for example of IGSN in a publication). Links between IGSNs, dataset DOIs, and publication DOIs will in the future allow researchers to find and access with a single query and without ambiguity all data acquired on a specific sample across the entire literature (Fig. 1).

To make data centres and scientific web portals effective ways of data sharing, scientists need to prepare their data for online publication. As an incentive to authors data publication should have the rank of a citeable publication, adding to his reputation and ranking among his peers. For data to be citeable it is necessary that they can be referred to in a persistent way. Since the location of internet resources (URL) may easily change, the use of persistent identifiers (e.g. Handle, DOI, URN, PURL, LSID, and others) as a way to locate the desired dataset reliably via the internet over a long time is a prerequisite for on-line data publication. While the technical aspects of a persistent identifier system are relatively simple, ensuring the global uniqueness and longevity of the PID-URL key-value pair requires a suitable governance structure [3].

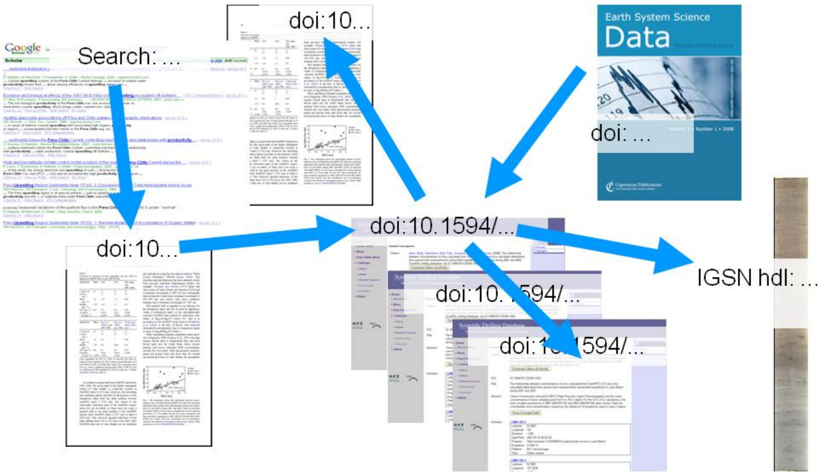

Figure 1: Persistent identifiers are keys to linking literature, data, and specimens. The semantic relationships between objects can be encoded in the identifier metadata.

A key learning emerging from the last ten years of ubiquitous networking in a department that bridges between research departments and infrastructure providers (Computing Centre, Library and Information Services) is that it is relatively easy to build isolated, monolithic, single-purpose applications for data management on a per-project basis (often termed “silos”, or more ironically “cylinders of excellence”). This approach is not only inefficient in the long run, but it generally removes the possibility of realizing the additional potential that emerges from repurposing data and applications across projects. It is particularly important to provide ongoing access to rare and expensive to obtain observational data, but the principle applies more generally as modern analysis and data mining techniques can uncover information in most recombined data.

The introduction of interoperable persistent identifier systems assists in the use of data through providing key elements to their provenance and thus provides us with the potential for new insights from repurposing of existing data. The identification of objects and agents can be used to track the provenance of data right to the physical specimen studied and the sensor used in the acquisition of data. In linking physical specimens with data, the IGSN implements a component of the Internet of Things for the geosciences.

REFERENCES

1. Lehnert, K.; Vinayagamoorthy, S.; Djapic, B.; Klump, J. The Digital Sample: Metadata, Unique Identification, and Links to Data and Publications. EOS Trans. Am. Geophys. Union 2006, 87, Abstract IN53C–07. http://www.agu.org/meetings/fm06/fm06-sessions/fm06_IN53C.html.
2. Dere, A. L.; White, T. S.; April, R. H.; Reynolds, B.; Miller, T. E.; Knapp, E. P.; McKay, L. D.; Brantley, S. L. Climate dependence of feldspar weathering in shale soils along a latitudinal gradient. Geochim. Cosmochim. Acta 2013, 122, 101–126. doi:10.1016/j.gca.2013.08.001.
3. Bütikofer, N. Catalogue of criteria for assessing the trustworthiness of PI systems; nestor-Materialien; Niedersächsische Staats und Universitätsbibliothek Göttingen: Göttingen, Germany, 2009; 18 p. http://nbn-resolving.de/urn:nbn:de:0008-20080710227.

ABOUT THE AUTHORS

Jens Klump has degrees in geology and oceanography from University of Cape Town and a PhD from Bremen University. In 2001 he joined the German Research Centre for Geosciences in Potsdam and got involved with the development of the publication and citation of research data through Digital Object Identifiers. This project sparked further work on research data infrastructures, such as enterprise data management systems and long-term digital archives. Jens' work on research data infrastructures focuses on the upstream components, such as automated data and metadata capture, sensor data integration, both in the field and in the laboratory, data processing workflows and provenance. Jens is the current vice-president of the International Geo Sample Number Implementation Organization. The organization coordinates the development and introduction of persistent identifiers for physical specimens of research materials. In early 2014 he was appointed OCE Science Leader for Earth Science Informatics in the CSIRO Earth Science and Resource Engineering. He is based in Perth WA.

Simon Cox trained as geophysicist, with a PhD from Columbia University. His work on informatics started with the Australian Geodynamics CRC, and he became involved in metadata standards on the Dublin Core Advisory Council. Work on XML standards for mineral exploration data led on to the GeoSciML project, and participation with the Open Geospatial Consortium, where he co-edited the Geography Markup Language standard. He developed Observations and Measurements as an OGC and ISO standard, which forms the basis for operational systems in diverse fields including air-traffic, water data transfer and environmental monitoring applications. He spent a year as a senior fellow at the EC Joint Research Centre in Italy working on integration of GEOSS and INSPIRE. He currently has leadership positions in the OGC, ISO/TC 211, the Research Data Alliance, and served on the council of the IUGS Commission for Geoscience Information and the International Association for Mathematical Geosciences. In 2006 he was awarded OGC's Gardels medal, and he presented the 2013 Leptoukh Lecture for the American Geophysical Union. Simon is currently based in CSIRO Land and Water in Melbourne, working on a variety of projects across environmental informatics, linked data and semantics.

Lesley Wyborn is a granite specialist by training and joined the then BMR in 1972. She has held a variety of positions as the organization changed to AGSO then GA. She has been involved in eResearch projects since 2000. With CSIRO, she helped develop the DIICCSRTE-funded AuScope Grid, Spatial Information Services Stack, Australian Spatial Research Data Commons and the Virtual Geophysics Laboratory Projects. She is a member of the Australian Academy of Science 'Data for Science Committee' and internationally she is on the Executive Committee of the American Geophysical Union Earth and Space Science Informatics Group and is associated with the NSF Geoinformatics for Geochemistry and EarthCube Projects.