Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings

Volume 15 *Seoul, South Korea* Article 43

2015

An Open Source Web Service For Registering And Managing Environmental Samples

Anusuriya Devaraju
CSIRO Mineral Resources

Jens Klump
CSIRO Mineral Resources

Pavel Golodoniuc
CSIRO Mineral Resources Flagship

AT5-01

AN OPEN SOURCE WEB SERVICE FOR REGISTERING AND MANAGING ENVIRONMENTAL SAMPLES

Anusuriya Devaraju, Jens Klump and Pavel Golodoniuc

CSIRO Mineral Resources Flagship,
PO Box 1130, Bentley,Western Australia, 6102.
Email: {anusuriya.devaraju, jens.klump, pavel.golodoniuc}@csiro.au

ABSTRACT

Records of environmental samples, such as minerals, soil, rocks, water, air and plants, are distributed across legacy databases, spreadsheets or other proprietary data systems. Sharing and integration of the sample records across the Web requires globally unique identifiers. These identifiers are essential in order to locate samples unambiguously and to manage their associated metadata and data systematically. The International Geo Sample Number (IGSN) is a persistent, globally unique label for identifying environmental samples. IGSN can be resolved to a digital representation of the sample through the Handle system. IGSN names are registered by end-users through allocating agents, which are the institutions acting on behalf of the IGSN registration agency. As an IGSN allocating agent, our goal is to implement a web service based on existing open source tools to streamline the processes of registering IGSNs and for managing and disseminating sample metadata. In this paper, we present our ongoing work onthe design and development of the web service, and its data schema and database model for capturing key aspects of environmental samples. We show how existing controlled vocabularies can be incorporated into the service development to support the metadata registration of different types of samples. The proposed sample registration and curating approach has been trialed in the context of the Capricorn Distal Footprints project on a range of different sample types, varying from water to hard rock samples. The initial results demonstrate the effectiveness of the service while maintaining the flexibility to adapt to various media types, which is critical in the context of a multi-disciplinary project.

1. INTRODUCTION AND MOTIVATION

In the earth science disciplines, physical samples are essential for understanding the complexity of our environment and its resources. For example, direct sampling techniques such as drilling and core sampling are employed to identify mineral and energy resources,insects and plants specimens are collected tounderstand their complex molecular interactions, and scientific ocean drilling programs provide insights into the seafloor and sub-seafloor microbial communities.While huge efforts have been put into collecting physical samples, these usually lie isolated; they may be kept by sample curators, laboratories, state agencies, or museums, and their metadata may existon proprietary data systems or on researchers' personal hard drives.Environmental samples are also diverse, as their collection, curation, and analysis vary from one domain to another. Each curator may follow their own way of documenting the samples. The descriptions may be incomplete, and no common service is available to discover the samples and their descriptions. It is also possible that the sample naming conventions used are only unique locally. According to Ramdeen (2015), "this can lead to complicated hybrid collections […] that, owing to their complicated curation schemes and lack of standardization may become lost". Inaccessible, irrecoverablesamples are unusable and have little value until they are released to scientists and public access(Geological Society of America (GSA), 2012)(Ramdeen, 2015). When sharingmetadata of samples in a globally distributed environment, the unique identification of

samples becomes essential. Persistent identifiers such as Digital Object Identifiers (DOI)[1] have proven successful in enabling users to access a digital resource via a persistent link on the Web. In a similar manner, assigning globally unique identifiers to physical samples will facilitate access to manysamples that are scattered across various repositories. Theseidentifiers can be used to locate samples unambiguously and to gather their associated metadata and data systematically. As a result, this new way of publishing samples and their metadata will create opportunities to re-use existing samples in terms of"new societal issues, environmental concerns, scientific interpretations, and analytical techniques"(Geological Society of America (GSA), 2012).

The International Geo Sample Number (IGSN) is a persistent and unique alphanumeric codefor identifying environmental samples and specimens. A sample's IGSN number can be resolved to a digital representation of the sample through the Handle[2] system. The IGSNsof physical samples are registered by end-users (e.g., individual researchers, data centers and projects)through allocating agents. Allocating agents arethe institutions acting on behalf of the main registration agency (IGSN e.V.)[3]. For further details on IGSN, see Section 2.

The IGSN is based on precursor work at Lamont-Doherty Earth Observatory (LDEO), which was developed as the System for Earth Sample Registration (SESAR)[4]. SESAR was developed with the requirements of individual investigators in geochemical research in mind. There are shortcomings in its technical implementation. Our work is expansion of precursor work in SESAR to suit Australian geo community needs. The Internet of Samples in the Earth Sciences (iSamples)[5] is an EarthCube Research Coordination Network programthat aims toconnect sample collections across the Earth sciences through cyberinfrastructures. One of the key questions addressed by the program is identifying metadata profiles and tools required to facilitate samples discovery and interoperability across domains. Our work complements this issue by developing a metadata schema,which can be used to represent various physical samples, and by implementing relevant tools, e.g., an allocating service and a metadata store.

This paper presents our ongoing work on developing an allocating agent web service and a metadata schema to streamline the processes of registering IGSNs and managing and disseminating sample metadata in CSIRO. The service supports a Representational State Transfer (REST) API for namespace governance, registration of IGSN and associated metadata. The schema defines descriptive metadata that are required for IGSN registration by the allocating agent. We will use the registration of water samples collected as part of the Capricorn Distal Footprints project[6] to demonstrate the solutions developed.The paper is organized as follows: Section 2 provides an overview of IGSN e.V., including organization and governance, syntax, and metadata. Section 3 presents a description of the service and the schema developed. This is followed by their implementation and results in Section 4. Finally, Section 5 concludes the paper with some directions for future work.

[1]http://www.doi.org/
[2]http://www.handle.net/
[3]http://www.igsn.org/
[4]http://www.geosamples.org/mysesar
[5]http://earthcube.org/group/isamples
[6]http://www.sief.org.au/FundingActivities/RP/Distal.html

2. IGSN OVERVIEW

Figure 1 shows the entities that form the IGSN architecture. The IGSN Registration Agency (IGSN e.V.) is theimplementing organization that operates the main registry and resolver services. At this level, the registry only imposes registration metadata (e.g.,the IGSN number and its landing page, the registrant, and the registration date). The Handle system is similarto Domain Name System (DNS), which resolves a URL to an Internet Protocol (IP).An allocating agent is the member institution that is authorized by the IGSN e.V. to register the IGSN in the permitted namespace. Examples of allocating agents are the Interdisciplinary Earth Data Alliance (IEDA), the Commonwealth Scientific and Industrial Research Organisation (CSIRO), and the German Research Centre for Geosciences (GFZ).[7]The difference between the top-level service and the allocating service is that the former only requires basic IGSN *registration metadata*[8], whereas the latter captures samples' *descriptive metadata*.Clients (e.g., individual data centers or sample curators) register samples together with metadatathrough an allocating agent. To enable this, an allocating agentshould providea service that registers identifiers from the top-level registry, keeps a repository of samples registered through the agent, and maintains a metadata portal to make samples searchable and accessible.This paper presents the development of an allocating service and itsdescriptive metadata schema.The metadata portal is not covered in this paper, but it is subject to future investigation. A client also hosts the landing pages of the registered samples. These pages may include more detailed metadata of samples (e.g., a description of the variables, the sample data provider,and links to actual datasets and related documentations).

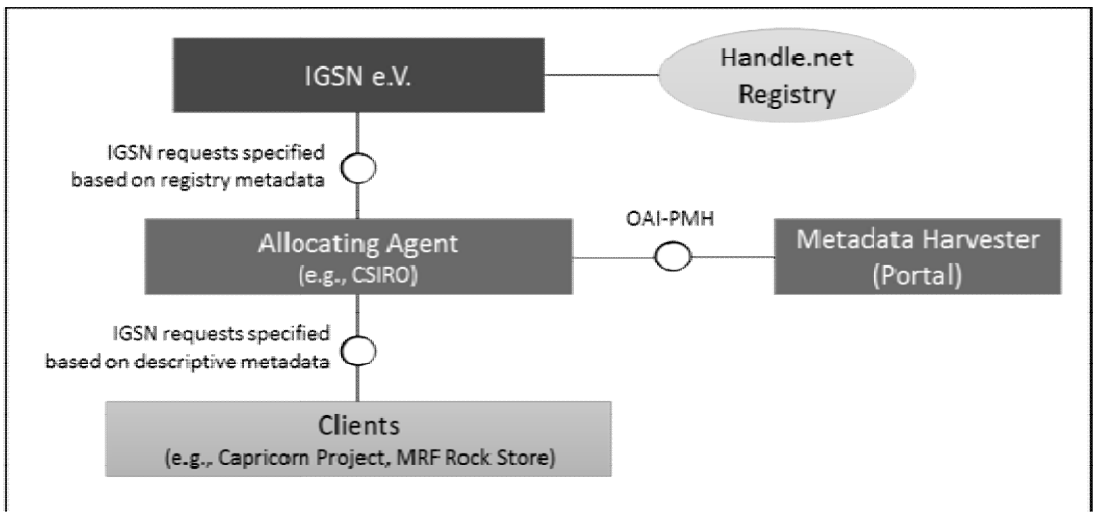

Figure1. Hierarchical architecture of the IGSN registration.

An IGSN must be unique and is case insensitive. It consists of <*namespace*><*code*> (see Table 1).The allocating agent should be certain that the element <*code*> is unique within its namespace. The recommended format of the IGSN is a nine-character string that comprises a namespace identifier and a sample number. However, it is also stated that the communities may adopt their own IGSN formats if necessary(IGSN e.V., 2015). In our service, we do not restrict the total length of an IGSN, as we see the possibilities of projects

[7]Existing allocating agents : http://dokuwiki.gfz-potsdam.de/datawiki/doku.php?id=igsn:allocating_agents
[8] For further details on the IGSN registry metadata schema, see http://trac.gfz-potsdam.de/igsn/wiki

registering samples with longer numbers.

Table 1. IGSN syntax with examples.

Element	Restriction	Descriptions and Examples
<IGSN>	*<namespace><code>*	An IGSN takes the form of an alphanumeric character divided into two elements: *<namespace><code>*. An example of IGSN is CSCAP00001.
<namespace>	UPPER (A-Z)	A namespace refers to the prefix of an allocating agent (e.g., CS stands for CSIRO).
<code>	Any combination of (A-Z), (a-z), (0-9), hyphen-minus (-), and dot (.)	A code is a combination of the sub-namespace and the sample number specified by the client. For example, 'CAP' stands for Capricorn Distal Footprints project,and '00001' is the local sample number assigned by the client.

3. DEVELOPING AN ALLOCATING SERVICE FOR CSIRO

This section describes the descriptive metadata model and summarizes operations supported by the allocating service.

3.1 Descriptive Metadata

One important concern for building the allocating service is thedescriptive metadata model (schema) and its database model. The schemadefines the essential characteristics of samples. The allocating service leverages the schema to validate and handle sample information sent by a client. The contributions made in terms of the schema development are summarized as follows:

a. Since the potential projects that will use the allocating service involvevarious types of samples, it is important to ensure that the schema can be adapted to respond to the samples' diversity and requirements.Therefore, we have used the records of different types of existing samples and inputs from the domain experts (e.g., sample collectors and curators) as a basis to identify the core elements representing a sample. Figure 2 illustrates a partial view of the descriptive metadata schema developed.Metadata of multiple samples can be defined based on the descriptive schema. The elements in the schema are generic enough to be reused for various samples. They can be grouped into identification, collection, curation, and related resources (for examples, see Table 2).
b. Some of the elements in the metadata schema are specified as mandatory (e.g., *sampleNumber, sampleName, isPublic, landingPage, sampleType* and *sampleCuration*).A client must supply these elements during the sample registration process. They are alsousefulin finding and retrieving physical samples via a metadata harvester (see Figure 1). The curation information is also required, as it indicates the storage and access to the physical samples.
c. A predefined list of values for the element*sampleTypes* and the attribute *featureType*has been specified, based on the ODM2 controlled vocabularies[9], to ensure the consistency and correctness of theirvalues. Here, the principle of Linked Data has been applied; as such,the predefined terms are identified by corresponding URIs. Therefore, more meaningful information about the terms can be determined.
d. We have re-used elements (e.g.,*sampleNumber* and *relatedResourceIdentifier*) and data

[9]http://vocabulary.odm2.org/

types (e.g., *dateType* and *relationType*) from the IGSN registry schema in our descriptive schema. We have extended*dateType*to describe both time instant and interval.

e. The schema also provides options to representdifferent types of location. For example, the location where a sample is collected can be specified as absolute location, bounding box, locality and relative location.

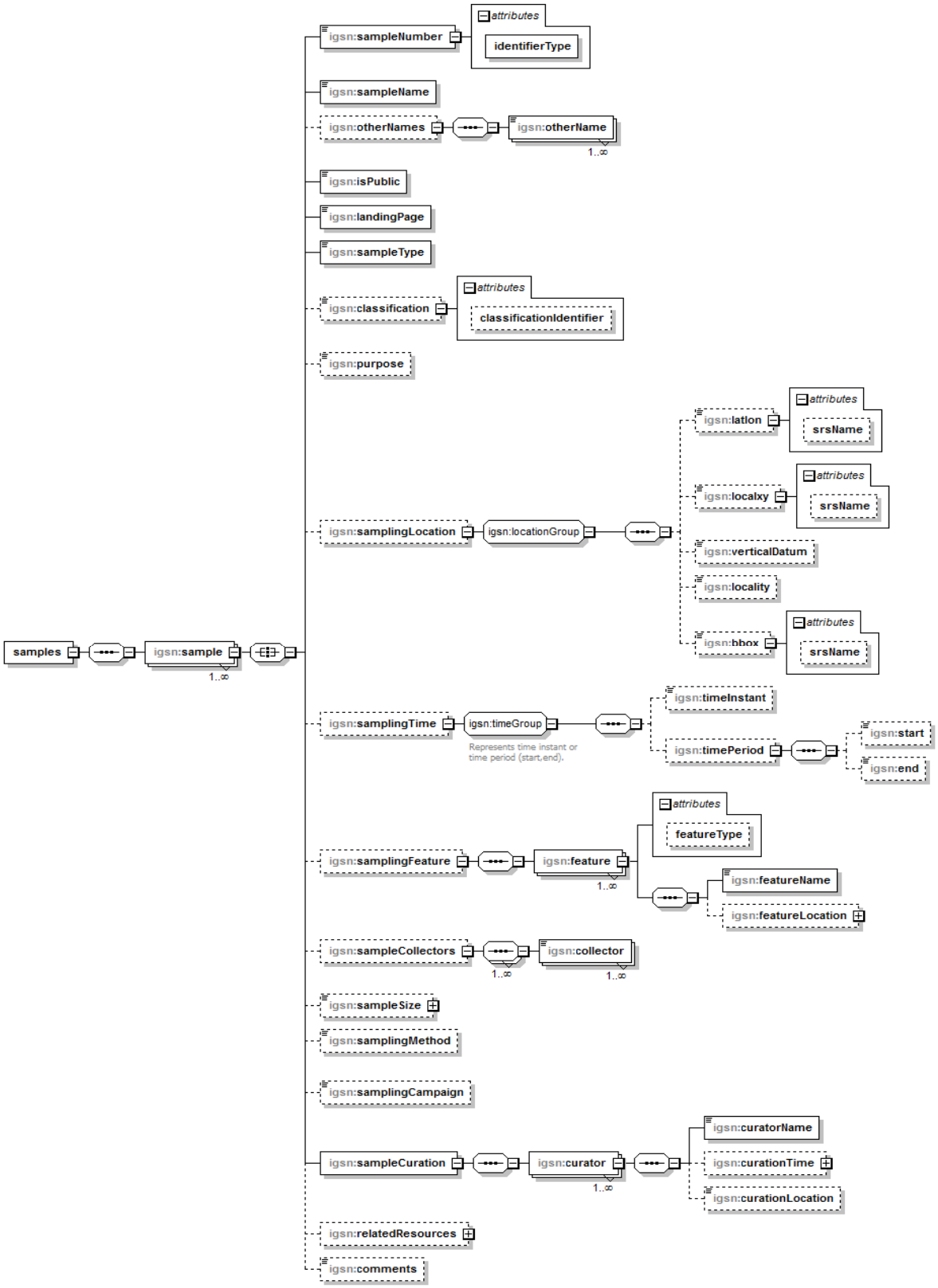

Figure 2. A partial view of the descriptive metadata schema.

3.2 Allocating Service

The allocating service enables a client to register sub-namespaces and multiple samples and retrieve the metadata of a particular sample programmatically. The service implementsa

REST API(Fielding, 2000). The service is accessible via the end point:http://{server-name}/igsn-service/. Table 3 summarizes three resources (*/subnamepsace, /igsn,*and */metadata*) supported by the service and their request and response information. Note that all requests are authenticated by passing in HTTP Basic Authentication headers.Follwing the principles shown in Figure 3, a client program registers IGSNs for a collection of samples via an allocating-agent. A client program will be only allowed to register IGSNs with allocated sub-namespaces. The sample descriptions (response body) sent by the client must conform to the descriptive metadata schema (Section 3.1). The allocating service will update a sample description instead of inserting a new description if the client specifies an existing IGSN.

4. IMPLEMENTATION AND RESULTS

This section demonstrates the results of the initial investigation and early adoption of the proposed metadata schema and IGSN allocation service for registration of samples within the context of the Capricorn Distal Footprints project.

4.1 Capricorn Distal Footprints

The Capricorn Distal Footprints project is a significant collaboration between industry, the Geological Survey of Western Australia, academia and CSIRO to address the issue of exploration through cover, which poses the biggest challenge for the Australian minerals industry, 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 project aims at understanding the metallogenic evolution of the orogen and provides exploration models to aid future discoveries. Regional geophysical interpretation is analyzed to establish links to the hydrogeochemical, regolith, and resistate mineral studies around known deposits to gain insight into the key indicators of mineralisation, not just within a few hundred meters of the deposits but much further afield. The regional study of such a scale requires analytical data of a large number of samples that include water, plant, soil, and rock samples. Apart from new analysis, historical datasets are obtained from other institutions and government agencies, which is often duplicated or hard to identify. All these difficulties led us to establish a sample curation system that would provide a robust framework of both technological mechanism and governance policies.

IGSN system offers a viable and broadly accepted solution for geo-sample identification in various science disciplines. In the initial stage of the project, the proposed approach was adopted for identification of water samples using IGSN. The proposed solution not only offers the persistent identifiers for samples but also offers a robust yet flexible mechanism to associate rich self-descriptive metadata with the sample. The Capricorn Distal Footprints project offers an ideal environment for application of IGSN system. The diversity of sample types requires adoption of multiple metadata profiles to provide ability to capture sample type specific intrinsic properties. The IGSN registration service is currently designed to handle generic metadata schema and the problem of sample type specific properties is yet to be solved.

Table 2. Examples of elements of the descriptive metadata. The rightwards arrow (→) indicates a child element.

	Examples (elements/attributes)	Obligation	Max Occurrence	Data Type
Identification	*sampleNumber*	M	1	CharacterString[a]
	landingPage	M	1	URI
	isPublic	M	1	Boolean
	sampleType	M	1	CV[b]
Collection	samplingFeature → feature : *featureType*(attribute)	O	1	CV[c]
	samplingMethod	O	1	CharacterString
	sampleCollectors → *collector*	O	N	CharacterString
Curation	sampleCuration →curator→ *curatorName*	M	1	CharacterString
Other related information	relatedResources →*relatedResourceIdentifier*	O	N	CharacterString
	comments	O	1	CharacterString

[a]The following XSD restriction applies to a sample number: <xs:pattern value="([A-Z]{2}[A-Za-z0-9\-.]*)"/>
[b] Controlled vocabulary of sample types: http://vocabulary.odm2.org/medium/
[c] Controlled vocabulary of feature types: http://vocabulary.odm2.org/samplingfeaturetype/

Table 3. A summary of the allocating service API.

Request URI	Descriptions	Method	Request Headers	Request Body	Response Headers	Response Body
/igsn-service/**subnamespace**/	This request will register a sub-namespace if the specified sub-namespace does not exist.	POST	Content-Type: text/plain	subnamespace={subnamespace}	-	HTTP status code (201 or non-201 with a short explanation of the status code)
/igsn-service/**subnamespace**/all	This will retrieve all registered sub-namespaces.	GET	Accept: application/json	-	Content-Type: application/json	HTTP status code (200 with JSON representing registered namespaces or non-200 status with explanation)
/igsn-service/**igsn**/	This request will register IGSNs for samples. It will update the sample description instead of inserting a new record if the IGSN exists.	POST	Content-Type: application/xml	UTF-8 encoded descriptive metadata	Content-Type: application/json	HTTP status code (201 with a list of successful and unsuccessful samples or non-201 with a short explanation of the status code)
/igsn-service/**metadata**/{igsn}	This request returns the descriptive metadata associated with a given {igsn}.	GET	Accept: application/xml	-	Content-Type: application/xml	HTTP status code (200 with XML representing a dataset or non-200 status code with explanation)

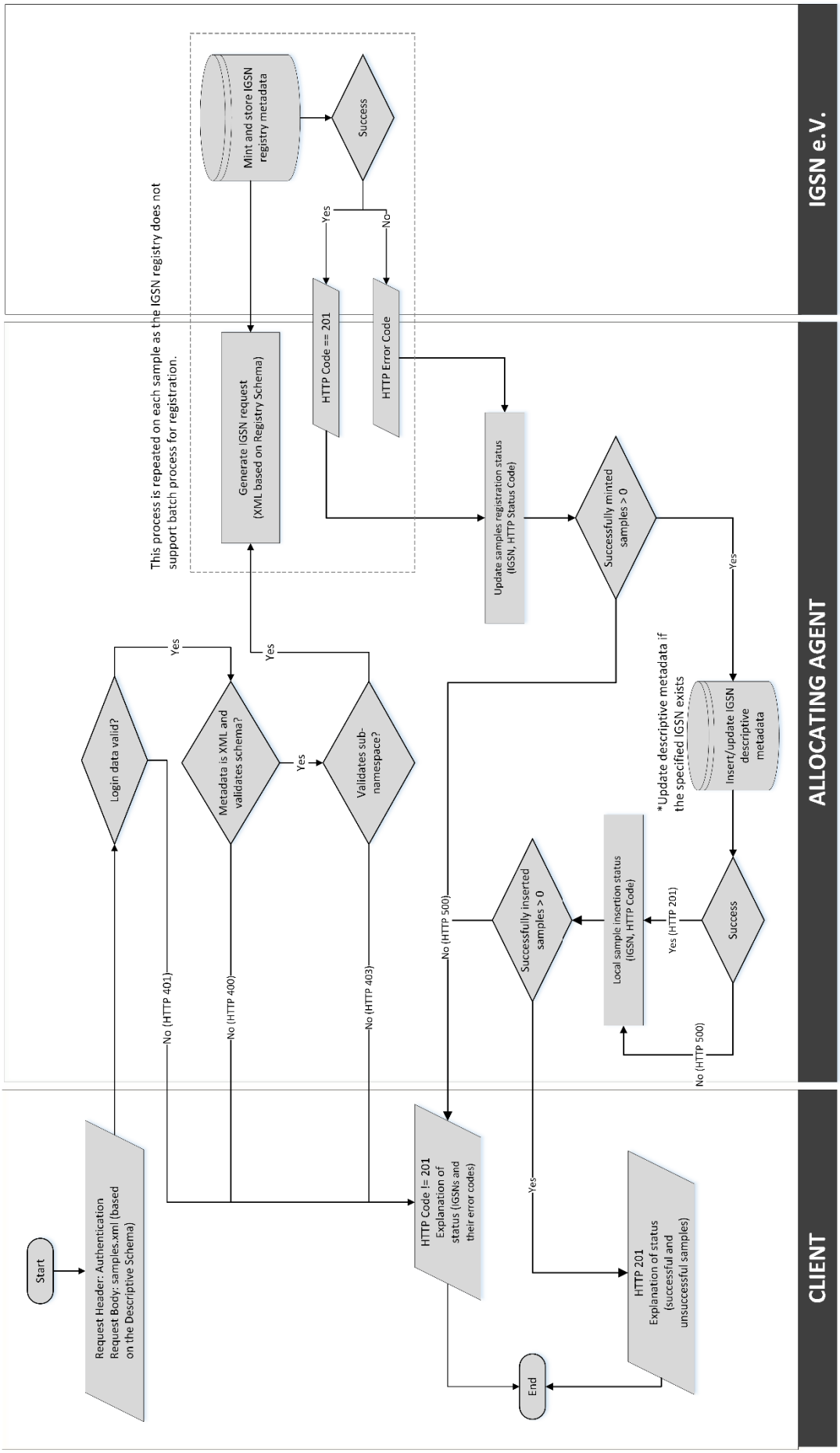

Figure 3. IGSN registration by a client program.

4.2 System Architecture and Results

Figure 4 illustrates the system architecture of the service developed. A relational database model was developed usinga MySQL database to store descriptions of samples and data centers. The allocating service is hostedon the Tomcat application server. All registration requests from the allocating service to the IGSN registry are sent via HTTPS protocol. The project curation system makes requests based on the API on Table 3. A client can also use the service in test mode by settingan optional query parameter, e.g., */igsn?testMode=true.* This means that the service will not register IGSNs from the top-level registry nor modify the descriptive metadata store. Figure 5 shows an example of descriptive metadata sent by the client program to register IGSNs for water samples from the Capricorn Distal Footprints project.TheIGSN registry shows that these samples have been successfully registered.An example of the sample IGSN is "10273/TEST/CSCAP0003", and its actionable link is: http://hdl.handle.net/10273/TEST/CSCAP0003.Note that the samples have been registered with the special test prefix "10273/TEST" via the allocating service. This prefix is reserved by the IGSN registry for testing purposes. We will register actionable IGSNs once the allocating service is publiclyavailable and the Capricorn Footprints project sample curation system is completed.

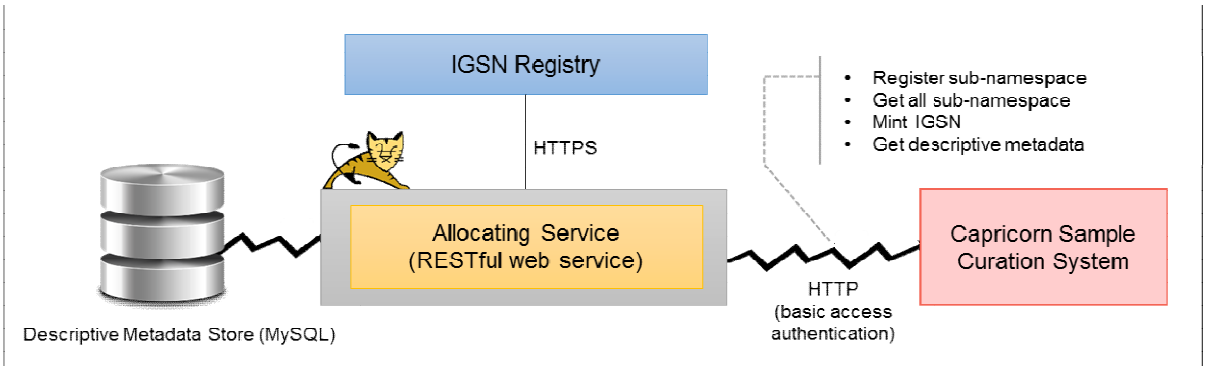

Figure 4. System architecture.

5. CONCLUSIONS

This paper describes the development of a web service for the registration of geological samples by CSIRO at an organizational scale. The solutions developed areuseful to connecting physical samples across the Earth sciences to the Web in a systematic manner. The proposed service and the schema were trialed within the scope of the Capricorn Distal Footprints project.Samples collected during the field sampling campaigns were registered through the CSIRO IGSN allocating service and therefore uniquely identified with global IGSN numbers. The descriptive schema has been developedbased on existing samples and insights from the domain experts. Although the results shown are based on water samples, the schema is extensible and can be applied to various sample types, which is essential in the context of a multi-disciplinary project.The descriptive schema is available through the GitHub repository[10]. While the basic requirements of the service have been achieved, there arestill areas for further development before making the service publically available:

a. The service will be tested with data from other sample stores in the Mineral Resources Flagship (MRF) of CSIRO, e.g. the sample repository of the Australian Resources

[10] https://github.com/kitchenprinzessin3880/csiro-igsn-schema.git

Research Centre (ARRC) and the National Collection for Mineral Reflectance Spectra. The IGSN has also been incorporated in laboratory information management systems at ARRC.

b. The allocating service will be extended to support updates and deletions of sample registrations. Following the IGSN top registry, the delete operation should only marka registered sample as “inactive”. A mechanism to activate the sample again will be included. A retrieval of IGSNs and their associated metadata based on a user defined polygon will also be supported.
c. The mapping between the descriptive metadata elements and the concepts specified in existing data standards (e.g., the Observations and Measurements—Sampling Features, ISO 19115 and GeoSciML) should also be documented to ensure the correct application of the schema across different science domains.
d. Currently, the descriptive schema incorporates the ODM2 controlled vocabularies to represent sample and feature types. The schema should be extended to vocabularies describing units, spatial reference system, vertical datum, and classification.
e. We willdevelop a web portal which harvests metadata from repositories of different allocating agents in Australia via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

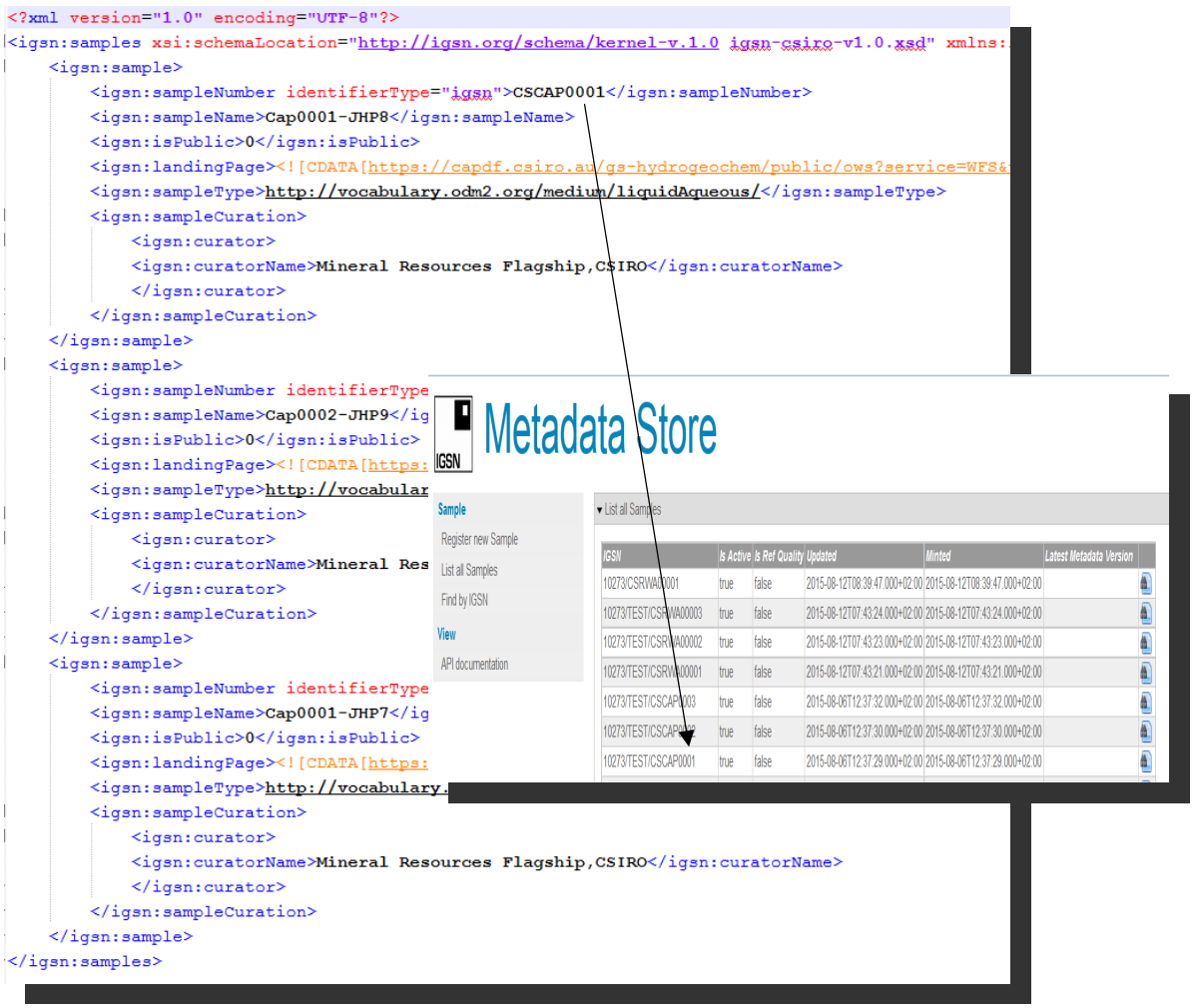

Figure 5. Sample registration request and results.

6. REFERENCES

Fielding, R. T., 2000. Chapter 5: Representational State Transfer (REST), *Architectural Styles and the Design of Network-based Software Architectures*. University of California, Irvine.

Geological Society of America (GSA), 2012. Position Statement - Geoscience Data Preservation. The Geological Society of America. Retrieved from http://www.geosociety.org/positions/position9.htm

Pearce, M. A., Hough, R., Ley, Y., Spinks, S. C., Thorne, R., White, A. J. R., Golodoniuc, P., Gray, D., Munday, T., 2015. The Capricorn Distal Footprints Project: An Orogen-scale approach to mineral systems. Proceedings of the GAC-MAC-CGU-AGU Meeting.

Ramdeen, S., 2015. Preservation challenges for geological data at state geological surveys. *GeoResJ*, 6, 213 – 220. http://doi.org/http://dx.doi.org/10.1016/j.grj.2015.04.002