TOWARDS A WEB-ENABLED GEO-SAMPLE WEB: AN OPEN SOURCE RESOURCE REGISTRATION AND MANAGEMENT SYSTEM FOR CONNECTING GEO-SAMPLES TO THE WEB

Anusuriya Devaraju[1], **Jens Klump**[1], **Victor Tey**[1], **Simon Cox**[2], **Ryan Fraser**[1]

[1]CSIRO Mineral Resources,
PO Box 1130, Bentley, 6102 Western Australia, Australia.
E-mail: {anusuriya.devaraju, jens.klump, victor.tey, ryan.fraser}@csiro.au

[2]CSIRO Land and Water,
Private Bag 10, Clayton South, 3169 Victoria, Australia.
E-mail: simon.cox@csiro.au

ABSTRACT

Within the earth sciences the curation and sharing of geo-samples is crucial to supporting reproducible research, in addition to extending the use of the samples in new research, and saving costs by avoiding sample loss and duplicating sampling activities. In the Commonwealth Scientific and Industrial Research Organisation (CSIRO), researchers gather various geo-samples as part of their field studies and collaborative projects. The diversity of the samples and their unsystematic management led ambiguous sample numbers, incomplete sample descriptions, and difficulties in finding the samples and their related data. These problems are also found in universities, research institutes and government agencies, which usually curate and manage diverse samples. To address this problem, we developed an open source registration and management system to identify geo-samples unambiguously and to manage their metadata and data systematically. The system supports the linking of samples and sample collections to the real world features from where they were collected, as well as to their data and reports on the Web. This paper describes the implementation of the system including its underlying design considerations, and its applications. The system was built upon the International Geo Sample Number persistent identifier system with Semantic Web technologies. It has been implemented and tested with individual users and three sample repositories in the organization.

1. INTRODUCTION

Geo-samples are physical specimens such as rock, soil, sediment, water and vegetation. They are valuable research assets. They are essential for understanding the natural world, required to reproduce scientific experiments, and serve as physical proofs for prior work. For example, groundwater specimens are used to estimate nitrate concentration in groundwater,

and biological specimens verify the existence of a species in a particular location. Currently, CSIRO collects and curates a huge number of specimens from various disciplines, such as mineral resources, land and water, energy, agriculture, and manufacturing. These physical specimens also include specimens contributed by other institutions and agencies. For example, the organization manages the National Soil Archive, which is a long-term storage facility of soil specimens from state agencies and national research programs (Karssies et al., 2011). In the future, thousands of samples will be collected to support research activities in the organization.

Geo-samples and their associated data are not commonly available to users other than those that collected the samples. Users may not easily discover them due to several reasons (Lehnert et al., 2006, Devaraju et al., 2016). Sample collectors may follow inconsistent cataloging practices, such as assigning different names or identifiers to the same specimen or the same name or identifier to two or more specimens. They may also make transcription errors while labeling the samples, and the sample metadata may also be incomplete or not recorded in a standard manner. Finally, an online catalog for discovering samples from various sample repositories is not available. To address these issues, we developed an open source resource registration and management system. This development is vital to facilitating the processes of identifying samples, as well as managing and disseminating their information on the Web. The main contributions of this work are as follows:

1. We developed an open source system that supports the effective management and discovery of physical resources such as geo-samples and sample collections[1]. In this paper, we refer to geo-samples and sample collections as physical resources.
2. We adopted the International Geo Sample Number (IGSN) to support the globally unique and persistent identification of physical resources.
3. We enhanced existing specimen-related controlled vocabularies and developed additional vocabularies. We incorporated the vocabularies into the developed metadata schema to ensure consistency in metadata entries, and on the web portal to improve sample discovery.
4. We demonstrated the application and the relevance of the system in the context of three sample repositories and individual sample curators.

2. INTERNATIONAL GEO SAMPLE NUMBER (IGSN)

The International Geo Sample Number (IGSN)[2] is a persistent and unique alphanumeric code for identifying physical samples and sample collections. Figure 1 illustrates the IGSN's hierarchical governance structure, which consists of the Implementation Organization of the IGSN (IGSN e.V.), allocating agents, and clients. The Implementation Organization governs

[1] A collection may be a group of arbitrary specimens or large collections of fragments from the same specimen, e.g., a storage tray of drill core sections and rock chips.

[2] http://www.igsn.org/

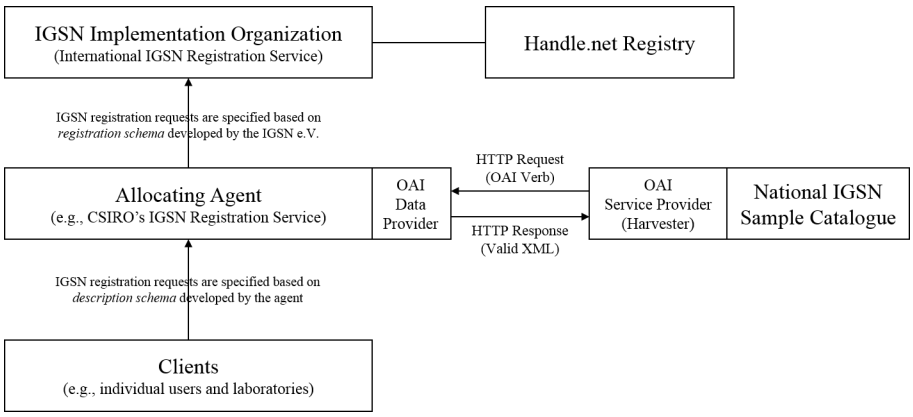

Figure 1: Hierarchical architecture of the IGSN.

and coordinates standards for identifying and citing physical samples, and operates the international IGSN registration service. The registration service uses the Handle.net System, which is a global persistent identifier resolver service (CNRI, 2010). The Handle.net resolves an IGSN of a sample (e.g., CSCAP876-MJ21)[3] to a landing page[4]. A landing page is the web page that contains the sample descriptions. Allocating agents are the member institutions that are authorized by the IGSN e.V. to register IGSNs within a permitted namespace. For example, CSIRO is one of three IGSN allocating agents in Australia besides Geoscience Australia (GA) and Curtin University. The IGSN e.V. allocated the namespaces 'CS', 'AU' and 'CU' to the allocating agents CSIRO, GA and Curtin University, respectively. Clients (e.g., individual researchers and laboratories) may register their samples with IGSN through an allocating agent. They also maintain the online landing pages of the samples they registered. In the CSIRO implementation, a client sends IGSN registrations to our registration service (allocating agent service) based on the *description schema* we developed. Then, the service forwards the registrations to the international IGSN registration service based on the *registration schema* developed by the IGSN e.V. (Figure 1). The registration schema covers minimal registration information (e.g., sample number, registrant, and datetime), whereas the description schema represents sample information such as identification, collection, curation, and related resources. This separation between registration information and description is important as it gives allocating agents greater flexibility in describing samples for different applications.

3. RESOURCE REGISTRATION AND MANAGEMENT SYSTEM

Figure 2 illustrates the architecture of the resource registration and management system. There are two ways to interact with the system. Individual users (e.g., sample collectors) may use a web-based user interface (Figure 3) to register the samples and manage their meta-

[3]To align with practices in DOI, ORCID, etc., IGSN now uses http://igsn.org/{igsn}, redirected to http://hdl.handle.net/10273/{igsn} as the URI from.

[4]https://capdf.csiro.au/data/?igsn=CSCAP876%2dMJ21

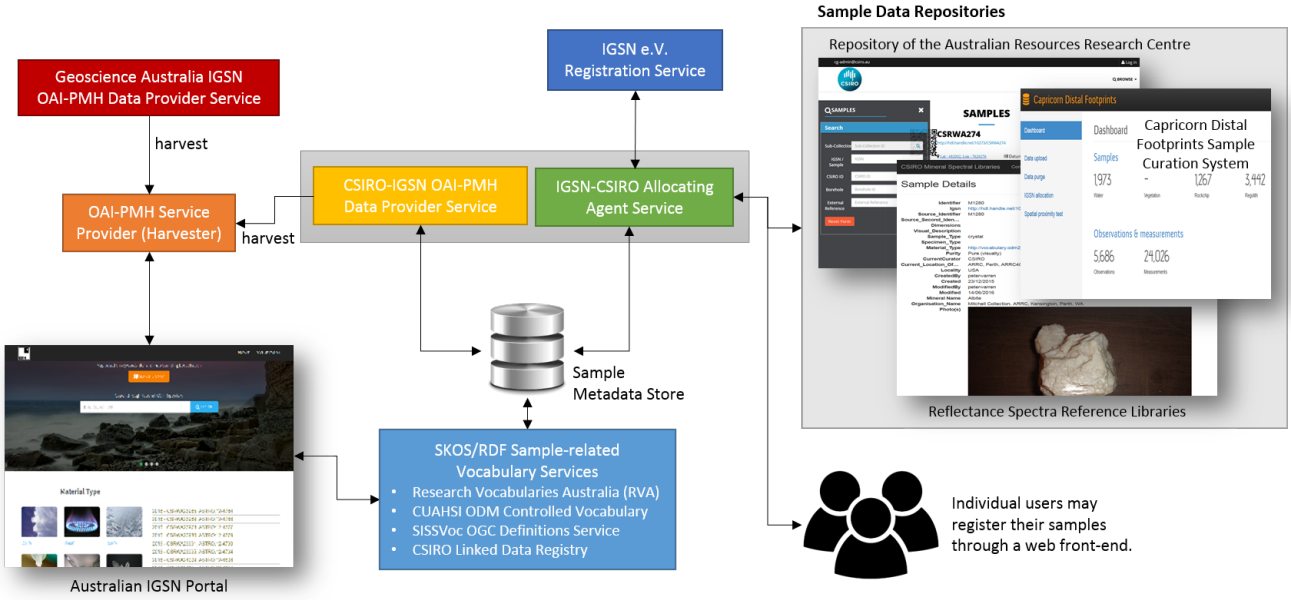

Figure 2: The architecture of the resource registration and management system.

data, whereas sample data repositories may perform the same operations programmatically, e.g., Capricorn Distal Footprints, Repository of the Australian Resources Research Centre (ARRC), and Reflectance Spectra Reference Libraries. Table 1 provides an overview of the samples registered through our system. The following are the components of the system developed. The links of the components are listed in Table 3. Their source repositories are available on the Github.[5]

a. **IGSN-CSIRO Allocating Agent Service.** The allocating agent service is a RESTful web service. Clients may authenticate to the service through the HTTP Basic Authentication. The allocating agent service enables clients to register their samples with IGSNs, retrieve and update metadata associated with registered samples, and obtain a list of all the prefixes of the IGSNs registered. Figure 4 illustrates the workflow of a client's sample registration through the allocating agent service.
In the CSIRO implementation, the IGSN of a sample (e.g., CSCAP876-MJ21) consists of a *prefix* and a *suffix*. A *prefix* is formed by an allocating agent's *namespace*, followed by a *sub-namespace* representing the agent's client. For example, in the prefix 'CSCAP', the namespace 'CS' refers to the allocating agent (CSIRO), and its sub-namespace 'CAP' denotes its client (Capricorn Distal Footprints). The suffix (e.g., 876-MJ21) in the example above refers to the local sample code specified by the client. In our system, clients may only register IGSNs with the sub-namespaces allocated to them. With the hierarchical namespace delegation pattern, we can systematically manage the allocation of specific namespace for different clients in the organization, while at the same time ensuring the global uniqueness of the sample identifiers.

[5] https://github.com/AuScope/igsn30

b. **Description Metadata Schema.** The IGSN requests sent by a client to the agent's registration service must be specified in XML, conforming to the description metadata schema we developed (Figure 5). The key features of the metadata schema are as follows:

- It covers the common concepts associated with geo-samples such as identification, sampling activity, curation, and related resources, so that it can be used to catalogue different sample types in the organization.
- It supports a batch registration as our applications require large batches of IGSN registrations.
- It has minimal restrictions on mandatory metadata elements (see Figure 5, rectangles with solid border). Some of these elements are required to obtain IGSNs from the international registration service (e.g., *resourceIdentifier* an *landingPage*), while the others are relevant for discovering samples (e.g., *materialTypes* and *curationDetails*).
- It captures the provenance of sample curation as geo-samples are often relocated from repository to another.
- It is also flexible in terms of representing spatial (i.e., coordinates and toponyms) and temporal information (i.e., instants and intervals based on the W3C Date and Time Formats). A coordinate reference system may be specified based on the EPSG Geodetic Parameter Dataset[6].
- It represents several relation types that allow a client to associate a sample or a collection with its related resources through their URIs, such as sub-sample, reference sample, dataset and publication. The information about related resources may exist in the client's sample data curation systems, or in institutional digital repositories such as the CSIRO's Data Access Portal[7] and the Publication Repository[8].

c. **Metadata Store.** The allocating agent service registers IGSNs with the international IGSN registration service and stores the sample metadata (extracted from the description XMLs sent by a client) in a PostgreSQL database. The database also captures client information, such as sub-namespace and authentication details.

d. **SKOS Controlled Vocabularies.** The description metadata schema leverages existing and new controlled vocabularies, which are expressed using the Simple Knowledge Organization System (SKOS). For a list of the vocabularies used by the system, see Table 2. Through the collaboration with the Geoscience Australia, we use the Australian National Data Service (ANDS) Vocabulary Service[9] to develop and maintain the new controlled vocabularies. Concepts in the vocabularies are identified with their persistent URIs in order to ensure machine actionability to the concepts. We include the URIs of the concepts in the description metadata schema in order to provide standardized information about permitted values for specific metadata elements, in addition to ensuring consistency in metadata entries. We also apply the vocabularies for searching and browsing samples on the National IGSN web portal.

[6] https://epsg.io
[7] https://data.csiro.au/
[8] https://publications.csiro.au
[9] http://www.ands.org.au/online-services/research-vocabularies-australia

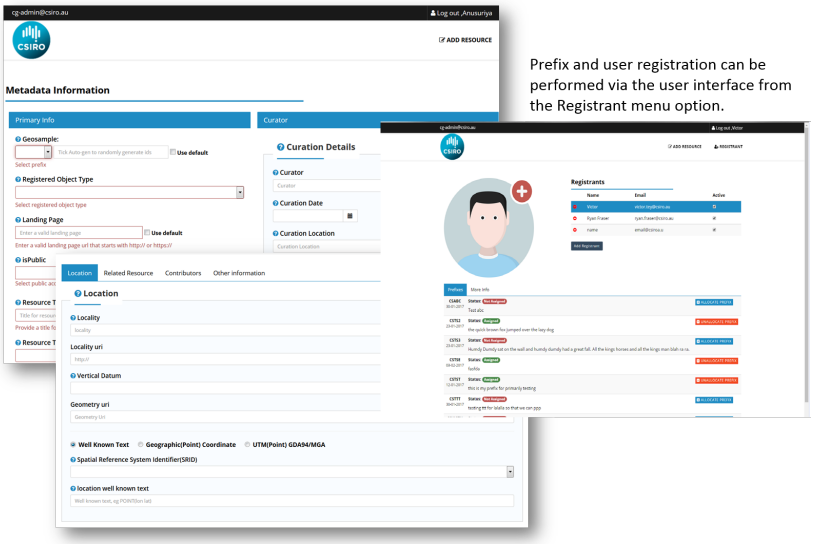

Figure 3: A web-based user interface for registering and managing physical resources.

Table 1: Local sample data repositories and their IGSN registrations (as at 31.05.2017)

Sample Data Repositories	Material Types	Registered samples
Repository of the ARRC	rock, mineral, soil	25652
Capricorn Distal Footprints	rock, vegetation, water, regolith	4232
Reflectance Spectra Reference Libraries	mineral, rock, synthetic material	94

e. **Metadata Harvesting and Discovery.** The OAI-PMH is a protocol for harvesting metadata catalogues from digital repositories (Lagoze et al., 2002). In this protocol, a *data provider* offers a catalogue of the repository holdings following the specifications of OAI-PMH, whereas a *service provider* operates a harvester that gathers metadata from one or more digital repositories (Lagoze et al., 2002). We implemented an OAI-PMH data provider service to disseminate the sample metadata records from the metadata store. We also developed an OAI-PMH service provider based on the PANGAEA Framework for Metadata Portals (panFMP) (Schindler and Diepenbroek, 2008). The service provider harvests metadata records from our metadata store through the data provider service, and from other allocating agents, e.g., Geoscience Australia. We developed the National IGSN web portal[10], which provides a common access to the sample metadata records harvested from different allocating agents.

[10] http://igsn.org.au

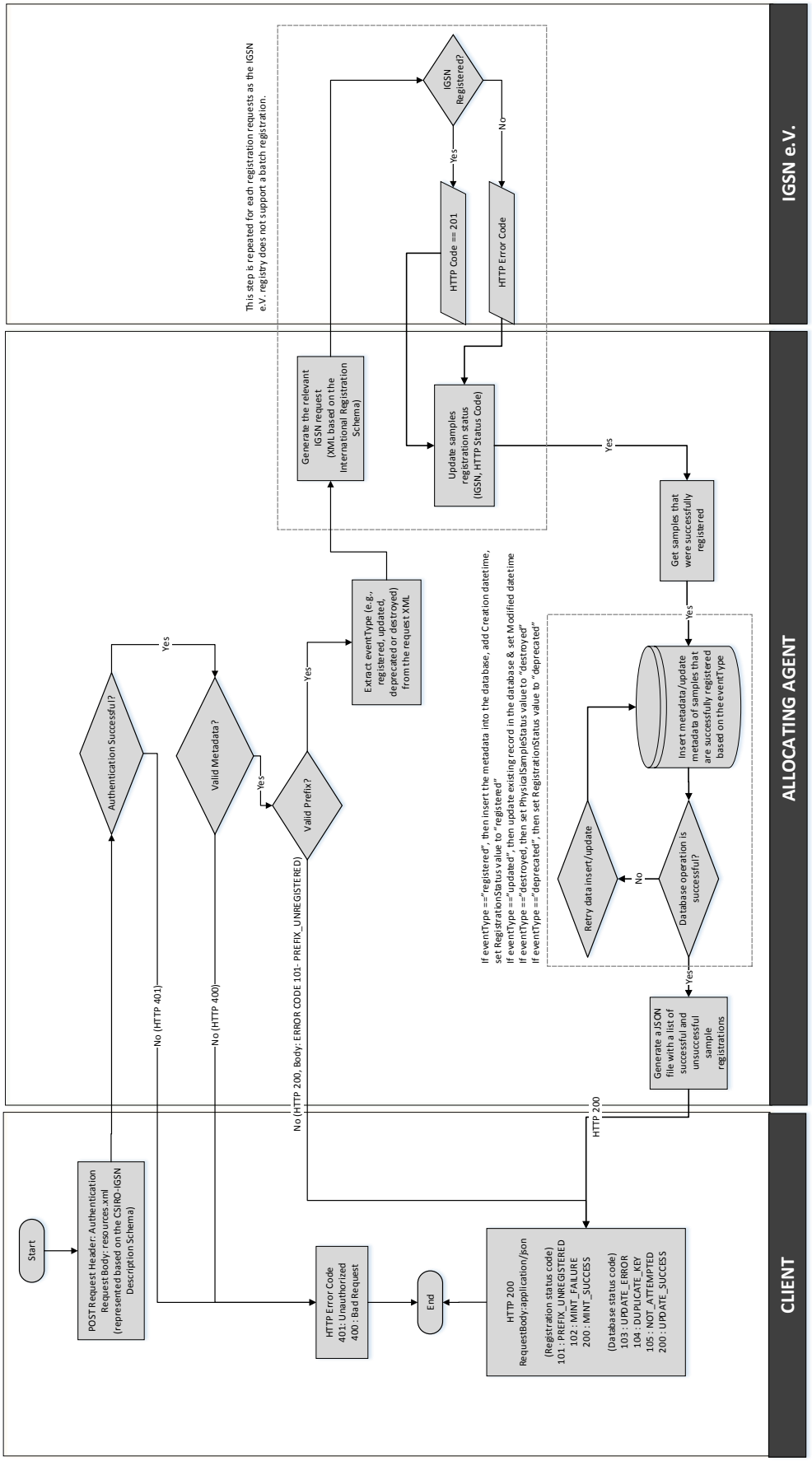

Figure 4: IGSN sample registration.

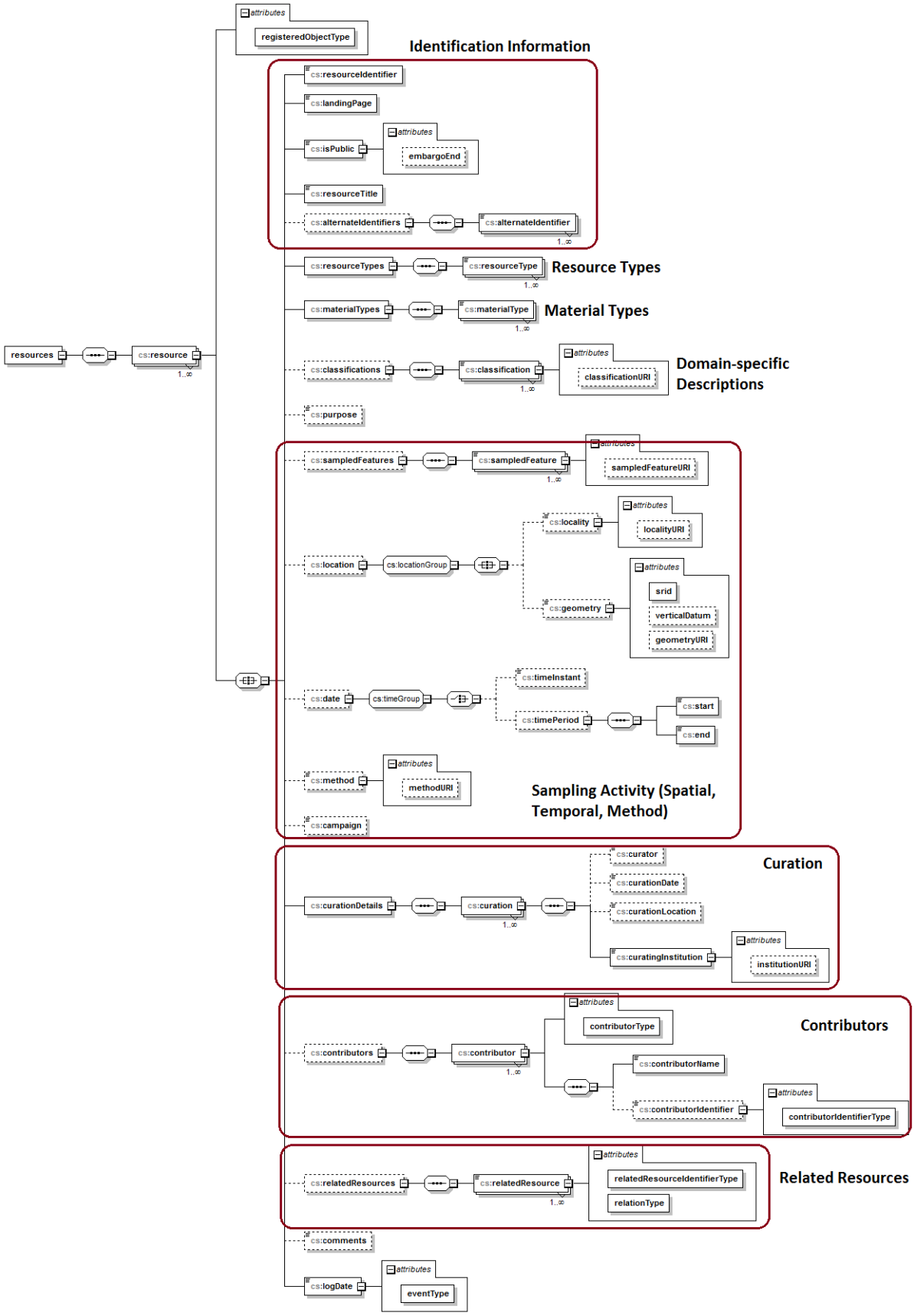

Figure 5: The CSIRO-IGSN description metadata schema (version 3.0).

Table 2: A list of existing and newly developed SKOS vocabularies. New vocabularies are indicated by asterisks*.

Vocabularies	Examples	Service	Provider
Material types[11]	soil, rock, vegetation.	Master Controlled Vocabulary Registry for ODM2	CUAHSI
Specimen types[12]	thin section, grab, dredge, cuttings.	Master Controlled Vocabulary Registry for ODM2	CUAHSI
Nil-reason types[13]	missing, unknown, withheld.	OGC definitions of nil reasons	CSIRO
Contributor types[14]	originator, custodian, point of contact.	Linked Data Registry	CSIRO
*Registration types[15]	physical sample, sample collection, sampling features.	ANDS Research Vocabularies Australia	CSIRO&GA
*Identifier types[16]	DOI, IGSN, LSID, ORCID	ANDS Research Vocabularies Australia.	CSIRO&GA
*Relation types[17]	isDerivedFrom, hasDocument, hasDigitalRepresentation.	ANDS Research Vocabularies Australia	CSIRO&GA

Table 3: A list of the system's components and their links.

Technical Components	Link
CSIRO Allocating Agent Service	https://igsn.csiro.au/igsn30/api
Description Metadata Schema	https://igsn.csiro.au/schemas/3.0/
Metadata Store (source)	https://github.com/AuScope/igsn30/tree/master/sql
CSIRO-IGSN OAI-PMH Data Provider	https://igsn.csiro.au/igsn30/api/service/30/oai
National IGSN Web Portal	http://igsn.org.au
OAI-PMH Harvester and National IGSN Web Portal (source repositories)	https://github.com/AuScope/NatPortalIGSN
Sample Registration and Management Web User Interface	https://igsn.csiro.au/igsn30

4. APPLICATIONS

The system has been tested with individual researchers and three sample data curation systems. The Capricorn Distal Footprints is a collaboration project between CSIRO, government agency, university and industry, which addresses the issue of exploration through cover by examining the geophysical and geological footprints of ore deposits at multiple scales across the Capricorn Orogen in Western Australia (Pearce et al., 2015). The ARRC Repository is a purpose-built facility managed by the CSIRO Business and Infrastructure Services (CBIS). The facility archives all kinds of legacy samples and collections (primarily rocks and minerals) gathered as part of research and development at the organization since the 1970s. The Reflectance Spectra Reference Libraries are used to assist interpretation of hyper- and multi-spectral remote and proximal sensing datasets (Laukamp et al., 2015). The following are applications of the developed system:

Tracking samples from the field to the repository. The CAPDF project uses a mobile field data collection application (Golodoniuc et al., 2017) as part of its field sampling activities. The local identifiers generated by the mobile application are prepended with a relevant namespace (CSCAP), and are then registered through our allocating agent service. This registration process takes place automatically when the records from the mobile application are uploaded into the CAPDF sample data repository. This gives the users an

[11] http://vocabulary.odm2.org/medium/
[12] http://vocabulary.odm2.org/specimentype/
[13] http://www.opengis.net/def/nil/OGC/0/
[14] http://registry.it.csiro.au/def/isotc211/CI_RoleCode
[15] http://registry.it.csiro.au/def/isotc211/CI_RoleCode
[16] pid.geoscience.gov.au/def/voc/igsn-codelists/identifierType
[17] pid.geoscience.gov.au/def/voc/igsn-codelists/relationType

advantage by ensuring consistent use of globally unique sample identifiers after the collection of samples. As a result, samples can be tracked easily at different stages of their life cycle, e.g., specimen handling and storage, laboratory analysis, and eventual disposal.

Asset management. Before we built the system, users had no way of knowing where to look for samples of interest in the sample repository, e.g., ARRC Repository. Each sample collector used their own system to label their samples and sample collections, and this resulted in duplicate or ambiguous labels. Previously, most of the labels on the sample containers were hand-written and without systematic sample identifiers, which may become impossible to place into context after a number of years. As a solution, existing samples and sample collections in the sample repository are registered with IGSNs, thereby allowing their unique identification. We printed QR code labels with relevant IGSNs and affixed them to the sample containers. As a result, users are now able to identify samples in a container by scanning its QR code, which then directs the users to a web page providing the sample information. Users may also use the sample discovery web portal to search for a sample or a sample collection, as well as its storage location (e.g., rack number) in the sample repository.

Web-based discovery of geo-samples and related resources. Previously, there was no way to locate samples from different sample data repositories, unless users searched these repositories individually. Through the sample discovery web portal, users are now able to discover the samples and their associated metadata from both the internal and external sample data repositories. This common access to samples can help to avoid duplicate field trips, create opportunities to re-use existing samples, and promote the reproducibility of sample-based data.

Outlook. Following the successful implementation of the system in CSIRO, we are now collaborating with the John De Laeter Centre for Isotope Research at Curtin University to adapt the system in the context of their Digital Mineral Library.

5. CONCLUSIONS

In Earth and environmental sciences, there has been extensive open source development for curating and publishing digital collections (e.g., datasets and documents). Nevertheless, there remains relatively little work on facilitating a globally unique and persistent access to geo-samples and their related data on the Web. To address this challenge, we developed an open source system to support the publication and management of geo-samples and sample collections in CSIRO. It comprises technical (e.g., client applications, metadata schema, metadata store, harvesting capabilities, web portal) and non-technical components (e.g., identifier governance, sample workflows). We built the system to accommodate the needs of both individual sample curators as well as sample repositories. The system assures the uniqueness of the samples, and connects their metadata and data systematically to the Web. The system also offers support to harvest sample metadata records from different sources, which can be aggregated to create new applications, such as the Australian IGSN sample discovery portal.

Through the applications of the system, we demonstrated that it is well-suited for a large organization, in which individual users, projects, and laboratories may all have different requirements for publishing and managing their specimens and collections. Using the development, we can now identify geo-samples unambiguously and discover them easily, and consequently avoid duplicate sampling activities and promote re-use of the samples for new purposes.

Acknowledgments

The IGSN implementation in CSIRO is part of the Research Data Services (RDS) project funded by the Department of Education as part of their Education Investment Fund (EIF) Super Science Initiative. The Capricorn Distal Footprints was funded by the Science and Industry Endowment Fund as part of The Distal Footprints of Giant Ore Systems: UNCOVER Australia Project (RP04-063).

References

Karssies, L., Jacquier, D., Wilson, P., Ringrose-Voase, A., Archive, C. N. S., Program, A. C. L. E., CSIRO National Soil Archive Manual, CSIRO Land and Water, URL www.clw.csiro.au/aclep/documents/CSIRO_National_Soil_Archive_Manual.pdf, 2011.

Lehnert, K. A., Vinayagamoorthy, S., Djapic, B., Klump, J., The Digital Sample: Metadata, Unique Identification, and Links to Data and Publications, EOS, Transactions, American Geophysical Union 87 (52, Fall Meet. Suppl.) (2006) Abstract IN53C–07, URL http://abstractsearch.agu.org/meetings/2006/FM/sections/IN/sessions/IN53C/abstracts/IN53C-07.html.

Devaraju, A., Klump, J., Cox, S. J., Golodoniuc, P., Representing and publishing physical sample descriptions, Computers & Geosciences 96 (2016) 1 – 10, ISSN 0098-3004, URL http://dx.doi.org/10.1016/j.cageo.2016.07.018.

CNRI, Technical Manual, Handle.Net Version 8.1, Corporation for National Research Initiatives (CNRI), URL http://handle.net/tech_manual/HandleTool_UserManual.pdf, 2010.

Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S., Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting, Tech. Rep., URL http://www.openarchives.org/OAI/2.0/guidelines.htm, 2002.

Schindler, U., Diepenbroek, M., Generic XML-based framework for metadata portals., Computers & Geosciences 34 (12) (2008) 1947–1955, URL https://doi.org/10.1016/j.cageo.2008.02.023.

Pearce, M. A., Hough, R., Ley, Y., Spinks, S. C., Thorne, R., White, A. J., Golodoniuc, P., Gray, D., Munday, T., The Capricorn Distal Footprints Project: An Orogen-scale approach to mineral systems, Proceedings of the GAC-MAC-CGU-AGU Meeting .

Laukamp, C., Lau, I., Mason, P., Warren, P., Huntington, J., Green, A., Whitbourn, L., Wright, W., Connor, P., CSIRO Thermal infrared spectral library - Part 1: Evaluation and status report 2015, Tech. Rep., CSIRO, URL https://doi.org/10.4225/08/5852dae88924b, 2015.

Golodoniuc, P., Devaraju, A., Klump, J., Case study: Curation and publication of physical samples using persistent identifiers, in: Geophysical Research Abstracts, vol. 19, EGU General Assembly 2017, 2017.