Computers & Geosciences 96 (2016) 1–10

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Case study

Representing and publishing physical sample descriptions

Anusuriya Devaraju [a,*], Jens Klump [a], Simon J.D. Cox [b], Pavel Golodoniuc [a]

[a] CSIRO Mineral Resources, PO Box 1130, Bentley 6102, Western Australia, Australia
[b] CSIRO Land and Water, Private Bag 10, Clayton South, 3169 Victoria, Australia

ARTICLE INFO

Article history:
Received 18 December 2015
Received in revised form
28 April 2016
Accepted 26 July 2016
Available online 29 July 2016

Keywords:
Physical samples
Specimens
Metadata model
IGSN

ABSTRACT

This paper presents a metadata model for physical samples, developed by CSIRO for its role as an allocating agent. The model is essential for connecting various samples to the Web in a systematic manner. It serves as a basis for registering and publishing samples from researchers and laboratories in CSIRO with the International Geo Sample Number (IGSN). The model is simple, extensible and publicly available. We specify how existing controlled vocabularies are incorporated into the model development, and discuss their relevance and limitations. We also describe the mappings between the developed model and existing standards. This is necessary to extend the model's adoption across various science domains. The model has been implemented and tested in the context of two large sample repositories in CSIRO. The results demonstrate the effectiveness of the metadata model while maintaining its flexibility to adapt to various sample types.

1. Introduction

Physical samples (also called specimens) are important observational units in environmental sciences. For example, water samples are collected to assess the quality of surface-water and ground-water resources, animal tissues are obtained to determine the presence of pesticide residues, and sampling of soils, rocks and sediments is undertaken to target areas for mineral exploration. Samples are diverse. They may be collected and curated by various entities, e.g., researchers, laboratories, educational institutions, state agencies, industry associations and mining companies, and museums. Samples may not be easily discoverable due to their missing unique identification and the lack of standards and metadata catalogs representing them (Hills, 2015; Lehnert et al., 2006). For example, each curator may follow their own method of documenting samples, and different samples may hence have identical names. They may be renamed, but the changes are not recorded explicitly. Documentation for samples may be distributed across spreadsheets, legacy databases, or other proprietary data systems, and are not commonly available to users other than those that collected the samples. A research project may lose its scientific validity as a result of inaccessible and irrecoverable samples (Geological Society of America (GSA), 2012; Jeram, 1995). The original observational conditions cannot be reproduced without samples for subsequent examination, and consequently the project results cannot be repeated (Jeram, 1995).

The International Geo Sample Number (IGSN)[1] is a globally unique identifiers for physical samples. These identifiers are actionable such that they provide persistent link to sample descriptions on the Web (Fig. 1). IGSN is governed by an international implementation organization (IGSN e.V.). Clients, e.g., individual researchers and laboratories may obtain IGSNs for their samples via the registration service developed by an allocation agent. An allocating agent is the member institution that is authorized by the IGSN e.V. to operate a sample registration service. For further details about IGSN, see Section 2.1.

This paper presents a *description metadata model* (or schema) to represent physical samples. The model was developed by CSIRO for its role as an allocating agent. It is used to handle sample registration requests from CSIRO's clients (e.g., individual curators and larger collaborative projects), and to disseminate sample metadata through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Lagoze et al., 2002) implementation[2]. The metadata model was developed to meet the following requirements:

* Corresponding author.
E-mail address: anusuriya.devaraju@csiro.au (A. Devaraju).

[1] http://www.igsn.org/
[2] https://igsn.csiro.au/csiro/service/oai

http://dx.doi.org/10.1016/j.cageo.2016.07.018
0098-3004/

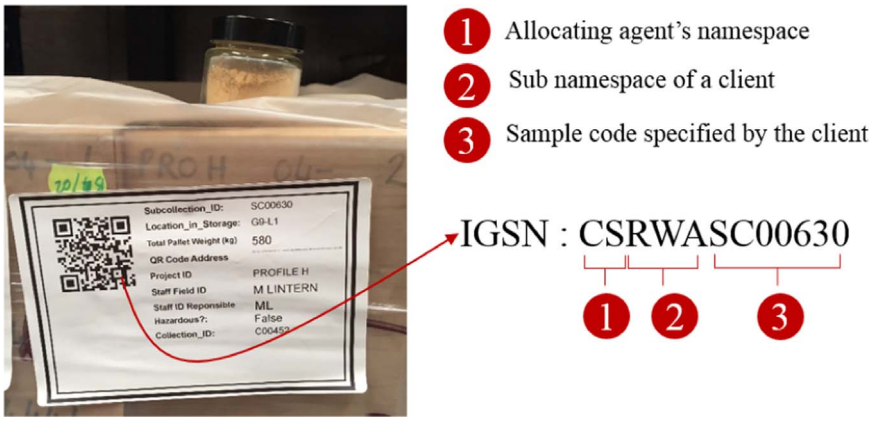

Fig. 1. An IGSN takes the form of an alphanumeric character divided into two elements: ⟨*namespace*⟩⟨*code*⟩. A *namespace* refers to the prefix of an allocating agent. It is assigned by the IGSN e.V. to the respective allocating agent, e.g., *CS* stands for CSIRO. An allocating agent ensures that the element ⟨*code*⟩ is unique within their namespaces. In the CSIRO implementation, a *code* is formed by a sub-namespace and then followed by a sample code (different strategies will be used by other allocating agents). A sub-namespace uniquely represents a client in the context of the allocating agent, whereas a sample code is the local sample identifier designated by the client. For example, *RWA* is the sub-namespace of the Rock Store. The Rock Store is a purpose built facility that curates collections of the Australian Resources Research Centre (ARRC). *SC00630* is the local code representing a sample collection. The resolvable handle URI for an IGSN is consists of a resolving service, the IGSN Handle prefix (10 273) and the IGSN, e.g., http://hdl.handle.net/10273/CSRWASC00630.

- *Generality*: Our use cases involve curating and publishing various physical samples and collections (e.g., rock, soil, vegetation, and water) in CSIRO. Therefore, the metadata model should be domain-independent. The focus is on representing common sample concepts in order to be general for modeling different sample types.
- *Extensible*: The metadata model should be flexible to define new concepts or modify existing concepts. This is important to accommodate new samples that will be collected by CSIRO and to promote the model's adoption by other allocating agents.
- *Validity & Practicality*: The model should be fully implemented and validated. It should be made accessible in a common format that is easier to use for clients and applied in the context of different use cases.

The rest of the paper is organized as follows. Section 2 provides an overview of IGSN and discusses the related work. Section 3 presents the metadata model. Section 4 covers the mappings between the model and existing standards. This is followed by the application of the metadata model (Section 5). Section 6 summarizes the contributions of the model and future work.

2. Related work

This section provides an overview of IGSN. It discusses existing work on describing and publishing physical samples with persistent identifiers.

2.1. IGSN Implementation

The main implementation organization (IGSN e.V.) developed a persistent identifier system for physical samples as DataCite[3] had not been founded at the time. Initially, Digital Object Identifiers (DOI) for datasets were registered through the German National Library of Science and Technology (TIB Hannover)[4] and sample registrations were outside the scope of the center. CrossRef DOI were ruled out due to the registration cost, e.g., USD 1 per DOI (J. Klump, personal communication, April 1, 2016).

Fig. 2 depicts the hierarchical architecture of the IGSN registration. IGSN e.V. operates the top-level registry and resolver services. It uses the Handle.net system (CNRI, 2010) to resolve each individual IGSN handle (e.g., 10273/CSRWASC00630) to a landing page[5] for the item denoted by the handle. The landing page contains more detailed (domain-specific) sample description. An allocating agent represents IGSN e.V. and is responsible for operating a registration service for clients, allocating sub-namespaces to its clients, maintaining a repository of descriptions of samples registered through the agent and providing interfaces for metadata harvesting. Examples of allocating agents are Commonwealth Scientific and Industrial Research Organisation (CSIRO), Interdisciplinary Earth Data Alliance (IEDA), Arizona State Geological Survey, Geoscience Australia and German Research Centre for Geosciences (GFZ). Clients may obtain IGSNs for their samples via the registration service managed by an allocation agent. They send IGSN requests to the agent's service based on the *description schema* developed by the respective allocating agent. The agent's service registers IGSNs with the top-level registry based on the *registration schema*[6]. The registration schema covers registration information (e.g., sample number, registrant and log), and excludes sample descriptions to allow much greater flexibility in describing samples for different use cases. Clients are also responsible for maintaining the landing pages of the registered samples.

2.2. Community efforts and metadata standards

The Internet of Samples in the Earth Sciences (iSamples)[7] is an EarthCube Research Coordination Network program that aims to preserve environmental samples through the use of cyber-infrastructures. One of the key aspects addressed by the program is identifying metadata profiles and tools required to facilitate sample discovery and interoperability across domains. Our work complements this aspect by developing a metadata schema that can be used to represent various physical samples, and relevant tools such as a sample registration service, a metadata store and interfaces for metadata harvesting.

When comparing the existing specifications for representing geospatial data, thus far little research has been undertaken on

[3] https://www.datacite.org/
[4] http://www.tib.eu/
[5] https://rockstore.csiro.au/arrc/#/browsesubcollections/CSRWASC00630
[6] http://schema.igsn.org/registration/
[7] http://earthcube.org/group/isamples