A Digital Repository for Physical Samples: Concepts, Solutions and Management

Anusuriya Devaraju[1(✉)], Jens Klump[1], Victor Tey[1], Ryan Fraser[1], Simon Cox[2], and Lesley Wyborn[3]

[1] CSIRO Mineral Resources, P.O. Box 1130, Bentley, WA 6102, Australia
{anusuriya.devaraju,jens.klump,victor.tey,ryan.fraser}@csiro.au
[2] CSIRO Land and Water, Private Bag 10, Clayton South, VIC 3169, Australia
simon.cox@csiro.au
[3] The Australian National University, 56 Mills Road, Acton 2600, Australia
lesley.wyborn@anu.edu.au

Abstract. Physical samples are important resources for sample-based data reuse. They may be utilized in the reproduction of scientific findings, depending on their availability and accessibility. Although several solutions have been developed to curate and publish digital collections (e.g., publications and datasets), considerably less attention has been paid to providing access to physical samples, and linking them to data, reports, and other resources on the Internet. Some progress has been made to bring physical samples into the digital world; for example, through the web-identifier schemes, sample metadata standards and catalogues, and specimen digitization. Existing studies based on the above examples are either project or domain-specific. Also, a particular challenge exists in providing citable and resolvable identifiers for physical samples outside the context of an individual project or a sample data repository. Within the Commonwealth Scientific and Industrial Research Organisation (CSIRO), further work is needed in order to connect the various types of physical samples collected by different entities (individual researchers, projects and laboratories) to the Web, and enable their discovery. We address this need through the development a digital repository of physical samples. This paper presents technical and non-technical components of the repository. They were applied to unambiguously identify the various physical samples and to systematically provide continuous online access to their metadata and data.

Keywords: Physical sample · Specimen · Persistent identifier · IGSN · Sample data curation · Institutional repository

1 Introduction

Physical samples (also called physical specimens) are information sources that come from the Earth's environment. Sampling activities are conducted for scientific research and monitoring purposes. For example, core samples are collected

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 74–85, 2017.
DOI: 10.1007/978-3-319-67008-9_7

to investigate the physical and chemical nature of rocks, soil specimens are gathered to calibrate soil-water measuring instruments, and specimens of plants and insects serve as reference materials to understand the biodiversity of a specific area. The reproducibility of scientific findings and the interpretation of sample-based data requires access to the physical samples and sample metadata, respectively [13]. In addressing the importance of samples, organizations have issued policies and regulations. For example, the NSF Data Sharing Policy asserts that [1], "investigators are expected to share with other researchers, [...] the primary *data, samples, physical collections* and other supporting materials created or gathered in the course of work under NSF grants" [16, p. 66]. The Australian Antarctic program (AAp) Data Policy (2015) indicates that chief investigators "are responsible for ensuring that all *data and samples* generated as part of their research are adequately managed for long-term re-use" [15].

Motivation. Physical samples are important research assets of CSIRO. Within the organization, the community of sample users includes individual researchers, projects, and laboratories, all of which collect or generate samples as a part of their field studies or analytical processes. Examples of samples include groundwater, drill cores, seabed cores, soil archives, sediment, biological specimens, and synthetic materials. The organization curates a large number of legacy samples as well as thousands of samples that will be collected in the future. As in most organizations, there are two main challenges to identifying and discovering samples [6,12]. First, sample collectors may follow their own naming convention to identify samples; therefore, sample names can often be ambiguous. For instance, different collectors may refer to different samples using the same name, or the same sample could be named differently based on analytical procedures performed. This naming ambiguity is also applicable to sample-related physical resources, such as sample collections[1] and sampling features[2]. Herein, we will refer to physical samples, sample collections, and sampling features as 'physical resources' in this paper. The second challenge concerns the discovery of samples. Sample descriptions are often only available to the sample owner. They may not be easily discoverable by other users due to the lack of online catalogues that offer access to the sample metadata.

Contributions. In this paper, we address the challenges through the development of a digital repository, which supports the effective management and discovery of physical resources and their metadata across the Web. The key features of the repository are (a) globally unique and persistent identification of the resources, (b) technical solutions (e.g., tools, data stores, web services and a web portal) that are domain independent, extensible, and easily accessible by members of the organization, and (c) interoperability with sample data repositories managed by other institutions. Given the diverse research communities in

[1] A collection may be a group of arbitrary specimens or an aggregation of specimens, e.g., rock chips.

[2] A sampling feature is an entity that is designed to observe some domain features. This may refer to the 'locations' where a sample was collected from such as drill-holes, wells, sections, and soil pits.

CSIRO, and the changes in staff over the years, exclusively technological solutions are inadequate. Therefore, another important aspect we considered when developing the digital repository is the non-technical solutions, such as an organizational and governance framework that supports the operation of the technical solutions and the governance of the physical resource identification used in the organization.

Outline. This paper provides an overall view of the digital repository, its related work Sect. 2, and its technical Sect. 3 and non-technical solutions Sect. 4. In addition, we describe the applications of the digital repository in the context of different sample repositories in the organization. We summarize the paper Sect. 5 by detailing its contributions.

2 Related Work

This section provides an overview of IGSN and summarizes related work.

2.1 International Geo Sample Number (IGSN)

Persistent identifiers, such as Digital Object Identifiers (DOI), have proven successful in providing long-term access to digital resources by maintaining the link between a digital resource and its location on the Web [8]. Similarly, assigning globally unique identifiers to physical samples will facilitate unambiguous and systematic access to the samples [6]. IGSN[3] is a persistent, globally unique code for the identification of physical samples and sample collections. The use of the IGSN is not limited to the geosciences but is also relevant to other sciences dealing with specimens, such as biology and oceanography. The IGSN initiative is represented by organizations in various parts of the world, including North America, Europe, Asia, Africa and Oceania.

Figure 1 illustrates the system architecture of the IGSN registration. The Implementation Organization of the IGSN (IGSN e.V.) governs and promotes standard methods for identifying and citing physical samples, and operates the international (top-level) IGSN registration service [10]. The international registration service is modelled after DataCite and utilizes the Handle.net System[4], which is a global persistent identifier resolver service [7]. An allocating agent is a member institution that is authorized by the IGSN e.V. to register the IGSN within an allocated namespace. CSIRO is one of three IGSN allocating agents in Australia alongside Geoscience Australia and Curtin University. In the CSIRO implementation, a client (i.e., individual users or laboratories) may send IGSN registrations to the agent's service based on the *description schema* developed by the respective allocating agent. Then, the agent service forwards the registrations to the international registration service based on the*registration schema*[5].

[3] http://www.igsn.org/.
[4] https://www.handle.net/.
[5] http://schema.igsn.org/registration/.

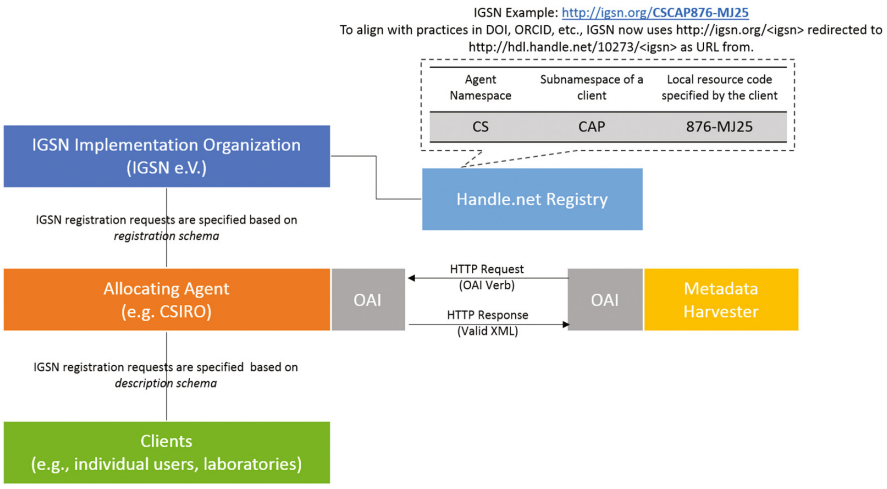

Fig. 1. The hierarchical architecture of the IGSN registration. A namespace refers to the prefix of an allocating agent, e.g., the IGSN e.V. allocated the prefix *CS* to CSIRO. A subnamespace uniquely represents a client. For example, *CAP* is the subnamespace of the Capricorn Distal Footprints project.

The registration schema only covers registration information (e.g., sample number, registrant and log), and excludes sample descriptions to allow greater flexibility in describing samples for different use cases. This separation of registration and description of objects differs from the practice in DOI registration where the registration agents send a standardized set of metadata to the DOI registry as part of the registration process. The Handle.net resolves each individual IGSN handle (e.g., 10273/CSCAP876-MJ25) to a landing page[6] for the resource identified by the handle. Landing pages include more detailed (domain-specific) information of the registered resources and are maintained by the respective client.

2.2 Related Work

Sample Registration Systems. Several organizations have introduced IGSNs to publish their physical samples information. Among these are the System for Earth Sample Registration (SESAR)[11], the Integrated Ocean Drilling Program (IODP) [2] and the International Continental Scientific Drilling Program [3]. The IGSN was developed as SESAR[7] in precursor work at Lamont-Doherty Earth Observatory (LDEO). SESAR was developed with the requirements of individual investigators' geochemical research in mind and have several technical limitations [6]. This work is an expansion of precursor work in SESAR. The existing systems were developed for fairly specific use cases in single research domains.

[6] http://capdf.csiro.au/igsn/CSCAP876-MJ25.
[7] http://www.geosamples.org/mysesar.

In contrast, the solutions we developed in CSIRO are domain-independent, i.e., they support representation and registration of various specimen types. Following the IGSN recommended practice, we facilitate the specimen discovery through the meta-data harvesting capabilities across the IGSN communities in Australia.

Specimen Metadata Information Model. There are several metadata schemas representing physical samples. However, some of them are domain specific (e.g., Darwin Core (DwC) [17]), while others have specific design considerations (e.g., modelling sampling features and observation procedures). We developed a comparison between the description metadata schema and the existing schemas in [6]. In this implementation, the sample description metadata schema supports the registration of physical resources through the CSIRO allocating agent service, and the dissemination of resource records through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[9] implementation[8]. The description schema adapts some concepts from the DataCite Metadata Schema (v4.0) [5]. It is closely aligned with ISO 19156:2011 (Geographic information - Observations and Measurements (O&M)) [4]. The DataCite Metadata Schema includes the core metadata elements for identifying and describing digital resources, whereas ISO 19156:2011 defines a common set of sampling feature types including *SpatialSamplingFeatures* and *Specimen*. Although the description schema shares some similarities with the two schemas, it differs from them in several aspects. First, it represents the common properties of the three physical resources – physical sample, sample collection and sampling features. We represent new metadata elements, and modify existing elements from the DataCite schema (e.g., cardinality and restrictions) to meet the requirements of the resources. For particular metadata elements (e.g., material, specimen and identifier types), we develop and set controlled vocabularies (expressed as Simple Knowledge Organization System (SKOS)[9] concepts) as their acceptable values. For more information about the description schema and its contributions, see Sect. 3.

Metadata Harvesting and Dissemination. There are several ways to harvest sample metadata catalogues, common examples are the OGC Catalogue Services for the Web (CSW)[10] and OAI-PMH[11]. In the OAI-PMH framework, a *service provider* deploys a client application (harvester) that requests metadata from one or more *data providers*. A data provider operates the metadata catalogue of a repository, which serves the OAI-PMH requests (see Fig. 1). We developed an OAI-PMH provider service to disseminate the sample metadata records in our digital repository in two metadata profiles such as Dublin Core, CSIRO-IGSN Description Schema. We also developed an OAI-PMH service provider which harvests metadata records from our metadata store through the data provider service, and from other allocating agents, e.g., Geoscience Australia.

[8] https://igsn.csiro.au/igsn30/api/service/30/oai.
[9] https://www.w3.org/2004/02/skos/.
[10] http://www.opengeospatial.org/standards/cat.
[11] https://www.openarchives.org/pmh/.

3 Solutions

Figure 2 illustrates the architecture of the digital repository that supports sample registration and discovery in CSIRO. Its components are listed as follows:

a. **Clients:** The allocating agent registration service handles requests from two types of clients – sample data curation systems and individual researchers. Individual researchers may register their samples with IGSNs via a web form[12], while sample data curation systems request IGSNs programmatically. Current sample data curation systems in CSIRO are the Capricorn Distal Footprints project, Repository of the Australian Resources Research Centre (ARRC), and Reflectance Spectra Reference Libraries. Table 1 summarizes the local sample systems, material types, and IGSNs registered.
b. **CSIRO allocating agent service:** The allocating agent registration service[13] is a Representational State Transfer (RESTful) web service endpoint that enables clients to register IGSNs of physical resources, to request subnamespaces, and to retrieve resource metadata programmatically. The IGSN registration requests sent by the clients must be encoded in XML conforming to the CSIRO IGSN description metadata model.[14] The agent registration service mints IGSNs from the international registration service on behalf of the clients.
c. **CSIRO-IGSN Description metadata model:** The description metadata model represents the common concepts associated with physical resources such as identification, collection, curation, and related resources. It is designed to be general enough to catalogue different specimen types in the organization. The metadata schema serves as the basis for IGSN registration through the CSIRO agent registration service and to disseminate resource metadata through the OAI-PMH data provider service. Key features of the schema are that it supports batch registration of resources as our use cases may involve large batches of IGSN registrations, and it has minimal restrictions on which elements are required, e.g., resource identification, types and curation details. Some of the metadata elements are required to obtain IGSNs from the international registration service (e.g., *resourceIdentifier* and *landingPage*), while the others are relevant when discovering the resources through the web portal (e.g., *materiaTypes* and *curationDetails*). In addition, the schema offers flexibility to express both geographic and non-geographic location information (toponym), and time instants and intervals based on the W3C Date and Time Formats[15], which is a simpler profile of ISO 8601[16]. Physical samples are often relocated from one repository to another, therefore the schema captures the provenance of sample curation. It also represents several relation types

[12] https://igsn.csiro.au/igsn30/.
[13] https://igsn.csiro.au/igsn30/api/.
[14] The XML schema and its graphical representation are available at https://igsn.csiro.au/schemas/3.0/.
[15] https://www.w3.org/TR/NOTE-datetime.
[16] https://www.iso.org/standard/40874.html.

to associate a registered resource with its related resources, such as subsamples, digital resources (datasets, reports, images) and a reference resource[17]. It leverages existing and new controlled vocabularies that we developed in order to provide standardized information about the metadata elements and to ensure consistent metadata entry by clients. Digitization of specimens is beyond the scope of the project, although the digital images of specimens could be linked to their specimens registered in our system through the description metadata model.

d. **Controlled vocabularies.** To align with existing standards, we incorporated existing vocabularies into the description metadata schema, e.g., OGC definitions of nil reasons[18], material and specimen types defined by the CUAHSI's Observations Data Model (ODM2)[19], the contributor types from the CSIRO Linked Data Registry[20] and EPSG Geodetic Parameter Dataset[21]. We also developed the missing SKOS-based vocabularies that are necessary to connect the registered resources to the Web of data, e.g., registration types, identifier types, and relation types. The new vocabularies were identified with their corresponding persistent URIs to ensure machine actionability to the vocabularies. We use the Research Vocabularies Australia (RVA)[22] system to maintain the new vocabularies.
e. **Metadata store:** Metadata are stored in a PostgreSQL database modelled after the description metadata model. The metadata store captures resource metadata and client information (e.g., subnamespaces).
f. **Metadata provider and harvester:** We implemented an OAI-PMH provider service[23] to disseminate the metadata of registered resources in the metadata store. We also developed an OAI-PMH harvester, which is based on the PANGAEA Framework for Metadata Portals (panFMP) [14]. It harvests sample metadata from our own repository and other allocating agents.
g. **National IGSN web portal:** panFMP is entirely web-service based and does not supply its own graphical user interface, therefore its index is queried through a web portal. The web portal[24] provides a common access to sample metadata harvested from OAI-PMH services operated by different allocating agents in Australia.

[17] A physical sample is usually compared with a reference sample.
[18] http://www.opengis.net/def/nil/OGC/0/.
[19] http://vocabulary.odm2.org/.
[20] http://registry.it.csiro.au/.
[21] https://epsg.io.
[22] http://www.ands.org.au/online-services/research-vocabularies-australia.
[23] https://igsn.csiro.au/igsn30/api/service/30/oai.
[24] https://igsn2.csiro.au/portal.

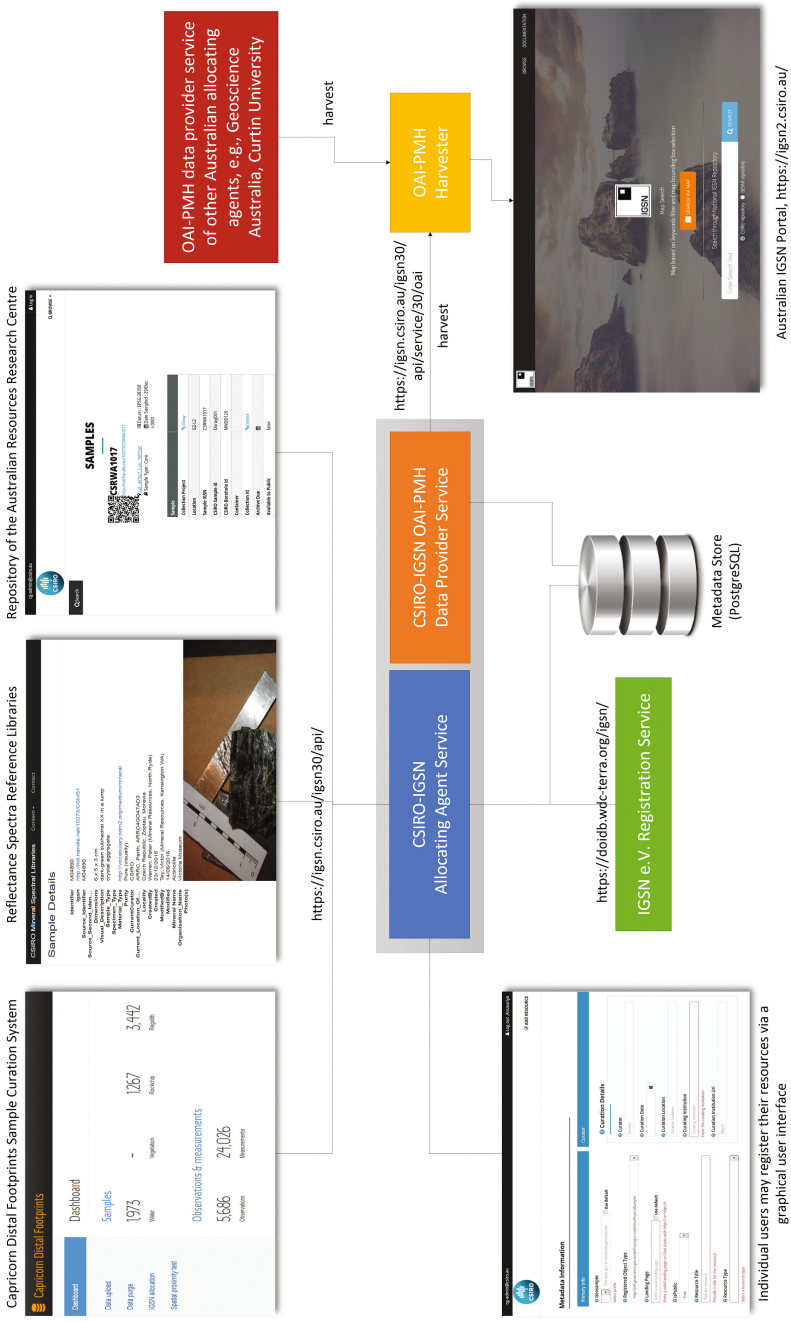

Fig. 2. Architecture of the CSIRO IGSN implementation.

Table 1. Local sample repositories and their IGSN registrations (as on 30.03.2017)

Repositories	Material types	IGSN Registered
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

4 Management

Use cases for samples in CSIRO range from individual researchers managing samples and their data manually (e.g., using spreadsheets) to projects and laboratories using sample curation systems. It is important to become familiar with how the users handle samples so that IGSN can be seamlessly integrated into their workflows and the existing workflows can be improved through technical solutions. To facilitate the integration of IGSN into our workflows, the following non-technical aspects were considered in this project.

a. **What can be identified with IGSNs?** To accommodate the needs of existing and potential sample applications, we allow IGSNs to be used to identify not only physical samples but also sample collections and sampling features. The IGSN Technical Documentation[25] makes recommendations concerning the format (identifier length) and semantic content of IGSN. In our implementation, we do not restrict the total length of an IGSN to allow repositories incorporate their existing identifiers into a globally unique IGSN. The IGSN identifiers are formed from a combination of the prefix of the allocating agent, client and the local sample identifier specified by the client.
b. **Identifier governance:** We established rules for assigning subnamespaces to different groups, collections and laboratories. The use of subnamespaces allows us to decouple the allocation of specific IGSNs in different parts of CSIRO, making it easier to ensure the global uniqueness of registered identifiers through a hierarchical delegation pattern [1].
c. **Integration of identifiers into new and existing systems:** For new sampling campaigns, we recommend that IGSN to be adopted at an early stage of the activity to ensure the consistent use of the identifiers throughout the sample life cycle, i.e., from collection and processing to curation. For existing sample curation systems, the local sample identifiers were extended using the IGSN namespace as a prefix to the local identifiers, thus making them globally unique. A similar method applies when individual researchers register their samples with IGSNs through the web form (Fig. 2). It is also possible to request the system to automatically generate unique identifiers for their samples. The web system also hosts the landing pages of registered resources, thus reducing the technical burden for users.

[25] http://igsn.github.io/syntax.

d. **Linking physical resources with their datasets:** For projects that have their own sample data curation systems, we recommend them to use the description metadata schema to associate a registered sample with the persistent URI of its datasets. For users who do not have a data curation system, we recommend to publish the specimen datasets via the CSIRO Data Access Portal, and then use the DOI generated by the data portal to link the dataset to its corresponding sample.
e. **Outreach:** We organized several events (presentations, meetings, and workshops) to introduce IGSN and to identify its potential application in the organization. The technical solutions are documented along with their source code, with examples available on a public repository[26]. This is important to reach the wider community, who may later adopt the solutions.
f. **National collaboration:** CSIRO has a joint project with the other allocating agents (Geoscience Australia and Curtin University) and is funded by the Australian Research Data Services program to implement IGSN for the Australian geoscience community. This collaboration effort involves representatives from academia, research and government agencies, and is essential to coordinating both IGSN-related activities and tool development at the national level, and to promote its implementation and governance in other sciences.

5 Conclusions

This paper described a successful implementation of persistent identifiers for physical resources (physical samples, sample collections, and sampling features) in a large organization. We developed a digital repository for the physical resources and specified its technical and non-technical components underlying the repository. The solutions developed have been applied to unambiguously identify physical resources from various studies, and to connect their metadata and data systematically to the Web. This improves the discovery of resources, and consequently facilitates their reuse and reproducibility.

The digital repository handles IGSN registrations from local sample data curation systems as well as from individual researchers. The hierarchical namespace delegation pattern is well suited for a large organization in which individual users, projects, and laboratories may all have different requirements for identifying and publishing their physical resources. The description metadata model is generic and extensible, and therefore suitable for representing the common properties of resources from different use cases. The digital repository harvests sample metadata from different sources, which can be aggregated to create new applications, for example, the Australian IGSN portal. Following the successful IGSN implementation in CSIRO, we are now collaborating with the John De Laeter Centre for Isotope Research at Curtin University to apply components developed in the context of their Digital Mineral Library. We reached out to some of the potential collections that could benefit from the CSIRO IGSN

[26] https://github.com/AuScope.

implementation, for example, the Australian National Soil Archive and the Australian National Insect Collection.

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

1. Bechtold, S.: Governance in namespaces. Loyola Los Angeles Law Rev. **36**(3), 1239–1320 (2003). doi:10.2139/ssrn.413681
2. Behnken, A., Wallrabe-Adams, H.J., Röhl, U., Krysiak, F.: Application of the IGSN for improved data - sample - drill core linkage. In: EGU General Assembly Conference Abstracts. EGU General Assembly Conference Abstracts, vol. 18, p. 16688, April 2016
3. Conze, R., Lorenz, H., Ulbricht, D., Elger, K., Gorgas, T.: Utilizing the international geo sample number concept in continental scientific drilling during ICDP expedition COSC-1. Data Sci. **16**, 2 (2017). doi:10.5334/dsj-2017-002
4. Cox, S.J.D.: Geographic Information - Observations and Measurements (OGC Abstract Specification Topic 20) (same as ISO 19156:2011) (2011)
5. DataCite Metadata Working Group: DataCite Metadata Schema 4.0. Technical report, DataCite e.V., Hannover, Germany, May 2016. doi:10.5438/0012
6. Devaraju, A., Klump, J.F., Cox, S.J.D., Golodoniuc, P.: Representing and publishing physical sample descriptions. Comput. Geosci. **96**, 1–10 (2016). doi:10.1016/j.cageo.2016.07.018
7. Klump, J., Cox, S.J.D., Wyborn, L.A.I.: Connecting geology with the internet of things. In: Towards Unified Global Research. Melbourne, VIC, Australia, October 2014. http://eresearchau.files.wordpress.com/2014/07/eresau2014_submission_80.pdf
8. Klump, J., Huber, R.: 20 years of persistent identifiers - which systems are here to stay? Data Sci. J. **16**, 9 (2017). doi:10.5334/dsj-2017-009
9. Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S.: Implementation guidelines for the open archives initiative protocol for metadata harvesting. Technical report (2002). http://www.openarchives.org/OAI/2.0/guidelines.htm
10. Lehnert, K.A., Klump, J., Arko, R.A., Bristol, S., Buczkowski, B., Chan, C., Chan, S., Conze, R., Cox, S.J., Habermann, T., Hangsterfer, A., Hsu, L., Milan, A., Miller, S.P., Noren, A.J., Richard, S.M., Valentine, D.W., Whitenack, T., Wyborn, L.A., Zaslavsky, I.: IGSN e.V.: registration and identification services for physical samples in the digital universe. In: American Geophysical Union, Fall Meeting 2011 (2011)
11. Lehnert, K., Carbotte, S., Ryan, W., Ferrini, V., Block, K., Arko, R., Chan, C.: IEDA: integrated earth data applications to support access, attribution, analysis, and preservation of observational data from the ocean, earth, and polar sciences. Geophysical Research Abstracts 13 (2011)

12. 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.), Abstract IN53C-07 (2006). http://abstractsearch.agu.org/meetings/2006/FM/sections/IN/sessions/IN53C/abstracts/IN53C-07.html
13. McNutt, M., Lehnert, K.A., Hanson, B., Nosek, B.A., Ellison, A.M., King, J.L.: Liberating field science samples and data. Science **351**(6277), 1024–1026 (2016). http://science.sciencemag.org/content/351/6277/1024
14. Schindler, U., Diepenbroek, M.: Generic XML-based framework for metadata portals. Comput. Geosci. **34**(12), 1947–1955 (2008). doi:10.1016/j.cageo.2008.02.023
15. The Australian Antarctic program (AAp): The Australian Antarctic program data policy 2014 (applied to projects approved between 2 April 2013 and 21 June 2015). Online (June 2015). https://data.aad.gov.au/aadc/about/data_policy_2014.cfm
16. The National Science Foundation: Proposal and award policies and procedures guide (part ii - award & administration guide). Online February 2014. https://www.nsf.gov/pubs/policydocs/pappguide/nsf14001/aagprint.pdf
17. Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Dőring, M., Giovanni, R., Robertson, T., Vieglais, D.: Darwin core: an evolving community-developed biodiversity data standard. PLoS ONE **7**(1), e29715 (2011). doi:10.1371/journal.pone.0029715