CODATA INTERNATIONAL SCIENCE COUNCIL

Towards Globally Unique Identification of Physical Samples: Governance and Technical Implementation of the IGSN Global Sample Number

RESEARCH PAPER

JENS KLUMP
KERSTIN LEHNERT
DAMIAN ULBRICHT
ANUSURIYA DEVARAJU
KIRSTEN ELGER
DIRK FLEISCHER
SARAH RAMDEEN
LESLEY WYBORN

Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Persistent unique identifiers (PID) are a critical element in digital research data infrastructure to unambiguously identify, locate, and cite digital representations of a growing range of entities – publications, data, instruments, organizations, funding awards, field programs, and others. The IGSN was developed as the International Geo Sample Number to provide a persistent, globally unique, web resolvable identifier for physical samples. IGSN is both a governance and technical system for assigning globally unique persistent identifiers to physical samples. Even though initially developed for samples in the geosciences, the application of IGSN can be and has already been expanded to other domains that rely on physical samples and collections. This paper describes the current architecture and technical implementation of IGSN, how IGSN relates to other sample identifiers, and how its technical systems are supported by an international governance structure.

CORRESPONDING AUTHOR:
Jens Klump
Mineral Resources, CSIRO, Perth WA, Australia
jens.klump@csiro.au

KEYWORDS:
persistent identifier; metadata; physical specimen; research data infrastructure

TO CITE THIS ARTICLE:
Klump, J, Lehnert, K, Ulbricht, D, Devaraju, A, Elger, K, Fleischer, D, Ramdeen, S and Wyborn, L. 2021. Towards Globally Unique Identification of Physical Samples: Governance and Technical Implementation of the IGSN Global Sample Number. *Data Science Journal*, 20: 33, pp. 1–16. DOI: *https://doi.org/10.5334/dsj-2021-033*

INTRODUCTION: PERSISTENT IDENTIFIERS FOR SAMPLES IN RESEARCH

Physical samples (aka. physical specimen, Haller et al, 2017) are at the heart of many scientific disciplines – they are the raw material, and basic element for reference, study, and experimentation and representative of a wider population or a larger spatial context (Devaraju et al, 2016): in particular in the natural and environmental sciences, material sciences, agriculture, anthropology, archaeology, and biomedicine. They include biodiversity samples, synthetic materials, rock or mineral samples, soil or sediment cores, seed accessions, water quality samples, archaeological artefacts, human tissue samples, and many more physical artefacts. As the canonical reference for many scientific observations and measurements, information about their origin, composition and whereabouts is required as part of the transparency of scientific experiments and resultant publications. Several publications and policies (e.g. Australian Antarctic Data Centre, 2015; CODATA, 2019; McNutt et al, 2016; National Science Foundation, 2020) have highlighted the importance of curating and publishing sample information, which reflects similar developments in recent years for research data. However, compared to the development of research data infrastructures, the development of infrastructures that enable physical samples to be discovered, described, and reused beyond disciplinary or institutional boundaries is only in its infancy (see e.g. Davies et al, 2021; Lannom et al, 2019).

A fundamental first step towards the discoverability of samples over the Web in an unambiguous way is a mechanism for persistent identification of samples. Organisations such as museums, geological surveys, and networked research programmes like the International Ocean Discovery Program (IODP) have systems in place for the unique identification of their samples. These systems, however, are limited to the scope of the organisation, they do not extend beyond institutional boundaries. Taking the step beyond institutional boundaries, additional challenges, e.g., ambiguous sample names arise. For example, ***Figure 1*** shows the sampling locations of samples with the same label 'M1' specified in literature and collected in the EarthChem database. Similarly, EarthChem lists eleven alternative names for the sample ARGAMPH-003 (*https://explore.earthchem.org/specimen/33522*), collected from the East Pacific Rise by dredging (*http://igsn.org/SIO0000003*). Addressing these types of ambiguities was a primary motivation for the development of a globally unique identifier, then called the International Geo Sample Number (IGSN) (see Lehnert et al, 2004; Lehnert and Klump, 2008).

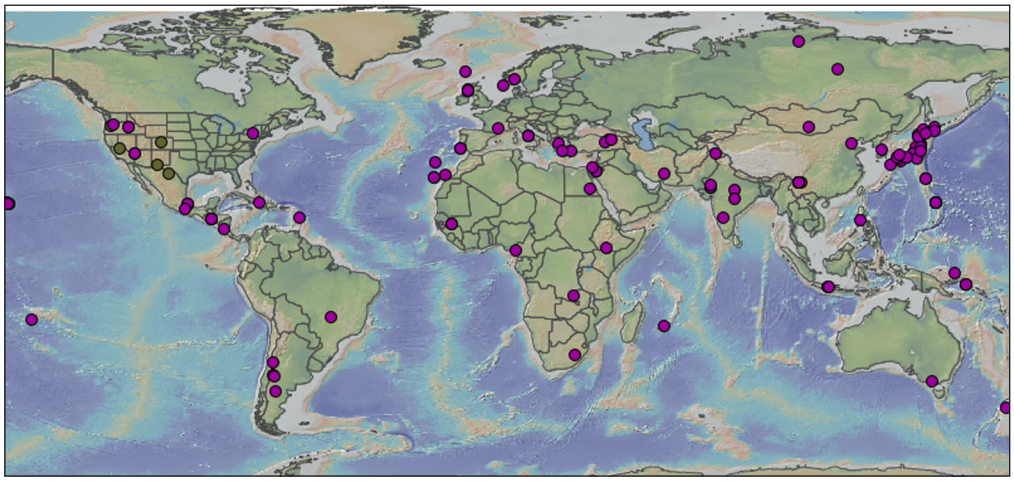

Figure 1 Locations of samples labelled "M1" in publications and collected in the EarthChem database.

IGSN is now both a governance and technical system for assigning and preserving globally unique persistent identifiers to physical samples and collections. ***Figure 2*** shows an example of a sample identified by an IGSN.

IGSN is governed by an international body, the IGSN Implementation Organization (IGSN e.V., *http://www.igsn.org*) (Lehnert et al, 2011). Even though developed in the geosciences, the application of IGSN as an identifier for physical samples is not limited to the geosciences but is increasingly adopted by other domains handling physical samples. To reflect the broadened scope of its application, the IGSN e.V. is currently considering changing the name of the identifier, retaining the acronym "IGSN". Since their invention in 2004 (Lehnert et al, 2004), the number of IGSN registrations has grown to 9.9 million (status October 2021).

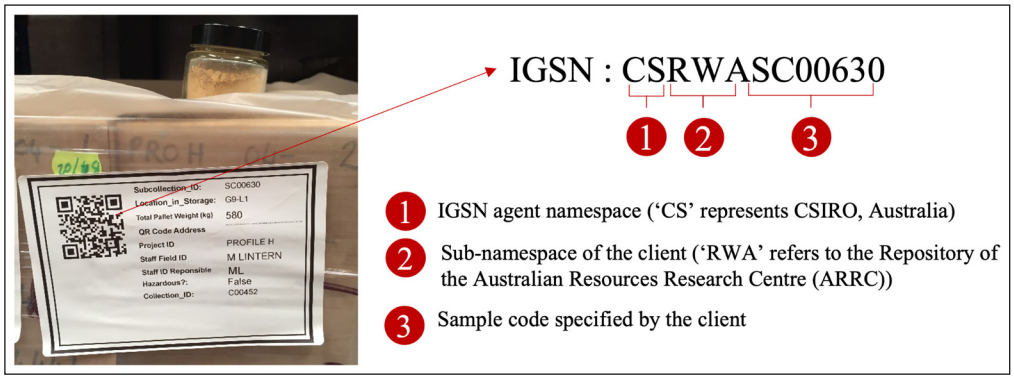

Figure 2 A rock sample collection curated at the Repository of the Australian Resources Research Centre (ARRC). Its IGSN 'CSRWASC00630' consists of the top-level namespace 'CS' administered by the IGSN agent 'CSIRO', the sub-namespace 'RWA' identifying the client (here ARRC), and the sample code 'SC00630'. Sub-namespaces and sample codes are managed by the IGSN agent who must ensure that the combinations are unique within their system. Modified after (Devaraju et al, 2016).

In September 2021, IGSN e.V. and DataCite entered a partnership that will transfer the minting of IGSN identifiers into the DataCite infrastructure and services. IGSN identifiers will become DataCite DOIs, and any DataCite member will be able to register identifiers for samples as IGSNs through DataCite. This paper describes the development of IGSN up to this point.

In the past years, we have seen progress on curating and publishing collections and samples using persistent identifiers, the International Committee for Documentation (CIDOC) published a 'statement on linked data identifiers for museum objects' (International Council of Museums, 2012). The statement recommends actionable URI (Universal Resource Identifier) for collection objects but does not provide further guidance on URI syntax or appropriate identifier systems. At the same time, community-specific sample identifier systems have been introduced, most actively pursued in life sciences and geosciences. For example, the bioinformatics and biodiversity communities created an identifier system (Life Sciences Identifier, LSID) to identify samples and biological taxa. Due to various socio-technical reasons, LSID was not adopted, and the community pragmatically decided to discontinue LSID in favour of "Cool URIs" (Groom et al, 2017; Güntsch et al, 2017). Since there is no way to tell which URIs are "Cool URIs", this approach comes with the risk that the chosen URI will not be persistent (Klump and Huber, 2017). The Food and Agriculture Organisations of the United Nations (FAO) recently decided to use DOIs to identify food crops (Alercia et al, 2018).

GOVERNANCE OF IGSN

Like all persistent identifier systems, IGSN is a socio-technical system. This means that IGSN needs a governance framework that ensures the persistence and uniqueness of the minted identifiers (Golodoniuc et al, 2017; Klump and Huber, 2017). The IGSN governance framework defines the role of IGSN e.V., allocating agents, and clients (see ***Figure 3***). The IGSN e.V. and allocating agents develop relevant best practices in collaboration with IGSN communities, oversee the allocation of namespaces for the IGSN, coordinate the description of samples with standardised metadata, using standardised vocabularies to facilitate machine-readability and semantic cross-linking of resources (Genova et al, 2017).

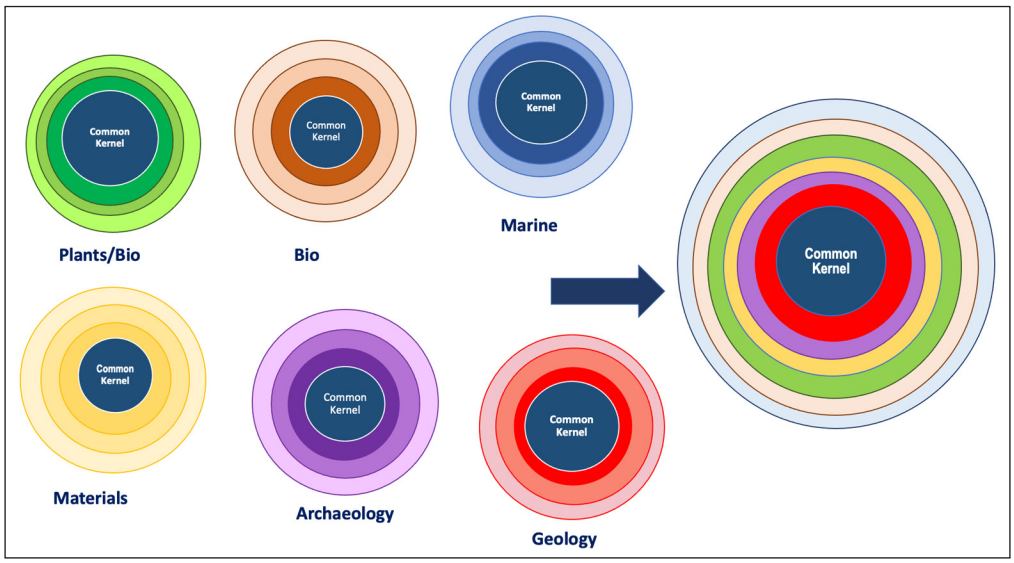

Figure 3 Some communities will need extensions around a core set of descriptive metadata, while other communities will need a separate core set of metadata elements and specific extensions to describe their samples adequately.

ENSURING UNIQUENESS AND PERSISTENCE

Starting an identifier system based on the Handle.net system (Kahn and Wilensky, 1995) is not too difficult. The challenge lies in creating a governance system that supports the goals of global uniqueness of the identifier and its persistence (Bütikofer, 2009). Like other persistent identifier systems, IGSN relies on its governance to ensure that an IGSN always resolves to a Web resource with a URL. This does not mean that the sample itself has to be persistent. Often samples are destroyed in an analytical process or are discarded (e.g., water samples), and the Web resource representing the sample should provide the user with information on the current status of the sample.

In the context of IGSN, these requirements are best met by a hierarchical delegation model (Bechtold, 2003) to assign namespace governance and responsibilities for IGSN namespaces. An IGSN is composed of a namespace, sometimes with a sub-namespace, and a sample code (***Figure 2***). This structure can be likened to the structure of telephone numbers, which consist of an international country code, an area code and the telephone number of the subscriber. In a hierarchical namespace governance model, the IGSN agent does not need to negotiate the allocation of individual identifiers with IGSN e.V., but may solely negotiate them with its clients in its allocated namespace. By delegating parts of the namespace governance to IGSN agents the communication overhead between the agents and the IGSN registry is minimised while offering more flexibility for agents. This is analogous to the current practice for assigning DOI, where DOI agents are allocated a prefix namespace, in which they can mint identifiers as needed.

Namespaces are governed and assigned to agents by IGSN e.V. Within their namespace, each agent may have their own naming convention and must ensure the uniqueness of the assigned IGSN. To minimise administrative efforts, most agents extend the hierarchical delegation pattern by using sub-namespaces followed by unique sample codes as illustrated in ***Figure 2***. This use of prefixes also allows the integration of local naming conventions already in use, thus making it easier to transform the locally unique identifier into one that is globally unique. An example was described by Conze et al (2017) for the International Continental Drilling Program (ICDP). In this example, the IGSN identifiers integrate the established ICDP naming conventions and IGSN can therefore be generated directly from the database without any changes to already established working procedures, naming conventions, or data systems. The hierarchical delegation pattern can also be used to assign sub-namespaces to teams in field sampling campaigns, allowing them to create unique IGSNs offline while in the field and then register them later.

Unlike most other persistent identifiers, IGSNs are not only used by machines but are often written and transcribed by humans onto sample labels, sample bags, and the like. The labelling of sample containers or the incorporation of IGSNs in tables in research articles puts practical limits on the number of characters that can be used. Out of these practical considerations, IGSN e.V. suggests a length between 9–12 characters for an IGSN, but this is not a binding requirement. To reduce the risk of mistyping, an IGSN is case insensitive and IGSN e.V. recommends that care is taken in the use of characters that can easily be confused, in particular in handwriting, such as '1' and 'I', or '0' and 'O'.

An IGSN is resolved as a URI through *http://igsn.org/*, which uses an HTTP redirect to the underlying Handle.net address. For example, MBCR5034RC57001 refers to a sample from the International Ocean Discovery Program (IODP), registered by MARUM (Center for Marine Environmental Sciences, University of Bremen) on behalf of IODP. The resulting IGSN URI for the identifier is *http://igsn.org/MBCR5034RC57001*. The browser resolves this URI to the URL of its corresponding landing page by redirecting the request to Handle.net which then further redirects to the URL of the landing page. Since igsn.org simply redirects to Handle.net, it is also possible to resolve an IGSN through any Handle.net resolver (e.g. *http://hdl.handle.net/10273/MBCR5034RC57001*).

DESCRIBING SAMPLES FOR DISCOVERY AND REUSE

A key aspect for the discovery of samples on the Web is their digital representation through a suitable metadata model (Devaraju et al, 2016). There are several domain-specific metadata models available such as Darwin Core (DwC), an extension of Dublin Core, and community-driven metadata standard for sharing biodiversity data (Wieczorek et al, 2012). Similarly, the

Biological Collections Ontology (BCO) is an application ontology to link biodiversity collections from various resources, including samples of organisms, ecological surveys and samples in metagenomic studies (Walls et al, 2014). These disciplinary metadata models can be regarded as disciplinary or domain-specific supplement to the IGSN Description Metadata Schema (see also ***Figure 3***).

For the geosciences, the System for Earth Sample Registration (SESAR, *http://www.geosamples.org*) developed a metadata model schema to describe basic concepts of geological samples (Lehnert, 2011). The U.S. Geoscience Information Network (USGIN) hosts several common content models for the geoscience domain, including the USGIN content model for Physical Samples (Hills, 2015). In a more general context, ISO 19156:2011 (Observations and Measurements, O&M) (*OGC Observations and Measurements v2.0 also published as ISO/DIS 19156*, 2013) includes a common concept for 'Specimen' with minimum attributes such as materialClass, samplingLocation, samplingTime and size. In the O&M model, a 'Specimen' is a specialization of 'SamplingFeature' which is further classified into various spatial sampling features such as cross-sections, transects and boreholes. The Sensor, Observation, Sample, and Actuator (SOSA) ontology provides constructs to represent sampling information including the relation between a sample and its feature of interest upon which the sampling activity was carried out. All of these models have in common that they interpret a sample as representing some larger feature of interest (Cox, 2020).

The description schemas discussed above may be well suited for samples in the earth and environmental sciences, but will often struggle to accommodate use cases from other disciplines, e.g. archaeology, veterinary medicine or material science. Some of the extended scope may be accommodated in community-specific extensions, while other user communities might need entirely different sample description schemas. Other use cases need to integrate with existing schemas and vocabularies through crosswalks (Damerow et al, 2021). ***Figure 3*** illustrates the concept of community extensions to core description schemas ('bullseye') and the parallel existence of multiple description schemas.

TECHNICAL IMPLEMENTATION

The use of IGSN in a broadening range of research domains and by a diverse range of stakeholders requires a high degree of flexibility. Many of the concepts employed in the implementation of the IGSN are derived from the lessons learned while implementing Digital Object Identifiers (DOI) for the publication and citation of research data, especially from the example of DataCite e.V. as an operator of this system (Brase, 2009) and its precursor (Klump et al, 2006). This includes the choice of Handle as the underlying persistent identifier protocol, which was chosen in 2008 to keep IGSN as much as possible interoperable with DataCite.

The concept of IGSN started in 2004 in a precursor project as the System for Earth Sample Registration (SESAR) (Lehnert et al, 2004). At this time DataCite had not yet been founded and the few first DOIs for datasets were registered through the German National Library of Science and Technology (TIB) (Brase, 2004). TIB saw the need for a persistent identifier system for samples but considered this use case to be out of scope of their DOI operations. Becoming a member of the International DOI Foundation (IDF) to be able to mint DOI independently of TIB was ruled out due to the high fees for IDF membership. Therefore, the consortium decided to base the IGSN on a generic implementation of the Handle.net System, which went into operation in 2008 (Lehnert and Klump, 2008). Following the example of DataCite, the governance and operation of the central IGSN services were incorporated into what was then called the International Geo Sample Number Implementation Organization (IGSN e.V.) (Lehnert et al, 2011).

THE IGSN SYSTEM ARCHITECTURE

IGSN e.V. developed the IGSN registry based on DataCite's Metadata Store (*https://mds.datacite.org/*, Fenner et al, 2019). The IGSN registry provides a REST API and a web user interface to manage IGSN registrations. An IGSN agent forwards the IGSN registrations including the registration metadata to the IGSN registry based on the Registration Metadata schema (see section below). All IGSNs are registered in the Handle System through a Handle Server.

It is the role of the IGSN agent to ensure the uniqueness of the identifier and the persistence of the associated landing page describing the sample (Devaraju et al, 2017). The Handle.net Registry manages the resolution of IGSN identifiers to the URL of the landing page describing the sample. IGSN operates in the handle-namespace <10273>. Further technical details on the registry (API) are available online (*http://igsn.github.io/registration/*). ***Figure 4*** gives a schematic overview of the IGSN system architecture for the minting of IGSN and syndication of IGSN catalogues by IGSN Agents.

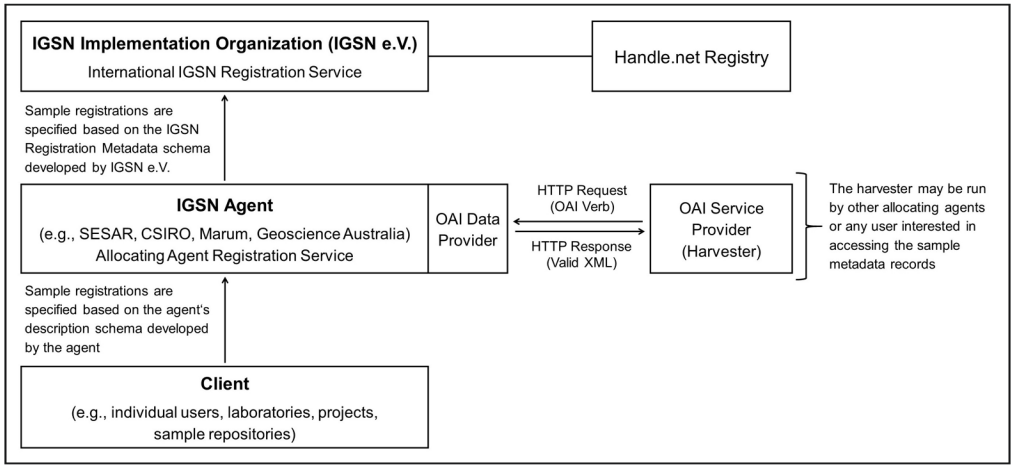

Figure 4 Schematic overview of the IGSN architecture for the minting of IGSN and metadata syndication. Modified after (Devaraju et al, 2017).

IGSN METADATA MODELS

A distinctive feature of the IGSN approach to metadata is its separation of registration metadata from description metadata. Other PID registries (e.g. DataCite, ORCID) require the transmission of a common set of metadata as part of the registration process, which is then incorporated into a central catalogue. In contrast to this common practice, the IGSN registration process separates the registration of the identifier and the provision of description metadata using two different schemas. Separating the registration metadata of the identifier from the description of the object gives the IGSN system the flexibility to accommodate a greater variety of applications, which may require different metadata profiles to describe their samples, e.g. for different disciplines or use cases. This approach also matches the standard registration models in ISO 19135 (ISO, 2013) and ISO 11179 (Pon and Buttler, 2009) more closely (Devaraju et al, 2017).

The IGSN Registration Metadata schema (*http://schema.igsn.org/registration/*) describes:

1) the registrant information;
2) any state changes of the identifier, e.g., submitted, registered, deprecated; and
3) its association with other identifiers such as other IGSN, DOI, etc.

These relations are important to determine the lineage of a sample and how it relates to other objects. Through this metadata element, an IGSN can be linked to other entities, such as a parent-sample, derived child-samples, aggregates of samples (e.g. drill core sections or dredges), sampling features (e.g. outcrops, drill holes). The same element can be used to link to datasets associated with a sample, or publications in which the samples are referenced. The samples are cross-linked to other entities by referencing their persistent identification and describing the nature of this relationship through a controlled vocabulary. This procedure follows common practice among PID providers (see e.g. DataCite Metadata Working Group, 2018); related work in this direction is the patterns proposed by (Cox, 2020) to capture a chain of samples.

The IGSN Description Metadata schema (*http://schema.igsn.org/description/* and *http://igsn.github.io/metadata/*) is developed by the IGSN members with inputs from a community of practice in the earth and environmental sciences. It is used to catalogue a minimum set of descriptive properties of samples and sample collections, such as sample type, material type, contributor, and sampling activity, to aggregate catalogues of samples across IGSN agents into overarching portals. This schema was deliberately kept general to allow the compilation of a global catalogue of, e.g., geological and biological samples and sample collections. It is based on the principles of the DataCite Metadata Schema (DataCite Metadata Working Group, 2016) and modified in terms of cardinality and restrictions on particular metadata elements, while new

elements (e.g., geolocation, collection methods, materials) were added to represent essential sample information that goes beyond the requirements of a bibliographic catalogue. Wherever possible, existing controlled vocabularies (e.g., representing sample and material types) like GeoSciML (Sen and Duffy, 2005), Eionet-GEMET (European Environment Agency, 2004), or the material and sample types as defined in the Observations Data Model (ODM) (Horsburgh et al, 2016) were incorporated into the scheme to enrich the metadata and promote consistency of metadata entries. Additionally, a number of vocabularies required to describe physical samples were requested to be added to the ODM registry. In the longer term, the IGSN initiative strives for machine-readable and well-governed standard vocabularies that can be applied to express different aspects of sample metadata. These vocabularies should follow standards, e.g., Simple Knowledge Organization System (SKOS), and should be identified with URIs.

Metadata schemas always encode an information model with a set of applications in mind (Devaraju et al, 2016).

METADATA SYNDICATION

The original system design uses the Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH) (Lagoze et al, 2002) to share metadata and disseminate catalogues of samples across IGSN agents. A useful feature of OAI-PMH is that it allows serving more than one metadata schema. IGSN agents can therefore develop domain-specific description schemas with their clients to serve their specific communities. Allowing application-specific metadata profiles gives IGSN agents greater flexibility to describe samples for different applications, e.g. allowing harvesting of certain sample types with their domain-specific description metadata required for domain-specific catalogues and applications (Devaraju et al, 2017). These communities are centred around use cases as communities of practice. In addition to IGSN common and community-specific profiles, it is good practice that OAI-PMH servers also offer metadata following the Dublin Core schema to allow harvesting of metadata by generic OAI-PMH clients that are not aware of the IGSN description schema. IGSN provides a mapping of IGSN Descriptive Metadata elements to Dublin Core elements online (see *http://igsn.github.io/oai/*).

The open nature of using a common protocol for sharing metadata and assembling sample catalogues allows anybody with knowledge of the OAI-PMH endpoints to build applications that make use of these metadata. The AuScope Discovery Portal at *http://portal.auscope.org.au/* is an example of an application that aggregates IGSN catalogues hosted by different IGSN Agents by harvesting the metadata catalogues and making them available through a search portal, thus enabling the discovery of samples. However, OAI-PMH was never built to serve several million records, and the fact that some IGSN Agents catalogue millions of samples has shown the limitations of OAI-PMH as a way of syndicating very large volumes of metadata. It is therefore foreseeable that IGSN will abandon OAI-PMH as its mechanism for metadata syndication and adopt a standard based on common Web technologies and schema.org combined with a sitemap file offering a list of all records available at an agent site (Fils et al, 2020).

APPLICATIONS OF IGSN

LINKING PHYSICAL OBJECTS TO THE WEB

The digital representation of a sample in the IGSN system is its landing page. The presentation of the sample metadata, or IGSN Landing Page, differs among portals and catalogues. Similar to practices in the DOI system (TIB Hannover, 2012), IGSN agents are required to display a description of the sample that is identified by an IGSN. In the spirit of "intelligent openness" (Royal Society, 2012), parts of a metadata record can be withheld to protect sensitive information, e.g. of vulnerable sites or threatened species. Communities of practice, like the example from the earth and environmental sciences discussed above, agree on a core set of metadata elements to display. Additional elements can be added by individual agents to improve the discoverability of samples, such as sample images, maps, and display of the hierarchical relationship of objects. ***Figure 5*** shows the example of an IGSN Landing Page for a sample from the International Continental Scientific Drilling Programme (ICDP). SESAR landing pages include a QR code for the IGSN that encodes the URL of the landing page. Users can copy and paste the QR code into sample labels and thus directly access landing pages from their mobile app using QR code readers.

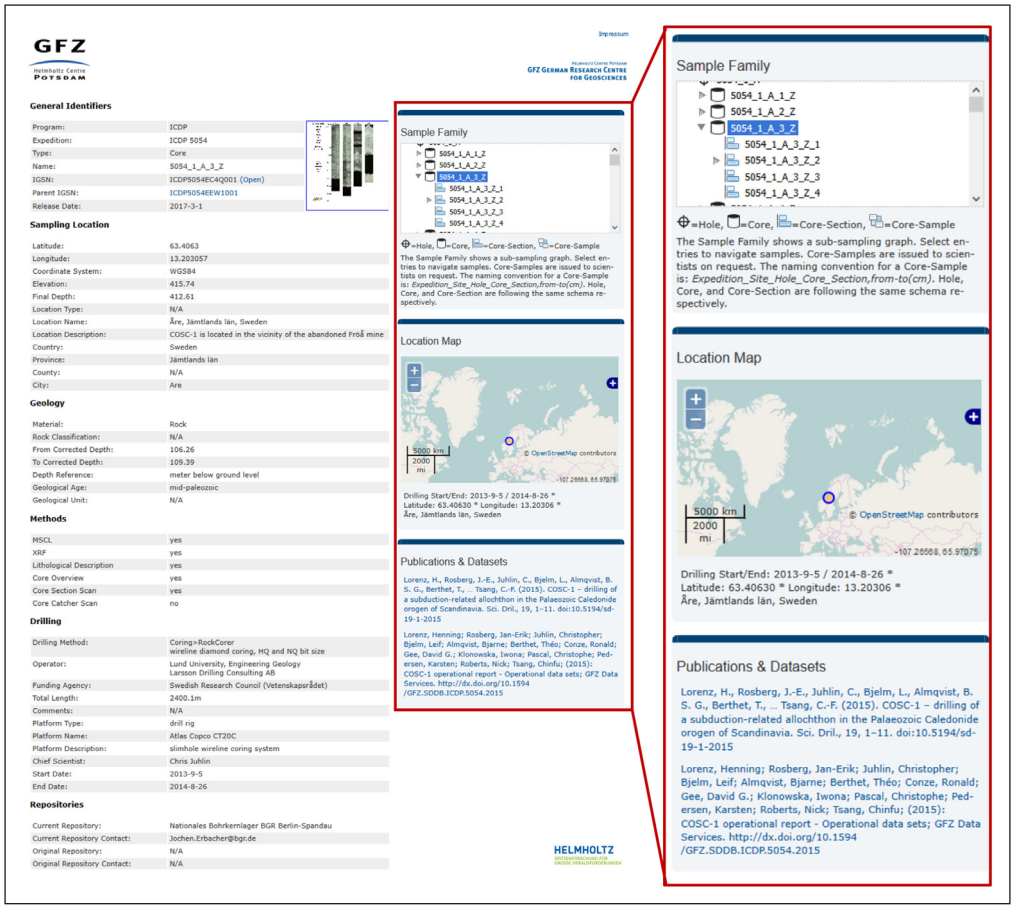

Figure 5 Example of an IGSN Landing Page published by GFZ Data Services for *http://igsn.org/ICDP5054EXF4601*, a core sample curated at the Core Repository for Continental Drilling of the German Federal Geological Survey in Berlin-Spandau, Germany.

LOCATING PHYSICAL SAMPLES

There are a number of ways on the side of the physical sample to link it with its digital representation (Kahn and Wilensky, 1995; Lannom et al, 2019) on the web using IGSN. The simplest method is to permanently affix a label to a sample, or by writing or engraving its IGSN onto it or its container along with its local accession or inventory number. Because space for labels is limited, in particular on small samples, QR tags or barcodes are more convenient as they offer the possibility to encode any identifying information in a machine-readable way. The technically most accessible way for using QR codes is to encode an IGSN as an actionable handle URI. Ideally, a label should show the QR code, the IGSN, as well as any inventory number in a human-readable way. ***Figure 6*** shows examples of this use case where labels with the IGSN in human-readable form and as a QR code have been attached to a sample and a box holding several samples.

Figure 6 An example of a coral sample and a rock collection identified through IGSN at the LDEO Core Repository and the Repository of the Australian Resources Research Centre (ARRC).

Right from the start, IGSN served as an identifier for both individual samples and for aggregations of samples. Over time other use cases arose, which will be discussed in this section. The first use case, identifying samples and linking them with their virtual representation is pretty straightforward. The idea of aggregate objects that combine several samples under one identifier follows the example of DataCite DOI used to aggregate several individually identified datasets under one DOI. Related to this idea is the identification of sampling features, e.g. boreholes, from which several related samples were taken. And finally, we will describe how to link samples through IGSN to other related objects, both physical and virtual.

Another form of aggregation of samples are 'collections' in the sense in which DataCite allows DOI objects to be aggregated under one identifier. An example from DataCite is the dataset published by König-Langlo and Gernandt (2009) of radiosonde ascends near the Georg Forster Antarctic Research Station. While each radiosonde dataset is identified by its individual DOI, the entire set of 426 radiosonde ascends is identified by a collective DOI. Following the same pattern, several samples, each identified by individual IGSN, can be aggregated into a collection of samples with its own IGSN identifying the collection.

LINKING DATA ACROSS REPOSITORIES AND CROSS-LINKING BETWEEN SAMPLES AND RELATED RESOURCES

An important aspect of Web resolvable identifiers is their ability to act as anchors for relations between objects such as samples, data, literature, instruments, authors, custodians, organisations, and many more. In this sense, IGSNs act as anchors for data and literature to the physical samples from which they were derived. IGSNs can, therefore, serve as anchors to the provenance of all related resources (texts, images, data, code, derived samples) underpinning the published research results.

A major step for the visibility and discoverability of IGSNs in data publications was the formal inclusion of IGSNs as RelatedIdentifierType in the DataCite 4.0 Metadata Schema (DataCite Metadata Working Group, 2016). This enables the integration of actionable IGSNs directly in the standardised and machine-readable metadata of research datasets (for example, see *http://igsn.org/ICDP5054EXF4601*) and allows IGSNs to be discovered in catalogues of research data repositories (e.g., GFZ Data Services (*http://dataservices.gfz-potsdam.de*), PANGAEA (*https://www.pangaea.de*), EarthChem Library (*https://www.earthchem.org/library*) and other portals harvesting metadata (e.g., DataCite Search, *https://search.datacite.org/*). ***Figure 7*** shows how IGSNs link from publications and data tables in publications can be used to link directly to sample descriptions.

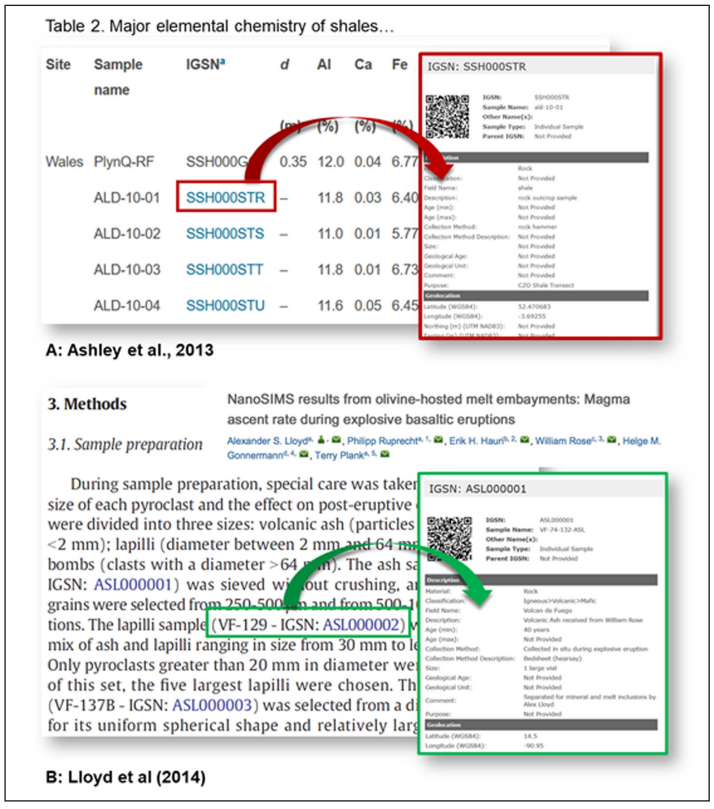

Figure 7 Referencing samples by using IGSNs in data tables (A: Dere et al, 2013) and in the body text (B: Lloyd et al, 2014) of journal articles. The executable IGSN links resolve to the online sample profiles (IGSN Landing Page).

As described for scholarly literature, where references are presented as actionable DOIs, it is recommended to also include actionable IGSN identifiers in datasets. This should be done in addition to the inclusion of IGSNs in the metadata of data publications via DataCite's "relatedidentifiertype". The IGSNs in datasets should be active Web links enabling cross-linking the data values with the online description of the sample. Such cross-linking benefits researchers since it enables unambiguous reference and effortless discovery of as well as direct access to contextual information about samples in the datasets. As an example, ***Figure 8*** displays the HTML view of a dataset published through the IGSN-external data repository PANGAEA.

Data

Download dataset as tab-delimited text *(use the following character encoding:* UTF-8: Unicode (PANGAEA default) *)*

1 Sample label	2 Depth [m]	3 Depth comp [mcd]	4 Vol [cm³]	5 Date (of sampling)	6 IGSN	7 Date (of freeze drying)	8 Date (of grinding)	9 Date (start of measurement)	10 TC [%]	11 S [%]	12 TOC [%]
347-M0059B-1H-1,63.5-65.5	60.635	61.931734	10	2014-01-25T16:20	IBCR0347EXIW701	2014-01-28	2014-01-28	2014-02-06T14:11	2.9010	0.13000	0.348
347-M0059B-2H-1,65-66	63.950	64.650001	10	2014-01-25T17:22	IBCR0347EXJY701	2014-01-28	2014-01-29	2014-02-06T14:13	2.1610	0.14800	0.798
347-M0059B-2H-2,65-66	65.450	66.150001	10	2014-01-25T17:25	IBCR0347EXLY701	2014-01-28	2014-01-28	2014-02-06T14:14	3.0810	0.18000	0.864
347-M0059B-3H-1,65-66	67.250	67.768696	10	2014-01-25T17:53	IBCR0347EXVZ701	2014-01-28	2014-01-28	2014-02-06T14:16	3.3470	0.15800	0.337
347-M0059B-3H-2,65-66	68.750	69.268696	10	2014-01-25T17:58	IBCR0347EXWZ701	2014-01-28	2014-01-28	2014-02-06T14:17	3.9110	0.16200	0.481
347-M0059B-3H-3,25-26	69.850	70.368696	10	2014-01-25T18:05	IBCR0347EX40801	2014-01-28	2014-01-29	2014-02-06T14:18	3.6300	0.15900	0.518
347-M0059B-4H-1,84.5-86.5	70.745	71.249452	10	2014-01-25T18:42	IBCR0347EXU0801	2014-01-28	2014-01-28	2014-02-06T14:20	3.9170	0.16900	0.475
347-M0059B-4H-2,85-87	72.240	72.744452	10	2014-01-25T18:43	IBCR0347EXV0801	2014-01-28	2014-01-28	2014-02-06T14:21	4.6630	0.16300	0.465
347-M0059B-5H-1,65-66	73.850	74.610540	10	2014-01-25T19:02	IBCR0347EX81801	2014-01-28	2014-01-28	2014-02-06T14:23	3.3890	0.14100	0.540
347-M0059B-5H-2,65-66	75.350	76.110540	10	2014-01-25T19:06	IBCR0347EXB1801	2014-01-28	2014-01-28	2014-02-06T14:24	4.3990	0.19700	0.409
347-M0059B-5H-3,18-20	76.380	77.140540	10	2014-01-25T19:11	IBCR0347EXH1801	2014-01-28	2014-01-28	2014-02-06T14:26	3.6480	0.14900	0.541
347-M0059B-6H-1,34.5-35.5	76.845	77.419787	10	2014-01-25T20:06	IBCR0347EX74801	2014-01-28	2014-01-28	2014-02-06T14:47	2.9570	0.13000	0.557
347-M0059B-6H-3,27-28	79.770	80.344787	10	2014-01-25T20:11	IBCR0347EX84801	2014-01-28	2014-01-29	2014-02-06T14:48	4.4070	0.25200	0.401
347-M0059B-7H-1,20-21	80.000	80.674786	10	2014-01-25T20:34	IBCR0347EXY5801	2014-01-28	2014-01-29	2014-02-06T14:50	4.2700	0.29800	0.488
347-M0059B-8H-1,85-86.5	83.950	83.586211	10	2014-01-26T11:31	IBCR0347EXLB801	2014-01-28	2014-01-28	2014-02-06T14:52	3.4970	0.46900	0.698
347-M0059B-8H-2,105-106.5	85.520	85.156211	10	2014-01-26T12:29	IBCR0347EXTC801	2014-01-28	2014-01-29	2014-02-06T14:53	0.9430	0.14800	0.111
347-M0059B-19H-1(CC),3-4.5	107.310	106.946211	10	2014-01-26T11:57	IBCR0347EXXB801	2014-01-28	2014-01-28	2014-02-06T14:55	3.5980	0.24900	0.417
347-M0059B-25P-1,13-14	117.430	117.066211	10	2014-01-26T11:52	IBCR0347EXRB801	2014-01-28	2014-01-28	2014-02-06T14:56	2.2670	0.44200	0.255
347-M0059B-29P-1,26-27	203.790	203.426211	10	2014-01-26T12:45	IBCR0347EXLD801	2014-01-28	2014-01-28	2014-02-06T14:58	11.8240	0.10000	0.003

Figure 8 An example dataset with related samples published at *https://doi.org/10.1594/PANGAEA.839477*. The data includes the variable "International Geo Sample Number" with actionable IGSN names (in green), e.g. *http://igsn.org/IBCR0347EXIW701*.

TRACKING SAMPLES FROM THE FIELD TO THE SAMPLE REPOSITORY

A central issue in managing and discovering samples is the use of ambiguous sample names. The most effective way to avoid this is by applying IGSNs at an early stage of the sample life cycle. Projects in Australia have used electronic field notebooks (Ballsun-Stanton et al, 2018) to document the sampling process and assign an IGSN (Golodoniuc et al, 2016; Noble et al, 2018; Reid et al, 2016) as part of their field sampling activities (***Figure 9***). In this case, the uniqueness of the IGSN is ensured by assigning sub-namespaces to sampling campaigns. Assigning an IGSN as early as possible in the sampling process allows samples to be tracked through different stages of their life cycle, e.g., sample handling and storage, laboratory analysis, and eventual disposal. Samples can also be tracked if they are moved to different laboratories or repositories. This practice aligns with the principles outlined in the W3C Working Draft on Extensions to the Semantic Sensor Network Ontology (Cox, 2020).

Figure 9 IGSN pre-allocation using the FAIMS electronic field notebook application (Ballsun-Stanton et al, 2018) during a geochemical soil sampling campaign in the Nullarbor Desert, South Australia (Noble et al, 2018).

REFERENCES TO SAMPLING FEATURES AND LOCATIONS

IGSN can be used to identify related entities that are closely linked to physical samples. Examples are boreholes, mines, outcrops, or other sites. They all have in common that they are not samples themselves, but, to follow O&M terminology, sampling features (Cox, 2020; Haller et al, 2019). From these sampling features, a number of samples could have been taken. In the example of a borehole, drill core is commonly not retrieved in one piece but in several segments or "runs". Each of these objects is individually identified, as are samples representing subsamples from these objects. The identifiers of these objects relate to each other, mirroring the hierarchical relationships between samples (see e.g. ***Figure 5*** and Conze et al, 2017).

CONCLUSIONS – LESSONS LEARNED AND OUTLOOK

The development of IGSN started from the practical question of how to uniquely identify geological samples, where the results from analyses on the samples were published and interpreted in multiple places in the literature at a later stage, to enable the results to be accurately correlated and to provide a two-way trail between the samples and results. Over time, additional use cases (e.g., samples of other earth sciences, synthetic materials and sampling features) were added to the scope of IGSN, and further ones are expected in the future. The system is designed to maintain flexibility to accommodate new uses of IGSN. The separation of registration and description metadata allowed us to start with the minting of IGSN while the development of a common description metadata schema was still in progress. The choice of OAI-PMH to syndicate metadata was made to allow the parallel use of generic and specific metadata schemas. However, as discussed above, we found that OAI-PMH does not scale well beyond a few million items and needs to be addressed in a growing system.

From the outset, the technical implementation followed pragmatic design decisions, which often were informed by equivalent experiences in the implementation of DataCite and its precursors, including the reuse of technical components that had already been developed for DataCite. IGSN itself is dedicated to and has certainly benefited from open science. In this spirit, all documentation and source code is made available online to promote reuse and collaborative development of the technical components. Since 2016, the DataCite Metadata Schema includes IGSN in their list of identifier types when pointing to related persistent identifiers (DataCite Metadata Working Group, 2016).

Our pragmatic approach resulted in a quick realisation of the basic service with additional features being added later. However, this pragmatic implementation came at the cost of legacy issues that need to be addressed to allow a sustainable operation of the service. Some of these steps are incremental improvements that will add to the functionality of the system. Among these improvements are the implementation of a global search portal leveraging the metadata syndication through OAI-PMH and the enrichment of the metadata schema through semantic alignment with SOSA/SSN.

Considering that the analysis of large numbers of samples formed the basis of datasets and subsequent interpretation of these data in a publication, the number of samples needed to be identified is potentially very large. The challenge of sharing metadata at very large scales was investigated as part of a project funded by the Alfred P. Sloan Foundation from 2018 to 2021 (Klump et al, 2020; Lehnert et al, 2020). The IGSN 2040 project's goal was to "achieve a trustworthy, stable, and adaptable architecture for the IGSN as a persistent unique identifier for material samples, both technically and organizationally". In the spring of 2019, the project hosted a technical workshop to discuss technological strategies which take advantage of modern technology such as cloud-based services and structured data "to achieve stable and trustworthy services of the IGSN to scale to the rapidly growing demands of its user community" (Klump et al, 2020b). The resulting recommendations are for the IGSN e.V. to transition from an XML-based metadata schema to a web architecture based on sitemaps and introduce the role of the 'Information Aggregator', which acts as a metadata harvester. In 2020, the IGSN 2040 project hosted a Technical Sprint, which successfully tested these recommendations with operational IGSN e.V. Allocating Agents (Fils et al, 2020). A follow-up Sprint is needed to test the role of the Information Aggregator.

Dealing with samples in scientific collections links IGSN deep into scientific practice and into the processes surrounding the curation of the record of science. IGSN, therefore, has to be regarded as a socio-technical system. While technical safeguards can be put into place to enforce basic rules, significant portions of the system governance rely on a social contract forming the base of a persistent system (Klump and Huber, 2017). Ensuring the uniqueness of minted identifiers in a partially asynchronous system, and maintaining the association between identifiers and online catalogues requires a decentralised system of governance. In the case of IGSN, we continued to develop the concept of ensuring the uniqueness of assigned IGSN through hierarchical namespace governance (Bechtold, 2003).

Over the past years, IGSN has grown dramatically from a niche solution for petrology to becoming a global identification system for samples with nearly 10 million registered objects. The uptake of IGSN by national geological survey organisations and major collections, as well as the integration of IGSN into the scientific record through links into the scientific literature, make IGSN a strong candidate solution for a globally unique identifier for physical samples (Hardisty et al, 2020; Thessen et al, 2019).

Through the IGSN 2040 project funded by the Alfred P. Sloan Foundation, IGSN investigated options for a sustainable business model and technical architecture. As an outcome from this project, IGSN e.V. and DataCite developed a Memorandum of Agreement to enter a partnership on persistent identifiers for physical samples. As a result, some of the details of the technical implementation of IGSN identifiers may change, but the overall principles will remain in place.

ACKNOWLEDGEMENTS

The authors would like to thank the members of IGSN e.V. for their continued support that allowed IGSN to grow over the years. We also like to thank our colleagues in the member organisations for their contributions to establishing IGSN and thank our colleagues from other institutions who contributed to the many discussions we had on persistent identifiers over the past years.

FUNDING INFORMATION

The original development of the IGSN was supported by the US National Science Foundation through grants to K. Lehnert (award nos. 04-45178, 05-14551, 05-50914, and 05-52123). Work on the Australian implementation was supported by the 2018-2020 Australian Research Data Commons (ARDC) Geoscience Data-enhanced Virtual Laboratory (GeoDeVL) Program. Designing the future of IGSN was supported by the Alfred P. Sloan Foundation in the IGSN 2040 project (Grant Agreement G-2018-11137).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Jens Klump: Conceived the IGSN system architecture, led the pilot test, synthesized results and wrote the manuscript.

Kerstin Lehnert: Conceived the IGSN, led the IGSN 2040 project and the development of new governance and business model, and contributed to writing and editing the manuscript.

Damian Ulbricht: Developed and maintained the technical framework for the GFZ implementation and the central IGSN Handle server and contributed to the writing and editing of the manuscript.

Anusuriya Devaraju: Contributed to the writing and editing of the manuscript. AD contributed to developing the IGSN description metadata and the implementation of the IGSN services at CSIRO.

Kirsten Elger: developed the GFZ allocating agent activities and contributed to the writing and editing of the manuscript.

Dirk Fleischer, Sarah Ramdeen and Lesley Wyborn: Contributed to the writing and editing of the manuscript.

AUTHOR AFFILIATIONS

Jens Klump *orcid.org/0000-0001-5911-6022*
Mineral Resources, CSIRO, Perth WA, Australia

Kerstin Lehnert *orcid.org/0000-0001-7036-1977*
Lamont-Doherty Earth Observatory, Columbia University, Palisades NY, USA

Damian Ulbricht *orcid.org/0000-0002-6298-758X*
Library and Information Services, GFZ German Research Centre for Geosciences, Potsdam, Germany

Anusuriya Devaraju *orcid.org/0000-0003-0870-3192*
TERN, University of Queensland, Brisbane, Australia

Kirsten Elger *orcid.org/0000-0001-5140-8602*
Library and Information Services, GFZ German Research Centre for Geosciences, Potsdam, Germany

Dirk Fleischer *orcid.org/0000-0003-0108-4675*
Center for Ocean and Society, Christian-Albrechts-Universität zu Kiel, Kiel, Germany

Sarah Ramdeen *orcid.org/0000-0003-1135-5942*
Lamont-Doherty Earth Observatory, Columbia University, Palisades NY, USA

Lesley Wyborn *orcid.org/0000-0001-5976-4943*
National Computational Infrastructure, Australian National University, Canberra, Australia

REFERENCES

Alercia, A, López, FM, Sackville Hamilton, NR and **Masella, M.** 2018. Digital Object Identifiers for food crops – Descriptors and guidelines of the Global Information System. Rome, Italy: Food and Agricultural Organization of the United Nations (FAO). Available at *https://agris.fao.org/agris-search/search.do?recordID=XF2018002129*.

Australian Antarctic Data Centre. 2015. The Australian Antarctic Program Data Policy. Kingston, TAS, Australia: Australian Antarctic Division. Available at *https://data.aad.gov.au/aadc/about/data_policy.cfm.*

Ballsun-Stanton, B, Ross, SA, Sobotkova, A and **Crook, P.** 2018. FAIMS Mobile: Flexible, open-source software for field research. *SoftwareX*, 7: 47–52. DOI: *https://doi.org/10.1016/j.softx.2017.12.006*

Bechtold, S. 2003. Governance in Namespaces. *Loyola of Los Angeles Law Review*, 36(3): 1239–1320. DOI: *https://doi.org/10.2139/ssrn.413681*

Brase, J. 2004. Using Digital Library Techniques – Registration of Scientific Primary Data. In: Jones, M, Fox, EA, Shen, R, Stern, J, Yoo, K-Y, Pfitzmann, B, Iwama, K, Gao, W, Li, M, Wille, R, Zhang, J and Varadharajan, V (eds.), *Research and Advanced Technology for Digital Libraries*, 488–494. Heidelberg, Germany: Springer-Verlag. DOI: *https://doi.org/10.1007/978-3-540-30230-8_44*

Brase, J. 2009. DataCite – A Global Registration Agency for Research Data. In: *Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 2009. COINFO '09*, 257–261. Presented at the Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 2009. COINFO '09. DOI: *https://doi.org/10.1109/COINFO.2009.66*

Bütikofer, N. 2009. Catalogue of criteria for assessing the trustworthiness of PI systems (No. 13). Göttingen, Germany: Niedersächsische Staats und Universitätsbibliothek Göttingen. Available at *http://nbn-resolving.de/urn:nbn:de:0008-20080710227.*

CODATA. 2019. The Beijing Declaration on Research Data (position paper). Paris, France: CODATA. Available at *http://www.codata.org/news/361/62/The-Beijing-Declaration-on-Research-Data* [Last accessed 19 November 2019].

Conze, R, Lorenz, H, Ulbricht, D, Elger, K and **Gorgas, T.** 2017. Utilizing the International Geo Sample Number Concept in Continental Scientific Drilling During ICDP Expedition COSC-1. *Data Science Journal*, 16(1): 1–8. DOI: *https://doi.org/10.5334/dsj-2017-002*

Cox, SJD. 2020. Extensions to the Semantic Sensor Network Ontology (W3C Working Draft No. TR/2020/WD-vocab-ssn-ext-20200116). Open Geospatial Consortium, Inc. Available at *https://www.w3.org/TR/2020/WD-vocab-ssn-ext-20200116/*.

Damerow, JE, Varadharajan, C, Boye, K, Brodie, EL, Burrus, M, Chadwick, KD, Crystal-Ornelas, R, Elbashandy, H, Eloy Albes, RJ, Ely, KS, Goldman, AE, Habermann, T, Hendrix, V, Kakalia, Z, Kemner, KM, Kersting, AB, Merino, N, O'Brien, F, Perznan, Z, Robles, E, Sorensen, P, Stegen, JC, Walls, RL, Weisenhorn, P, Zavarin, M and **Agarwal, D.** 2021. Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ecosystem Sciences. *Data Science Journal*, 20(1): 11. DOI: *https://doi.org/10.5334/dsj-2021-011*

DataCite Metadata Working Group. 2016. DataCite Metadata Schema 4.0. Hannover, Germany: DataCite e.V. DOI: *http://doi.org/10.5438/0012*

DataCite Metadata Working Group. 2018. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data (No. Version 4.2). Hannover, Germany: DataCite e.V. DOI: *https://doi.org/10.5438/bmjt-bx77*

Davies, N, Deck, J, Kansa, EC, Kansa, SW, Kunze, J, Meyer, C, Orrell, T, Ramdeen, S, Snyder, R, Vieglais, D, Walls, RL and **Lehnert, K.** 2021. Internet of Samples (iSamples): Toward an interdisciplinary cyberinfrastructure for material samples. *GigaScience*, 10(giab028). DOI: *https://doi.org/10.1093/gigascience/giab028*

Dere, AL, White, TS, April, RH, Reynolds, B, Miller, TE, Knapp, EP, McKay, LD and **Brantley, SL.** 2013. Climate dependence of feldspar weathering in shale soils along a latitudinal gradient. *Geochimica et Cosmochimica Acta*, 122: 101–126. DOI: *https://doi.org/10.1016/j.gca.2013.08.001*

Devaraju, A, Klump, JF, Cox, SJD and **Golodoniuc, P.** 2016. Representing and Publishing Physical Sample Descriptions. *Computers & Geosciences*, 96: 1–10. DOI: *https://doi.org/10.1016/j.cageo.2016.07.018*

Devaraju, A, Klump, J, Tey, V, Cox, SJD and **Fraser, R.** 2017. Towards a web-enabled geo-sample web: An open source registration and management system for connecting geo-samples to the web. In: *Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings*. Presented at the FOSS4G 2017. Boston, MA: OSGEO. p. Article 16. DOI: *https://doi.org/10.7275/R5P26W9H*

European Environment Agency. 2004. GEMET – GEneral Multilingual Environmental Thesaurus. Copenhagen, Denmark: European Environment Agency. Available at *https://www.eionet.europa.eu/gemet/en/about/.*

Fenner, M, Peters, S, Leinweber, K, Zukowski, E and **Garza, K.** 2019. *DataCite Metadata Store (MDS)*. Hannover, Germany: DataCite e.V. Available at *https://github.com/datacite/mds* [Last accessed 17 July 2020].

Fils, D, Robertson, JC, Ramdeen, S and **Klump, J.** 2020. Building Community and Road-Testing the New IGSN System Architecture. In: *American Geophysical Union Fall Meeting 2020. Presented at the American Geophysical Union Fall Meeting 2020*. San Francisco, CA: American Geophysical Union. Available at *https://agu.confex.com/agu/fm20/meetingapp.cgi/Paper/726092.*

Genova, F, Arviset, C, Almas, BM, Bartolo, L, Broeder, D, Law, E and **McMahon, B.** 2017. Building a Disciplinary, World-Wide Data Infrastructure. *Data Science Journal*, 16(16). DOI: *https://doi.org/10.5334/dsj-2017-016*

Golodoniuc, P, Car, NJ and **Klump, J.** 2017. Distributed Persistent Identifiers System Design. *Data Science Journal*, 16: 34. DOI: *https://doi.org/10.5334/dsj-2017-034*

Golodoniuc, P, Devaraju, A and **Klump, JF.** 2016. The implementation of IGSN in the context of Australian mineral exploration. In: *Geophysical Research Abstracts, EGU2016–1562. Presented at the European Geosciences Union General Assembly 2016*. Vienna, Austria: Copernicus Society. Available at *http://meetingorganizer.copernicus.org/EGU2016/EGU2016-1562.pdf.*

Groom, Q, Hyam, R and **Güntsch, A.** 2017. Data management: Stable identifiers for collection specimens. *Nature*, 546(7656): 33–33. DOI: *https://doi.org/10.1038/546033d*

Güntsch, A, Hyam, R, Hagedorn, G, Chagnoux, S, Röpert, D, Casino, A, Droege, G, Glöckler, F, Gödderz, K, Groom, Q, Hoffmann, J, Holleman, A, Kempa, M, Koivula, H, Marhold, K, Nicolson, N, Smith, VS and **Triebel, D.** 2017. Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. *The Journal of Biological Databases and Curation*, 2017(1). DOI: *https://doi.org/10.1093/database/bax003*

Haller, A, Janowicz, K, Cox, SJD, Lefrançois, M, Phuoc, DL, Lieberman, J, García-Castro, R, Atkinson, RA and **Stadler, C.** 2019. The Modular SSN Ontology: A Joint W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation. *Semantic Web*, 10(1): 9–32. DOI: *https://doi.org/10.3233/SW-180320*

Hardisty, A, Saarenmaa, H, Casino, A, Dillen, M, Gödderz, K, Groom, Q, Hardy, H, Koureas, D, de la Hidalga, AN, Paul, D, Runnel, V, Vermeersch, X, van Walsum, M and **Willemse, L.** 2020. Conceptual design blueprint for the DiSSCo digitization infrastructure – DELIVERABLE D8.1. *Research Ideas and Outcomes*, 6: e54280. DOI: *https://doi.org/10.3897/rio.6.e54280*

Hills, DJ. 2015. Let's make it easy: A workflow for physical sample metadata rescue. *GeoResJ*, 6: 1–8. DOI: *https://doi.org/10.1016/j.grj.2015.02.007*

Horsburgh, JS, Aufdenkampe, AK, Mayorga, E, Lehnert, KA, Hsu, L, Song, L, Jones, AS, Damiano, SG, Tarboton, DG, Valentine, D, Zaslavsky, I and **Whitenack, T.** 2016. Observations Data Model 2: A community information model for spatially discrete Earth observations. *Environmental Modelling & Software*, 79: 55–74. DOI: *https://doi.org/10.1016/j.envsoft.2016.01.010*

International Council of Museums. 2012. Statement on Linked Data identifiers for museum objects. Helsinki, Finland: International Council of Museums. Available at *http://network.icom.museum/fileadmin/user_upload/minisites/cidoc/PDF/StatementOnLinkedDataIdentifiersForMuseumObjects.pdf.*

ISO. 2013. Geographic information – Procedures for registration of items of geographic information (Draft No. ISO/DIS 19135-1:2013). Geneva, Switzerland: International Organization for Standardization (ISO).

Kahn, R and **Wilensky, R.** 1995. A Framework for Distributed Digital Object Services (Technical Note No. tn09- 01). Reston, VA: Corporation for National Research Initiatives. Available at *http://hdl.handle.net/cnri.dlib/tn95-01.*

Klump, J, Bertelmann, R, Brase, J, Diepenbroek, M, Grobe, H, Höck, H, Lautenschlager, M, Schindler, U, Sens, I and **Wächter, J.** 2006. Data publication in the Open Access Initiative. *Data Science Journal*, 5: 79–83. DOI: *https://doi.org/10.2481/dsj.5.79*

Klump, J and **Huber, RX.** 2017. 20 Years of persistent identifiers – Which systems are here to stay? *Data Science Journal*, 16(9): 1–7. DOI: *https://doi.org/10.5334/dsj-2017-009*

Klump, J, Lehnert, KA, Wyborn, LAI, Ramdeen, S and **IGSN 2040 Steering Committee.** 2020. Building a sustainable international research data infrastructure – Lessons learnt in the IGSN 2040 project. In: *EGU General Assembly 2020. Presented at the EGU General Assembly 2020*. Vienna, Austria: Copernicus Society. pp. EGU2020–12001. DOI: *https://doi.org/10.5194/egusphere-egu2020-12001*

König-Langlo, G and **Gernandt, H.** 2009. Compilation of ozonesonde profiles from the Antarctic Georg-Forster-Station from 1985 to 1992. *Earth System Science Data*, 1(1): 1–5. DOI: *https://doi.org/10.5194/essd-1-1-2009*

Lagoze, C, Van de Sompel, H, Nelson, ML and **Warner, S.** 2002. Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting. Ithaca, NY: Open Archives Initiative. Available at *http://www.openarchives.org/OAI/2.0/guidelines.htm*. DOI: *https://doi.org/10.1108/07378830310479776*

Lannom, L, Koureas, D and **Hardisty, AR.** 2019. FAIR Data and Services in Biodiversity Science and Geoscience. *Data Intelligence*, 2(1–2): 122–130. DOI: *https://doi.org/10.1162/dint_a_00034*

Lehnert, KA. 2011. Metadata Standards for Sample-Based Observations. In: *Geophysical Research Abstracts, EGU2011–13113. Presented at the European Geosciences Union General Assembly 2011*. Vienna, Austria: Copernicus Meetings. Available at *https://presentations.copernicus.org/EGU2011/EGU2011-13113_presentation.pptx.*

Lehnert, KA, Goldstein, SL, Lenhardt, WC and **Vinayagamoorthy, S.** 2004. SESAR: Addressing the need for unique sample identification in the Solid Earth Sciences. In: *AGU Fall Meeting 2004*, SF32A-06. Presented at the AGU Fall Meeting 2004. San Francisco, CA: American Geophysical Union. Available at *http://adsabs.harvard.edu/abs/2004AGUFMSF32A..06L* [Last accessed 10 May 2016].

Lehnert, KA and **Klump, J.** 2008. Facilitating Research in Mantle Petrology with Geoinformatics. *Presented at the 9th International Kimberlite Conference*. Frankfurt (M), Germany: Copernicus Society. p. 9IKCA-00250. Available at *http://www.cosis.net/abstracts/9IKC/00250/9IKC-A-00250-1.pdf.*

Lehnert, KA, Klump, J, Arko, RA, Bristol, S, Buczkowsky, B, Chan, S-L, Chan, S, Conze, R, Cox, SJD, Habermann, T, Hangsterfer, A, Hsu, L, Milan, A, Miller, S, Noren, A, Richard, S, Valentine, DW, Whitenack, T, Wyborn, LA and **Zaslavsky, I.** 2011. IGSN e.V.: Registration and Identification Services for Physical Samples in the Digital Universe. *Presented at the American Geophysical Union Fall Meeting*. San Francisco, CA: American Geophysical Union. pp. IN13B–1324. Available at *http://abstractsearch.agu.org/meetings/2011/FM/sections/IN/sessions/IN13B/abstracts/IN13B-1324.html.*

Lehnert, K, Wyborn, L, Klump, J and **Ramdeen, S.** 2020. IGSN 2040 Organizational Steering Committee Workshop Report. Potsdam, Germany: IGSN e.V. DOI: *https://doi.org/10.5281/zenodo.3724722*

Lloyd, AS, Ruprecht, P, Hauri, EH, Rose, W, Gonnermann, HM and **Plank, T.** 2014. NanoSIMS results from olivine-hosted melt embayments: Magma ascent rate during explosive basaltic eruptions. *Journal of Volcanology and Geothermal Research*, 283: 1–18. DOI: *https://doi.org/10.1016/j.jvolgeores.2014.06.002*

McNutt, M, Lehnert, KA, Hanson, B, Nosek, BA, Ellison, AM and **King, JL.** 2016. Liberating field science samples and data. *Science*, 351(6277): 1024–1026. DOI: *https://doi.org/10.1126/science.aad7048*

National Science Foundation. 2020. Proposal and Award Policies and Procedures Guide (No. 20–1). Washington, DC: National Science Foundation. Available at *https://www.nsf.gov/pubs/policydocs/pappg20_1/index.jsp.*

Noble, RRP, Gonzalez-Alvarez, I, Reid, N, Krapf, C, Pinchand, T, Cole, D, Lau, I, Fox, D, Brant, F, White, AJR, Klump, J and **Petts, A.** 2018. Regional Geochemistry of the Coompana Area (Technical Report No. EP187470). Perth, WA, Australia: Commonwealth Scientific and Industrial Research Organisation. DOI: *https://doi.org/10.25919/5c59cf28d6fd2*

OGC. Observations and Measurements v2.0 also published as ISO/DIS 19156 (Abstract Specification No. 10-004r3) 2013. Arlington, VA: Open Geospatial Consortium. Available at *https://www.ogc.org/standards/om* [Last accessed 13 July 2021].

Pon, RK and **Buttler, DJ.** 2009. Metadata Registry, ISO/IEC 11179. In: Liu, L and Özsu, MT (eds.), *Encyclopedia of Database Systems*, 1724–1727. Boston, MA: Springer US. DOI: *https://doi.org/10.1007/978-0-387-39940-9_907*

Reid, N, Ballsun-Stanton, B, White, AJR, Sobotkova, A and **Klump, JF.** 2016. A mobile app for geological/geochemical field data acquisition. In: *35th International Geological Congress. Presented at the 35th International Geological Congress*. Cape Town, South Africa.

Royal Society. 2012. Science as an open enterprise (No. 02/12). London, United Kingdom: The Royal Society. Available at *http://royalsociety.org/policy/projects/science-public-enterprise/report/.*

Sen, M and **Duffy, T.** 2005. GeoSciML: Development of a generic GeoScience Markup Language. *Computers & Geosciences*, 31(9): 1095–1103. DOI: *https://doi.org/10.1016/j.cageo.2004.12.003*

Thessen, AE, Woodburn, M, Koureas, D, Paul, D, Conlon, M, Shorthouse, DP and **Ramdeen, S.** 2019. Proper Attribution for Curation and Maintenance of Research Collections: Metadata Recommendations of the RDA/TDWG Working Group. *Data Science Journal*, 18(1): 54. DOI: *https://doi.org/10.5334/dsj-2019-054*

TIB Hannover. 2012. TIB DOI Registration. Hannover, Germany: TIB Hannover. Available at *http://www.tib-hannover.de/en/services/doi-service/doi-registration/.*

Walls, RL, Deck, J, Guralnick, R, Baskauf, S, Beaman, R, Blum, S, Bowers, S, Buttigieg, PL, Davies, N, Endresen, D, Gandolfo, MA, Hanner, R, Janning, A, Krishtalka, L, Matsunaga, A, Midford, P, Morrison, N, Tuama, ÉÓ, Schildhauer, M, Smith, B, Stucky, BJ, Thomer, A, Wieczorek, J, Whitacre, J and **Wooley, J.** 2014. Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. *PLOS ONE*, 9(3): e89606. DOI: *https://doi.org/10.1371/journal.pone.0089606*

Wieczorek, J, Bloom, D, Guralnick, R, Blum, S, Döring, M, Giovanni, R, Robertson, T and **Vieglais, D.** 2012. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLOS ONE*, 7(1): e29715. DOI: *https://doi.org/10.1371/journal.pone.0029715*

TO CITE THIS ARTICLE:
Klump, J, Lehnert, K, Ulbricht, D, Devaraju, A, Elger, K, Fleischer, D, Ramdeen, S and Wyborn, L. 2021. Towards Globally Unique Identification of Physical Samples: Governance and Technical Implementation of the IGSN Global Sample Number. *Data Science Journal*, 20: 33, pp. 1–16. DOI: *https://doi.org/10.5334/dsj-2021-033*

Submitted: 04 June 2021
Accepted: 14 October 2021
Published: 28 October 2021

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.