Scaling Identifiers and their Metadata to Gigascale: An Architecture to Tackle the Challenges of Volume and Variety

RESEARCH PAPER

JENS KLUMP
DOUG FILS
ANUSURIYA DEVARAJU
SARAH RAMDEEN
JESS ROBERTSON
LESLEY WYBORN
KERSTIN LEHNERT

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

ABSTRACT

Persistent identifiers are applied to an ever-increasing variety of research objects, including software, samples, models, people, instruments, grants, and projects, and there is a growing need to apply identifiers at a finer and finer granularity. Unfortunately, the systems developed over two decades ago to manage identifiers and the metadata describing the identified objects no longer scale. Communities working with physical samples have grappled with these three challenges of the increasing volume, variety, and variability of identified objects for many years. To address this dual challenge, the IGSN 2040 project explored how metadata and catalogues for physical samples could be shared at the scale of billions of samples across an ever-growing variety of users and disciplines. In this paper, we focus on how we scale identifiers and their describing metadata to billions of objects and who the actors involved with this system are. Our analysis of these requirements resulted in the definition of a minimum viable product and the design of an architecture that not only addresses the challenges of increasing volume and variety but, more importantly, is easy to implement because it reuses commonly used Web components. Our solution is based on a Web architectural model that utilises Schema.org, JSON-LD, and sitemaps. Applying these commonly used architectural patterns on the internet allows us to not only handle increasing variety but also enable better compliance with the FAIR Guiding Principles.

CORRESPONDING AUTHOR:

Jens Klump

Mineral Resources, CSIRO, Perth, WA, Australia

jens.klump@csiro.au

KEYWORDS:

persistent identifiers; metadata; web architecture; linked open data

TO CITE THIS ARTICLE:

Klump, J, Fils, D, Devaraju, A, Ramdeen, S Robertson, J, Wyborn, L and Lehnert, K. 2023. Scaling Identifiers and their Metadata to Gigascale: An Architecture to Tackle the Challenges of Volume and Variety. *Data Science Journal*, 22: 5, pp. 1–17. DOI: https://doi.org/10.5334/dsj-2023-005

1. INTRODUCTION AND MOTIVATION

The number of research artefacts that are identified through Persistent Identifiers (PID) and catalogued is rising rapidly. PIDs support a globally unique and unambiguous digital identification of resources on the Web. They have been around for over two decades (Klump & Huber 2017) and are now recognised as essential components of global research data ecosystems (Cousijn et al. 2021). Initially designed for scholarly literature and related resources, their application has expanded to various components of the scholarly ecosystem, including datasets, physical samples,[1] software, instruments, organisations, people, grants, and projects. Increasingly, identifiers are being used to track and link the research artefacts, starting with creating a digital resource (e.g., a dataset) through various stages of the research process - analysis, generation of further data, images, publication, and ultimately, its curation and preservation. In addition, there is a growing tendency to apply identifiers at a finer granularity and to use identifiers to link assets, such as through parent-child relationships. For example, not only can a collection of analyses have an identifier, its datasets and individual analyses within a dataset can be uniquely identified.

Physical samples play a key role in many disciplines, and often they serve as anchors in the physical world for data and its interpretation. To make these artefacts findable and accessible and to make their related data interoperable and reusable requires that this information is made available on the internet. PIDs for samples serve as this anchor. For example, physical samples can be linked to the virtual world by their digital representations (e.g., Lannom et al. 2019) and can be referenced by globally unique Web-resolvable persistent identifiers (e.g., Hardisty et al. 2021; Klump et al. 2021). The growing number of PIDs for samples, their description in online catalogues, and the required interoperability between these systems introduce three challenges: very large numbers of samples present the challenge of volume, while the many different use cases and ways to describe samples introduce the challenges of variety and variability. These dimensions were recognised early on as key challenges in the processing of large data volumes (Laney 2001). Other authors defined additional Big Data dimensions such as veracity and velocity (see e.g., Suresh 2014). However, velocity and veracity are not primary challenges faced by persistent identifier systems and federated catalogues. Therefore, in this paper, we will only discuss how we address the challenges of variety, variability, and volume.

For this paper, we will use physical samples and the International Generic Sample Number (IGSN) as an example of how the scaling problem in identifiers and metadata can be addressed. The IGSN was originally developed as the International Geo Sample Number to address the need for unambiguous identification of samples in geochemical databases and synoptic studies (Klump et al. 2021; Lehnert et al. 2004) but has since expanded in scope and is now used across many other disciplines. This paper will not discuss the IGSN in detail but will refer to Klump et al. (2021), where the technical details of the IGSN identifier, its underlying technology, and its governance are discussed.

To deal with the potential expansion of scale, the IGSN 2040 project (Lehnert et al. 2021) conducted a workshop in Canberra (Australia) to explore how metadata and sample catalogues could be shared at the scale of billions of samples (Klump et al. 2020). The aim of this workshop was to design a future system architecture for the minting and management of IGSN PIDs that would be able to scale to large numbers of samples, adapt to ever-changing applications for sample descriptions, and offer more vectors for discovery and services based on distributed sample catalogues. The outcome of the design process and discussions was a Web architecture-based approach that can accommodate both very large numbers of samples as well as thematic breadth across disciplines and use cases. Even though the IGSN technical development progressed in a different direction (Buys & Lehnert 2021), our proposal for a system architecture still has a broad application and is transferable to other research data infrastructures and PID systems.

The adoption of a new technology by a community depends heavily on how easy the technology can be adopted. The demise of the Life Science Identifier (LSID) showed that a technology

1 Different communities have their specific definitions of the terms 'specimen' or 'sample' to describe the sometimes complex relationships between a physical object of study and the 'thing' it specifically represents (see e.g., Haller et al. 2019). In this paper, we will use the term 'sample' for readability and follow its more common use in the geosciences community.

stack that is difficult to implement will ultimately fail to be adopted by the community (Klump & Huber 2017).

1.1 THE CHALLENGE OF VOLUME

In research areas like geology, biology, or archaeology, it is common that thousands of samples are analysed before the data are then compiled into a dataset and described in a publication, which in turn, can then be aggregated in synthesis studies. It is, therefore, a reasonable assumption that publications and data are underpinned by orders of magnitude larger numbers of samples (e.g., Hardisty et al. 2021; Lannom et al. 2019). There is growing evidence that the scale at which identifiers are being applied in the research ecosystem is rapidly moving from the millions to billions, that is, Gigascale. The growing adoption of PIDs for individual samples across disciplines made many aware of the need to accommodate very large numbers of identifiers. In their scoping report, the Distributed System of Scientific Collections (DiSSCo) consortium estimates that natural history collections hold approximately 1.5 billion objects (Hardisty et al. 2021). This growth in numbers would put considerable strain on current architectures for minting, managing, publishing, finding, and accessing identifiers, and most existing systems would not be able to scale.

The increase in volume could be accommodated by adopting scalable cloud-based infrastructures for minting and managing identifiers. However, tools commonly used for harvesting metadata associated with identifiers have not been able to scale, and experience with IGSN has already shown that most common implementations of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, Van de Sompel et al. 2004) become prone to failures once the number of items harvested from catalogues goes into millions of items.

1.2 THE CHALLENGES OF VARIETY AND VARIABILITY

Adoption of PIDs across research communities introduces the challenge that the increasing diversity of use cases cannot be accommodated by unified metadata schemas that attempt to generalise across use cases but need to be applicable in diverse communities (e.g., Damerow et al. 2021; Davies et al. 2021; Genova et al. 2017). This diversity in metadata needs to be reconciled with the Findable, Accessible, Interoperable, Reusable (FAIR) Principles of Wilkinson et al. (2016). To be able to scale to large numbers, the metadata has to be made machine-actionable (e.g., De Smedt et al. 2020; Devaraju & Huber 2021; Ganske et al. 2020). The expansion of scope, combined with the large increase in the number of samples, requires an approach to metadata that can accommodate many different applications from a wide range of disciplines.

Building consensus on metadata is hard work and expensive (Genova et al. 2017), and metadata, like any language, is a social process driven by context and purpose (Eco 1997; Parsons et al. 2022). In many systems, the definition of schemas for the description of samples is too narrowly focused on one use case. A single common schema cannot convey a meaningful description of a sample without running the risk of being either a large but sparsely populated collection of loosely defined elements or a small set of metadata unable to hold much information about the object it describes. This situation is a common challenge in Web-based applications and has therefore been addressed by search engine operators and others (Guha 2011; Noy & Brickley 2017).

In this paper, we propose an approach to metadata that is prescriptive with respect to protocols but at the same time open to new applications by prescribing only a high-level information model that uses well-known Web architecture patterns to connect between systems. This approach allows us to accommodate a greater heterogeneity and variety of metadata than in the prescriptive approaches common in many research data infrastructures today, which often end up being restrictive in the number of communities that can apply them.

We, furthermore, describe a governance framework that is designed to help user communities develop and codify their expectations towards metadata structures that allow them to build information aggregators and other value-added services to serve their use cases.

The paper is organised as follows: Section 2 specifies a set of design principles to guide our technological choices. In Section 3, using use cases for the application of IGSN as an exemplar,

we define the actors in the IGSN system – so-called personas – to explore their interactions with the IGSN system in the context of the use cases. From this, in Section 4, we determine a Minimum Viable Product of the system architecture that allows us to guide and prioritise the development of its components. Section 5 discusses our design choices and the broader implications of our proposal.

2. GUIDING PRINCIPLES

Before we defined the architecture of the envisaged infrastructure for the IGSN 2040 project, we defined a set of design principles based on a high-level design manifesto to guide our technical choices. The aim was to build a system that was easily extensible, scalable, sustainable, and flexible to accommodate the requirements of new science communities for their sample-based research.

We defined the following statements in our design manifesto following the example of Ross et al. (2020):

- Our domains deserve research-specific software.
- Diverse practices and limited resources require generalised software.
- Do one thing well with modular and federated software (but slice the pie thoughtfully).
- Open-source software supports open research and has other advantages (but is difficult to maintain).
- Scope requirements carefully.
- Invest in outreach and engagement.

To support diverse practices, the technical architecture has to be general in nature, which is best achieved by adopting common practices for sharing metadata on the web.

For many stakeholders, sample lifecycles are only one consideration amongst many. Thus IGSN needs to fit into existing technology stacks, and the development of bespoke new toolsets should be avoided. Additionally, the broad spectrum of actors in the samples community means the ability to scale is vital – it should be easy to get started by adopting the deep thinking into sample-based science that IGSN already offers while also making it possible to do something new that is relevant to a given community without holding them back in the future.

2.1 WEB ARCHITECTURE

Web architecture implementations are mature and optimised for large-scale interactions to support the scale of sample data expected. Additionally, the architecture can be equally implemented on commodity hardware or in dynamically scalable cloud infrastructure.

Enabling a semantic network on Web architecture brings key capabilities for addressing large-scale sample identifier services. Web architecture approaches have demonstrated the ability to scale to billions of resources and requests. Additionally, this capacity is a commodity and one that can be dynamically scaled in response to changing needs. Given the limited control over federated resources on the internet, error handling is an integral aspect of the Web architecture to allow the graceful handling of errors and redirection and forwarding of these events.

To enable a more adaptive approach to metadata, the proposed architecture uses Schema.org (https://schema.org/) together with JavaScript Object Notation for Linked Data (JSON-LD, https://json-ld.org/). Schema.org is a structured markup that standardises HTML tags for creating rich search engine results. JSON-LD is a lightweight linked data format and is a representation of the RDF data model, which does not require the same high degree of syntactic and semantic standardisation that is required in XML-based technologies and does not require establishing a consensus across the entire system. It allows the inclusion of a wide range of already published expressive semantic vocabularies into the Web architecture approach described above. This approach is similar to how internet search works without establishing a semantic consensus across the entire internet and thus enables both scale and reliability of the system.

Early on, the major internet search engine operators realised that Web crawlers were not the most efficient way of indexing the entire internet and that they often missed indexing the content of dynamically generated pages. To make it easier for search engines to find relevant content, the major search engine operators introduced the sitemaps protocol (sitemaps.org 2006). The Sitemaps protocol allows a Webmaster to inform search engines about URLs on a website that are available for crawling. While the sitemaps protocol has been successful for search engine operators, only a few projects or eResearch infrastructures have used this protocol. Examples of the use of sitemaps to aggregate research data are the Environmental Data Initiative (EDI) (Servilla et al. 2018) and Biositemaps (Dinov et al. 2008), although both services seem to operate no longer.

Despite the limited uptake of sitemaps.org among research data infrastructures, we propose that the sitemaps.org protocol can be used by information aggregators serving specific communities and their content providers. In this scenario, the PID registrar would be the curator of a list of domain-specific sitemaps that would allow information aggregators to build domain-specific search engines.

2.2 CONTEXT AND VOCABULARY

Whereas Web architecture brings scale and resilience, the semantic network brings context. This context provides an understanding of the meaning and network of connections and properties in human and machine-readable ways. Such context enables richer search and semantic relations to be provided. The context maps terms to Web-resolvable, and hopefully machine-readable, references.

As a foundation, these references may leverage Schema.org as a lightweight and broad upper-level vocabulary to help address general discovery and upper-level concepts (e.g., Michel & The Bioschemas Community 2018) and contribute to making samples findable. Complementing this is the ease with which this approach can be extended with research discipline or application-specific vocabularies to provide richer context to the samples to aid in addressing the remaining FAIR Guiding Principles such as access, interoperability, and reuse. This extensibility is a design feature of Schema.org.

This result is a common set of minimal registration metadata that is ubiquitous across the range of focused community terms and concepts. This facilitates simpler publishing and innovation to address community needs by defining a required core that can be extended by various communities of practice. This results in a clearer onboarding process for agents, science communities, and facilities, connecting them with a broader set of use cases. Communities that have already invested in the development of semantic types can preserve and leverage that work through community-specific extensions of the Schema.org core (e.g., Thessen et al. 2019).

Linked data represents accessible and connected data and combining this with relationships among the data facilitates a linked data Web as envisioned by Tim Berners-Lee (2009). Further, structured data on the Web and data on the Web best practices provide connections with Schema.org and other Web architecture-based linked data and semantic Web patterns. This union of semantic networks and Web architecture is a key principle of the proposed architecture.

2.3 GOVERNANCE

The governance of a (sample) PID system covers three main components and their interactions to ensure the effective management and use of the PIDs: (1) the roles and the actions of the entities involved, (2) the practices they must adhere to, and (3) the technology underpinning the processes.

An architecture that leverages multiple entities built on Web architecture needs strong governance to enable coordinated and efficient operation among the various communities and maintain interoperability. Where standards and projects guide the technical implementation of a PID-based sample system architecture, its governance focuses on guiding principles, best practices around the principles and community requirements that are fed back to the system

implementers. Specifically, things such as alignment to principles and functional purposes (e.g., requirements of the PID communities) and publishing guidance (e.g., identifier practice standards) are examples of value-adding activities provided by governance.

Guiding principles such as FAIR and Linked Open Data (LOD) are programmatic foundations, but the interfaces and functional needs of communities are equally important. These highly focused and practical needs represent the communities' goals. With guidance, these too can be expressed as principles, decoupled from the technical implementation.

Guiding the publishers of sample information to express and disseminate the sample records in a standardised way helps ensure their resources are connected to the community's desired goals. Another important use of this is to bridge the representation of data and the services and interfaces expressed by aggregators and employed by the sample community.

On a technical level, critical elements of a Web architecture approach such as the resolution of identifiers to resources, curation of sitemaps, core schema models, and patterns for community extensions to those core models are examples of crucial governance activities. Conversely, approaches to indexing or the development of interfaces to such indexes may vary based on community or sub-community needs and should be enabled by the architecture.

As part of the governance, the IGSN must also ensure its agents are accountable as information providers. Agents must ensure and maintain the quality and accuracy of the continued access to and adherence to standard procedures for PID metadata.

2.4 OUTREACH AND ENGAGEMENT

Sample communities are diverse, and therefore community engagement is vital to foster the adoption of new architecture and its services. Mediating and bridging between broad cross-cutting principles and focused community principles is an essential value-adding service of governance. Consensus-based decision-making and coordination are required to get the existing sample communities and potentially interested parties to transition into the new architecture. Like other federated information infrastructures, IGSN is a socio-technical system, and its success relies on the implementers of its components. A system that is bespoke and difficult to implement will struggle to find adoption. Our hypothesis was that using standard Web architecture components would make the implementation of the system easier for the stakeholders involved. To evaluate this hypothesis, the IGSN 2040 project organised a code sprint that was conducted in June of 2020 led by the authors of this paper together with members of the IGSN community, who between them allocated persistent sample identifiers (IGSN IDs), maintain a registry of these identifiers and aggregated information into federated catalogues (Fils et al. 2020). The code sprint was designed to build a rapid prototype to test the proposed Web architecture approach and structured data on the Web patterns using synthetic data, test the scalability of the system and uncover barriers to the adoption of the proposed technology stack.

The sprint also allowed the exploration of potential use cases and interfaces for harvested data. Outcomes of the sprint were a draft publication guidance documentation, a description metadata schema, and examples to support expressing sample metadata (Fils et al. 2020). Additional documentation of the evaluated architecture can be found in GitHub repositories associated with this sprint (Fils 2021; Robertson et al. 2020; Schindler & Devaraju 2020).

3. USER CASES AND PERSONAS: WHO IS IN THE GAME AND WHY

The design of the proposed system architecture is based on the analysis of services conducted during the 2019 IGSN 2040 workshop in Canberra, Australia (Klump et al. 2020). After defining the design principles for the future IGSN system architecture, workshop attendees identified the personas involved in the IGSN ecosystem. Next, they developed three workflows for use of IGSN. Finally, the group identified the services needed to support these workflows and mapped the personas to them. The personas and their interactions with the system architecture are summarised in Figure 1.

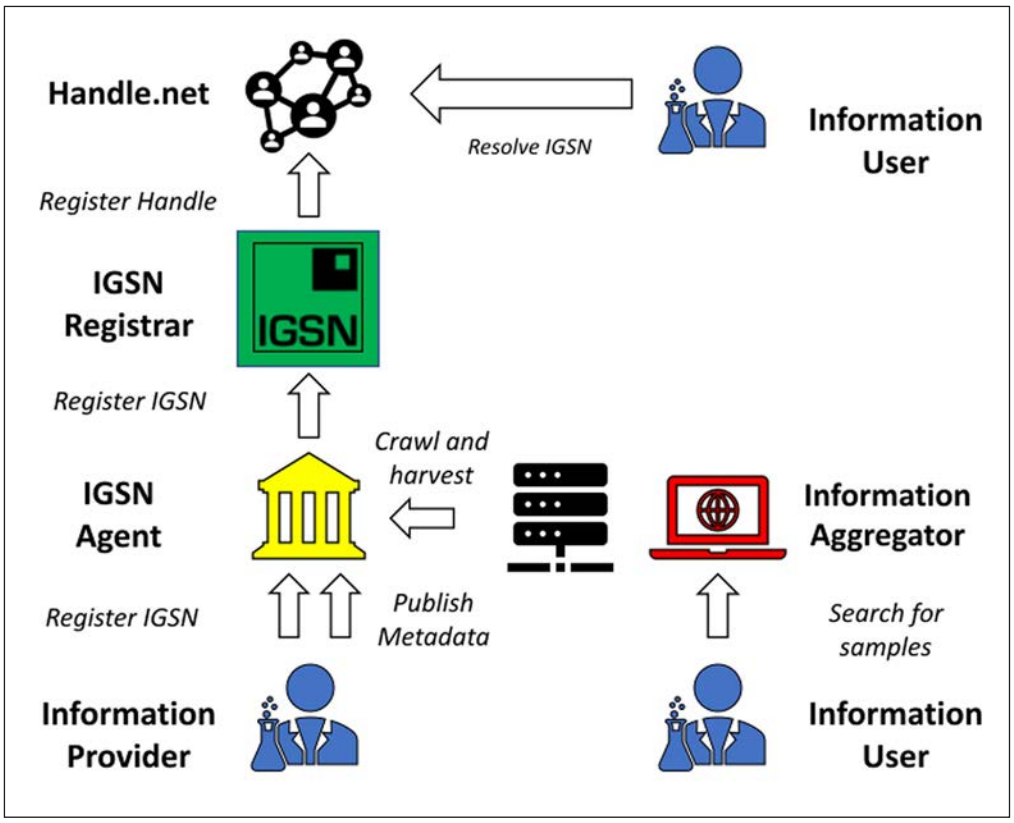

Figure 1 Schematic overview of the proposed system architecture based on Web architecture design principles. Colours used for personas are IGSN Registrar (green), IGSN Agent (yellow), information provider/user (blue), and information aggregator (red).

In this section, we discuss the concept of personas and describe the various personas involved in the proposed IGSN system. We provide two exemplary use case workflows: (1) linking samples to publications and (2) obtaining samples for reuse. We will describe the actors involved in these two use cases through the concept of personas.

3.1 PERSONAS

To model the users and their interaction with the system, we used the concept of 'personas'. This concept is used in user experience design to represent the goals and behaviour of a group of users, rather than an individual, and describes the role a person or organisation may take (Lidwell et al. 2010). Personas help guide the exploration of the context of behaviours and relationships between actors. The primary personas involved in the two use cases presented in this paper are IGSN Registrar, IGSN Agent, Information Aggregator, Information Provider, and Information User.

Any entity can interact with the system in multiple roles and can thus be expressed through more than one persona. We will consider the personas individually based on their interactions with the PID system.

a. IGSN Registrar

The IGSN Registrar represents two roles, one technical and the other in governance. The IGSN e.V., in partnership with DataCite (Buys & Lehnert 2021), fills this technical role by (a) operating the service for minting and resolving IGSN PIDs and (b) providing the central clearing point for linking IGSN Agents with other personas who want to use the data and samples. In the proposed system, the IGSN Registrar would maintain the sitemaps used by the Information Aggregators. The governance role includes mediating decisions for determining protocols and schemas for communication between IGSN Agents and Information Aggregators. The IGSN e.V. carries out the social aspect of coordinating the community in adhering to standards and best practices for sample publications and meeting functional requirements.

b. IGSN Agent

The IGSN Agent (e.g., institutions, agencies) provides IGSN registration services to Information Providers for samples and sample collections. To register an IGSN, Information Providers submit metadata to an IGSN Agent, which includes the required IGSN metadata kernel and community-developed descriptive metadata specific to the IGSN Agent. The IGSN Agents develop and

maintain controlled vocabularies and schemas used to validate the provided metadata. Once the submission and validation processes are complete, the IGSN Agent submits the registration metadata to the IGSN Registrar to mint the IGSN. The IGSN Agents also maintain the landing pages (Web pages to which the IGSN resolves), which contain the sample metadata in human and machine-readable formats and register their sitemaps with the IGSN Registrar. In the proposed architecture, these pages will be embedded with JSON and will be used alongside sitemaps to support harvesting by Information Aggregators (this is currently supported by OAI-PMH). IGSN Agents also provide administrative services to the Information Provider persona, which may include authentication and authorization for metadata management, processing namespace requests, and transfer of custodianship of registered metadata.

c. Information Aggregator

The Information Aggregator is both a consumer and secondary service provider. They firstly harvest sample metadata from one or more IGSN Agents to compile the structured metadata provided by them into an index (i.e., catalogue). Secondly, these indices are then leveraged to provide value-added services such as spatial indexes, clustering, or other. Many of these value-added services may be implemented to address the goals or functional needs of a particular community of practice. Information Aggregators interact with individual IGSN Agents to access their schemas and vocabularies. They connect with the IGSN Registrar to access sitemaps or validate an IGSN. Information Aggregators may be a third party, or IGSN Agents may also act in the role of an Information Aggregator.

d. Information Provider

Information Providers are the originators, producers or collectors of samples who document the information (metadata) about the sample, which then supports the ability of others to reuse the sample. Information Providers rely on IGSN Agents for cost-effective, persistent management of this metadata. Motivations include tracking samples through their lifecycle (from field to lab, analysis to publication, etc.), meeting the requirements of a funding agency or publisher, and supporting discoverability and access. Information Providers register sample metadata following the guidelines of the appropriate IGSN Agent for their community. Information Providers maintain the metadata over time, for example updating information to the IGSN Agent about the sample or adding DOIs for derived datasets. Information providers may also contribute to discussions on vocabularies and metadata schema requirements within their IGSN Agent community. Information Providers may also act as Information Users.

e. Information User

Information Users are looking for samples or information about samples. They may visit an Information Aggregator to search and discover samples based on criteria such as a specific location or sample type. Information Users may obtain an IGSN off a sample container or from a publication (generated by an Information Provider) and resolve the IGSN to evaluate the metadata in order to reuse the sample or verify information about it. Resolving the IGSN connects the Information User to the IGSN Registrar, leading to the landing page maintained by the IGSN Agent. The Information User may utilise resources provided by the IGSN Agent to understand the vocabularies and schema used on the landing page. Information Users may also become Information Providers by generating additional information which updates or augments existing sample information.

3.2 USER STORIES AND WORKFLOWS

Having defined the personas, we now explain how they interact in two example workflows:

1) The 'sampling to publication' workflow

The workflow 'sampling to publication' (Figure 2) begins with a researcher (Information Provider) collecting samples to address his research questions. The samples are described, and the metadata are registered with an IGSN Agent. The IGSN Agent is responsible for registering the samples with IGSN identifiers through the registration service supported by the IGSN Registrar, which includes providing to the IGSN Registrar the minimum metadata required to register a sample and obtain an IGSN and making the complete, descriptive metadata records submitted by the Information Provider about the samples available for harvesting to

Information Aggregators. The researcher drafts a manuscript that contains the interpretations of data and observations based on their samples. The researcher includes the IGSN identifiers of the samples in his publication. The identifiers are resolvable to the metadata landing pages maintained by the IGSN Agent. The reviewers of the manuscript (Information Users) verify the IGSN identifiers for the samples (manually or through an automated process). As part of the scientific review, they may also obtain richer descriptions of the samples from the metadata landing pages. Once the review is complete, the manuscript is published.

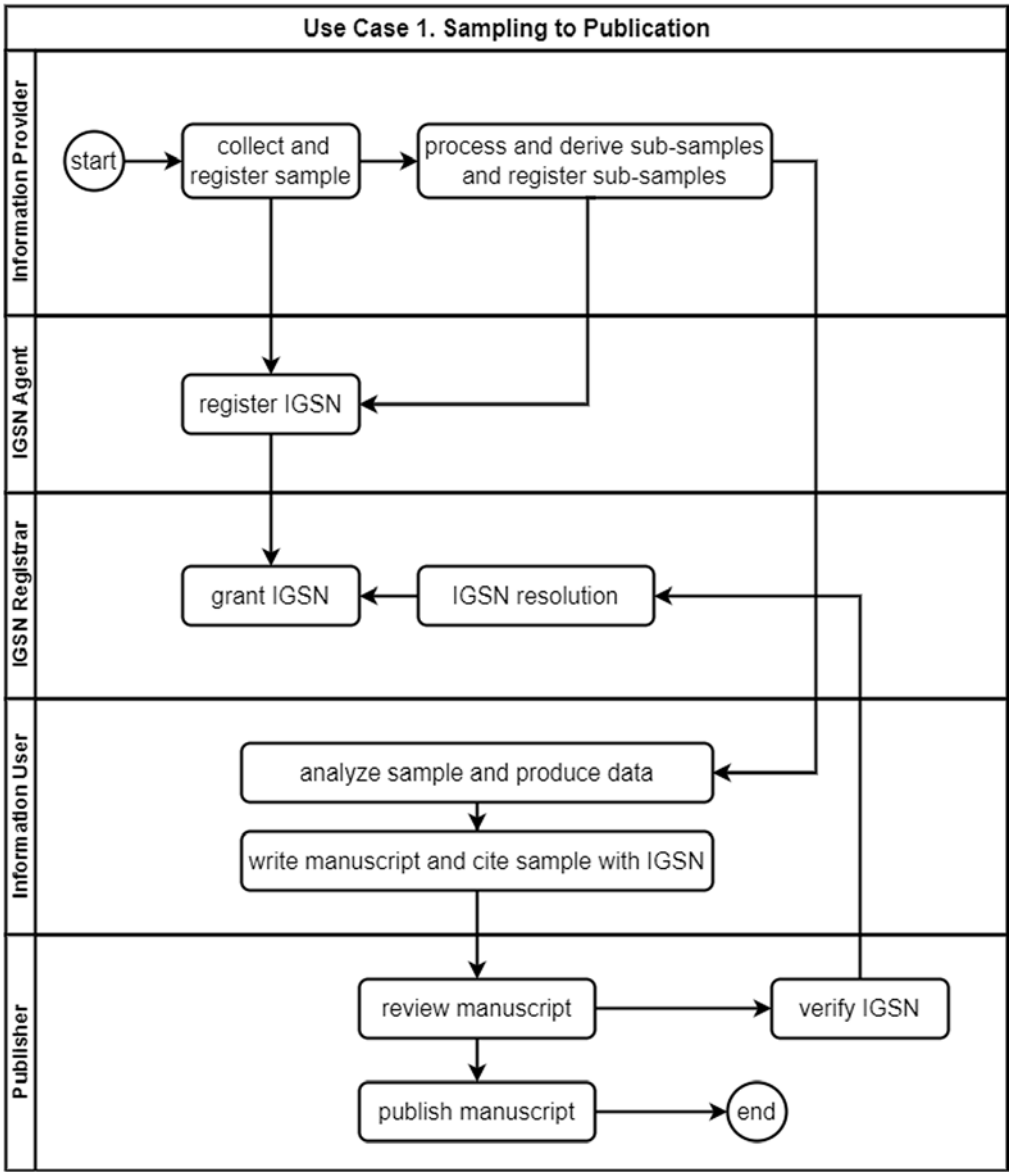

Figure 2 Workflow of linking samples to a publication.

2) The 'sample for reuse' workflow

In the workflow for obtaining a 'sample for reuse' (Figure 3), an entity, acting as an Information Aggregator, develops and hosts a sample catalogue by harvesting metadata from IGSN Agent(s). A researcher (Information User) then searches the catalogue or a data portal to discover samples that may be relevant to their work. They evaluate the metadata profiles created by the Information Providers to determine if the samples may fit their research needs. Once they identify the sample(s) of interest, they use the information provided in the record to contact the custodian of the sample and request the sample or a subsample for further studies.

4. RESULTS: IGSN SERVICES AND MINIMUM VIABLE PRODUCT

We used the personas and interactions described in the workflows above to define the essential elements of the IGSN ecosystem (Figure 4) that are needed to mint identifiers; enable the discovery of samples; disseminate metadata that aid the interpretation and potential reuse of the samples; and govern the IGSN system. Furthermore, we derived the core services necessary for IGSN to be able to serve its purpose, that is, the minimum viable product (MVP). The MVP is summarised in Figure 4 with the solid boxes representing the components of the MVP while the outlined boxes represent elements that can be added at a later stage.

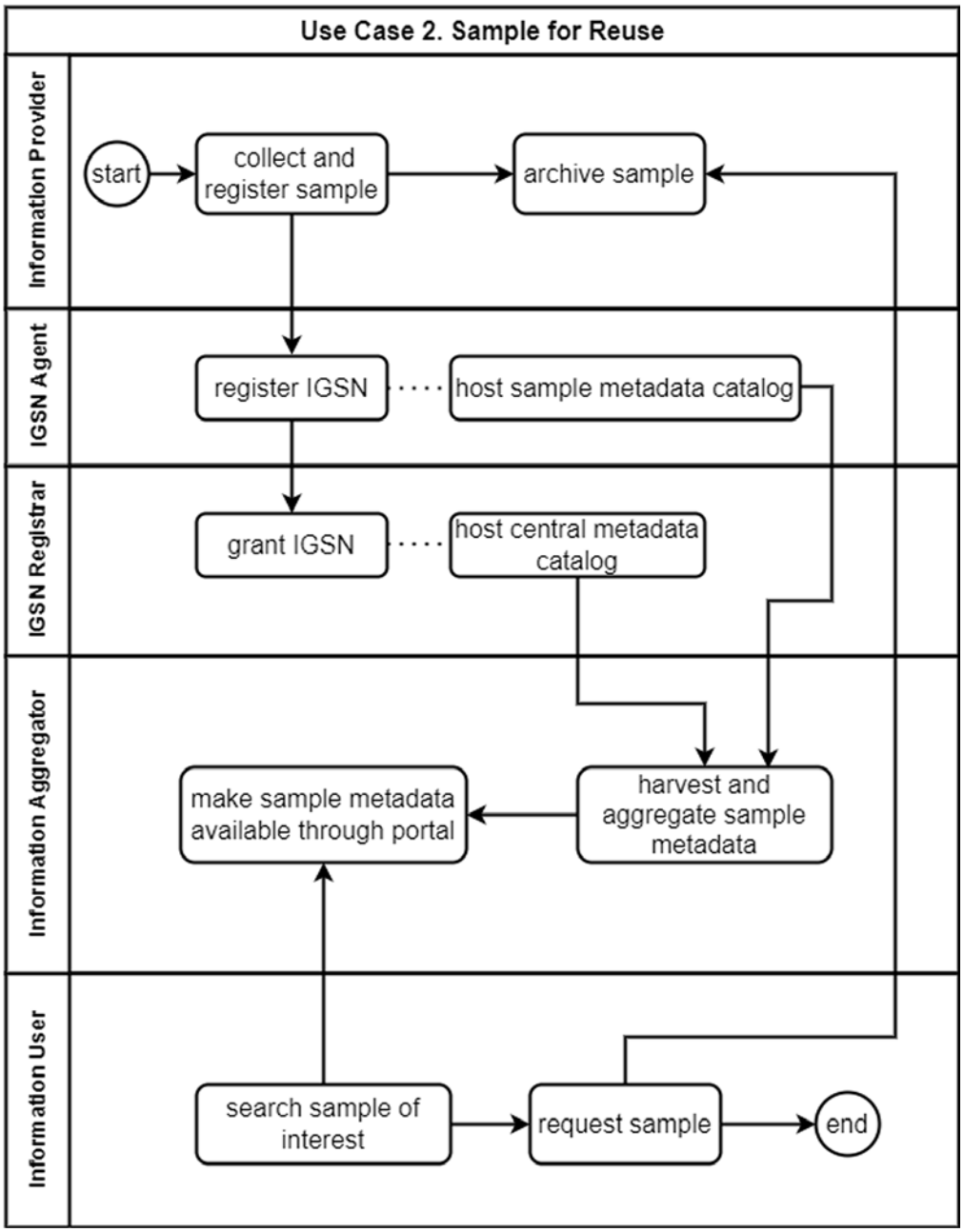

Figure 3 Workflow of reusing samples.

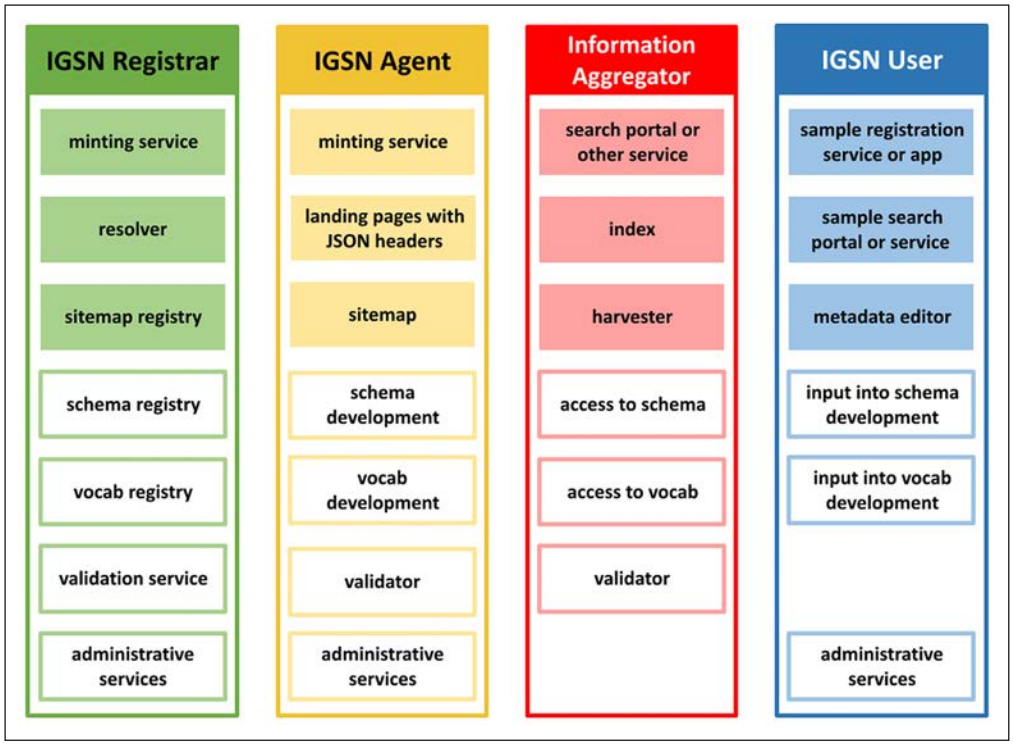

Figure 4 Elements of the IGSN ecosystem. Solid boxes are elements that are needed as a minimum for the system to function (minimum viable product, MVP). To simplify the figure, we portrayed the IGSN user to include the Information User and the Information Provider as one persona, assuming that researchers will be users and providers of information about a sample. The colours used in this figure correspond to those in Figure 1.

The system can be split into three groups of services:

- PID minting and resolution;
- discovery, reuse and governance; and
- administration.

PID minting and resolution continue to be based on the Handle System, either directly, through DOI or another Handle-based identifier system. Metadata syndication and aggregation use the metadata embedded in the landing pages based on JSON-LD and Schema.org. Harvesters are directed to the relevant Web resources through sitemaps, where the IGSN Registrar acts as a registry of sitemaps. These sitemaps can be tailored to be community or use case specific. The IGSN Registrar also hosts registries of the schemas and vocabularies used in the Schema.org and JSON-LD encoded metadata, respectively.

In the federated architecture outlined in the introduction, the role of the IGSN registrar becomes more that of a coordinator than of a regulator. In this scenario, IGSN does not maintain a central catalogue of all registered samples, but only mints and administers the identifiers. Instead of maintaining a central catalogue, the IGSN registrar acts as a reliable pointer to resources, such as a curated list of sitemaps, that can then be harvested by Information Aggregators.

5. DISCUSSION

IGSN is not the only system facing the challenge of accessing and indexing large and diverse metadata catalogues. Similar challenges are faced by trans-disciplinary systems like the DiSSCo (Hardisty et al. 2021), DataCite (Neumann & Brase 2014), Bioschemas (Michel & The Bioschemas Community 2018), and the European Open Science Cloud (EOSC) (Schwardmann et al. 2021). The challenges can be mapped to the dimensions scale, heterogeneity, and adoption.

With the proliferation of physical samples collected for various studies worldwide, it became apparent that discovering them over the Web was going to be challenging. One way of making them findable is through the syndication of metadata and the compilation of catalogues. Syndication of sample catalogues for indexing in federated portals had been a feature of the original IGSN system architecture (Klump et al. 2021). In its original design, IGSN used OAI-PMH because the protocol allows the syndication of metadata catalogues in more than one metadata schema (Van de Sompel et al. 2004). At the time when it was conceived, OAI-PMH was an elegant, lean, and fault-tolerant protocol for exchanging catalogues of several thousand objects. For the sharing of metadata and catalogues, XML is fairly verbose, and its mark-up adds significant bulk to the data payload, which is manageable with catalogues containing thousands to tens of thousands of entries but becomes a significant burden once catalogues scale to millions or billions of entries. Experience in the operation of the IGSN system has already shown that standard implementations of OAI-PMH become prone to failures once the number of harvested records goes into millions of items (Klump et al. 2021).

While the self-describing nature of XML seems to be an elegant solution, it turns out to be less so in practical application. The enforcement of schema constraints makes XML-based applications rigid in their information model and prone to runtime errors caused by syntactical inaccuracies (Barkstrom 2010). The simplifications introduced by JSON-LD and Schema.org make implementations easier without losing semantic richness.

Harvesting metadata is not unique to research applications but is also a core function of search engines. Indexing the internet at large led to the development of lightweight encodings based on JSON-LD. In early 2017, Google unveiled Google Dataset Search (Noy & Brickley 2017). Google Dataset Search builds on Schema.org, an initiative by Google, Bing, and Yahoo! to create and support a common vocabulary for structured data markup on Web pages (Guha 2011).

In 2018 the US National Science Foundation (NSF) EarthCube Science Support Office (ESSO) at UCAR, through pilot work funded by the office, developed a test leveraging Schema.org type Dataset and structured data on the Web to address issues of FAIR data for geoscience (Lingerfelt et al. 2018). From this evolved the ESIP Schema.org Cluster, which develops and publishes guidance for leveraging this approach for datasets (Jones et al. 2021). Additionally, EarthCube is extending this work as GeoCODES (https://geocodes.earthcube.org/) to develop

and provide a discovery portal and associated knowledge graph for the NSF geoscience data facilities.

A similar approach was taken by the Bioschemas community (Michel & The Bioschemas Community 2018). They noted that Schema.org is well suited for describing data repositories but lacks terms to describe other research entities, such as biological taxa, proteins, genes, and reagents, that are of interest to biosciences. An additional challenge was that metadata records were distributed across several repositories. This information is required for aggregators to harvest, parse, and index very large volumes of metadata (Thessen et al. 2018).

Scaling metadata to large numbers may appear as the primary challenge, but when we look at the use cases described in the section above, most use cases will not need to index the entire catalogue of samples on the internet. A scalable architecture must allow information aggregators to select only a subset specific to their application. Decentralising the architecture of cataloguing samples on the internet makes it easier to structure the system components into modules that can be combined and adapted as needed in a specific application. The MVP outlined earlier describes the minimum components and functionality needed for building applications based on distributed sample catalogues.

The other challenge is the thematic breadth of samples on the internet and the many ways to describe them in diverse and evolving use cases. The use of structured data on the Web provides a means to leverage web architecture approaches in the samples community. This approach has many benefits: it leverages existing developer tools, skills, and experiences and can be implemented on existing server architecture. Furthermore, it builds on existing vocabulary efforts (e.g., Albertoni et al. 2021; Guha 2011; Haller et al. 2019) and enables connecting samples in a semantic manner to a broader range of resources like institutions, people, and publications.

Landing pages can serve rich and diverse descriptions of samples. To make this content machine-actionable, it is important that communities of practice agree on the content that they expect to be displayed on landing pages. This approach has been pioneered in the community of earth system model data repositories (Ganske et al. 2020).

The above examples show how leveraging web architecture patterns around structured data for the Web gives access to the semantic Web and ways to encode the context around data. This makes building a network across science disciplines far easier as they do not require universal consensus among all stakeholders. In addition, the use of web architecture allows third parties like Google, Bing, DataOne, and others to access and use metadata to provide offerings based on open, well-known architectures. The IGSN 2040 code sprint in 2020 showed that using common Web technologies that are also used in search engines makes this approach flexible and easy to adopt (Fils et al. 2020).

The system architecture outlined in this paper has a broader impact than providing a technical solution to the challenge of indexing billions of objects across heterogeneous and distributed resources. By leveraging existing technology and approaches, a larger community is enabled to engage and make more samples discoverable and usable. This addresses the challenge of accessing very large numbers of records. The simplified architecture makes it easier to develop tools and interfaces, allowing the presentation of samples and their information in a manner aligned with a given community's needs. Furthermore, a simplified architecture aids sustainability from both a technical and financial perspective.

The open architecture of making metadata available as structured data on the Web democratises the publishing of knowledge by enabling more communities to represent their knowledge expressed in a way that is fit for their purpose and not limited by the least common denominator of an overarching standard. The nature of structured data on the web provides the ability to apply semantic context to samples. This means richer discovery and information about resources, their uses in the past and potential future uses becoming more readily available. Semantically enriched descriptions contribute directly to the implementation of the FAIR Guiding Principles (Wilkinson et al. 2016) beyond the aspects of findability (F) and access (A). Providing metadata that is interpretable by machines also makes them more interoperable (I) and makes them available for reuse (R) (Plomp 2020).

6. CONCLUSIONS AND OUTLOOK

The work portrayed in this paper was motivated by the need for scaling persistent identifiers to billions of objects and the means for describing and cataloguing physical samples in a wide range of use cases. The modules needed as the basis of the proposed system architecture were identified by defining the personas and exploring the processes in which the various users of the system interact with each other and the technical infrastructure.

Our analysis of the requirements for the identification and description of objects at the Gigascale showed that the necessary scalability can only be achieved by identifying the concerns of the different personas in the respective processes and analysing which information was needed to enable specific use cases. The wide variety and variability of use cases and discipline-specific profiles also mean that many players hosted in different communities are involved and are best served by a modular rather than a monolithic infrastructure. A decentralised infrastructure, however, requires a governance structure that defines and enforces conventions through community-driven processes. A role for the IGSN Registrar could be the curator of a list of sitemaps of its affiliated information providers (IGSN Agents) for its Information Aggregators.

At this scale, machine readability and fault tolerance are critical. To achieve the necessary balance between flexibility and structure, we looked at technologies that form the backbone of search infrastructures on the internet at large. The patterns developed around Schema.org provide this balance.

An additional advantage of the proposed architecture is that the necessary technology stack is composed entirely of established Web architecture patterns. As we demonstrated in our code sprint, using established Web architecture patterns makes adoption easier and provides an opportunity to leverage this pattern to enable scaling for both publication and harvesting of resources to very large numbers of objects. A structured data on the Web implementation as presented and discussed here is an approach that gives control of data and metadata back to the communities who are both closer to the data and will have more sense of ownership and responsibility for it. This open and easily accessible approach is a way to translate the FAIR Guiding Principles to the samples community for both humans and machines.

DATA ACCESSIBILITY STATEMENT

There is no data cited in this project. All code is available on GitHub at https://github.com/igsn and https://github.com/gleanerio.

ACKNOWLEDGEMENTS

The authors would like to thank the members of the IGSN 2040 Technical Committee and the participants of the IGSN 2040 Technical Sprint for their input in developing the concepts presented in this paper.

We would like to thank Josh Greenburgh from the Alfred P. Sloan Foundation for his interest in and support of the IGSN system.

Key components for our design were developed by the ESIP Schema.org Cluster and EarthCube Project 418.

Work on the IGSN was supported through AuScope and the Australian Research Data Commons (ARDC). AuScope and ARDC are programmes in Australia's National Collaborative Research Infrastructure Strategy (NCRIS).

The authors would also like to thank the reviewers of the Data Science Journal whose comments helped improve this manuscript.

FUNDING INFORMATION

The project, 'Defining the Future of the IGSN as a Global Persistent Identifier for Material Samples' was supported by the Alfred P. Sloan Foundation in the IGSN 2040 project (Grant Agreement G-2018-11137).

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 Project 418 on publishing JSON-LD metadata was supported by EarthCube (NSF award no. 1623751).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Jens Klump was PI on the IGSN 2040 Project, Leader of the Technical Committee and wrote the manuscript.

Doug Fils was part of the IGSN 2040 Technical Committee and contributed to the writing and editing of the manuscript.

Anusuriya Devaraju was part of the IGSN 2040 Technical Committee and contributed to the writing and editing of the manuscript.

Sarah Ramdeen was Project Coordinator on the IGSN 2040 Project and contributed to the writing and editing of the manuscript.

Jess Robertson was part of the IGSN 2040 Technical Committee and contributed to the writing of the manuscript.

Lesley Wyborn was PI on the IGSN 2040 Project and contributed to the writing and editing of the manuscript.

Kerstin Lehnert was PI on the IGSN 2040 Project and 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

Doug Fils orcid.org/0000-0002-2257-9127
Ocean Leadership, Washington, DC, USA

Anusuriya Devaraju orcid.org/0000-0003-0870-3192
Mineral Resources, CSIRO, Perth, WA, Australia; TERN, The University of Queensland, Brisbane, QLD, Australia

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

Jess Robertson orcid.org/0000-0002-4553-9697
Ministry of Business, Innovation and Employment, Wellington, New Zealand

Lesley Wyborn orcid.org/0000-0001-5976-4943
Australian Research Data Commons, Canberra, ACT, Australia

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

REFERENCES

Albertoni, R, Browning, D, Cox, SJD, Gonzalez-Beltran, A, Perego, A and **Winstanley, P.** 2021. Data Catalog Vocabulary (DCAT) - Version 3 (W3C Proposed Recommendation). Cambridge, MA: World Wide Web Consortium (W3C). Available at https://www.w3.org/TR/vocab-dcat-3/.

Barkstrom, BR. 2010. When is it sensible not to use XML? *Earth Science Informatics*, 4: 45–53. DOI: https://doi.org/10.1007/s12145-010-0063-2

Berners-Lee, T. 2009. Linked Data. *W3C Design Issues*. Available at https://www.w3.org/DesignIssues/LinkedData.html [Last accessed 29 October 2021].

Buys, M and **Lehnert, KA.** 2021. Partnership between IGSN and DataCite. *DataCite Blog*. [Last accessed 3 November 2021]. DOI: https://doi.org/10.5438/7z70-1155

Cousijn, H, Braukmann, R, Fenner, M, Ferguson, C, van Horik, R, Lammey, R, Meadows, A and **Lambert, S.** 2021. Connected research: The potential of the PID graph. *Patterns*, 2(1): 1–7. DOI: https://doi.org/10.1016/j.patter.2020.100180

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

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

De Smedt, K, Koureas, D and **Wittenburg, P.** 2020. FAIR digital objects for science: From data pieces to actionable knowledge units. *Publications*, 8(2): 21. DOI: https://doi.org/10.3390/publications8020021

Devaraju, A and **Huber, R.** 2021. An automated solution for measuring the progress toward FAIR research data. *Patterns*, 2(11): 100370. DOI: https://doi.org/10.1016/j.patter.2021.100370

Dinov, ID, Rubin, D, Lorensen, W, Dugan, J, Ma, J, Murphy, S, Kirschner, B, Bug, W, Sherman, M, Floratos, A, Kennedy, D, Jagadish, HV, Schmidt, J, Athey, B, Califano, A, Musen, M, Altman, R, Kikinis, R, Kohane, I, Delp, S, Parker, DS and **Toga, AW.** 2008. iTools: A framework for classification, categorization and integration of computational biology resources. *PLoS ONE*, 3(5): e2265. DOI: https://doi.org/10.1371/journal.pone.0002265

Eco, U. 1997. *The search for the perfect language*. 2nd ed. Oxford, United Kingdom: Blackwell.

Fils, D. 2021. gleanerio/gleaner-compose. Available at https://github.com/gleanerio/gleaner-compose [Last accessed 25 June 2021].

Fils, D, Klump, J and **Robertson, J.** 2020. Connecting data to the physical world: IGSN 2040 sprint outcomes and recommendations (Technical Report). DOI: https://doi.org/10.5281/zenodo.3905364

Ganske, A, Heydebreck, D, Höck, H, Kraft, A, Quaas, J and **Kaiser, A.** 2020. A short guide to increase FAIRness of atmospheric model data. *Meteorologische Zeitschrift*, 29(6): 483–491. DOI: https://doi.org/10.1127/metz/2020/1042

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

Guha, R. 2011. Official Google blog: Introducing schema.org: Search engines come together for a richer Web. *Google Blog*. Available at https://googleblog.blogspot.com/2011/06/introducing-schemaorg-search-engines.html [Last accessed 3 July 2020].

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, Addink, W, Glöckler, F, Güntsch, A, Islam, S and **Weiland, C.** 2021. A choice of persistent identifier schemes for the Distributed System of Scientific Collections (DiSSCo). *Research Ideas and Outcomes*, 7: e67379. DOI: https://doi.org/10.3897/rio.7.e67379

Jones, M, Richard, SM, Vieglais, D, Shepherd, A, Duerr, RE, Fils, D and **McGibbney, LJ.** 2021. Science-on-Schema.org v1.2.0. DOI: https://doi.org/10.5281/zenodo.4477164

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, Ulbricht, D, Devaraju, A, Elger, K, Fleischer, D, Ramdeen, S and **Wyborn, LAI.** 2021. Towards globally unique identification of physical samples: Governance and technical implementation of the IGSN global sample number. *Data Science Journal*, 20(33): 1–16. DOI: https://doi.org/10.5334/dsj-2021-033

Klump, J, Lehnert, K, Wyborn, L and **Ramdeen, S.** 2020. IGSN 2040 Technical Steering Committee Meeting Report. Potsdam, Germany: IGSN e.V. DOI: https://doi.org/10.5281/zenodo.3724683

Laney, D. 2001. 3D Data Management (No. 949). Stamford, CT: META Group. Available at https://web.archive.org/web/20120806062002/http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

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, 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*. *Presented at the AGU Fall Meeting 2004*. San Francisco, CA: American Geophysical Union. pp. SF32A-06. Available at http://adsabs.harvard.edu/abs/2004AGUFMSF32A..06L [Last accessed 10 May 2016].

Lehnert, K, Klump, J, Ramdeen, S, Wyborn, L and **Haak, L.** 2021. IGSN 2040 Summary Report: Defining the Future of the IGSN as a Global Persistent Identifier for Material Samples. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.5118289

Lidwell, W, Holden, K and **Butler, J.** 2010. *Universal Principles of Design, Revised and Updated*. 2nd ed. Beverley, MA: Rockport Publishers. Available at https://learning.oreilly.com/library/view/universal-principles-of/9781592535873/.

Lingerfelt, E, Fils, D and **Shepherd, A.** 2018. Project 418: A Funded Project of the EarthCube Science Support Office. *Presented at the AGU Fall Meeting 2018*. Washington, D.C.: American Geophysical Union. pp. IN31B-22. Available at https://agu.confex.com/agu/fm18/meetingapp.cgi/Paper/442533 [Last accessed 18 January 2022].

Michel, F and **The Bioschemas Community.** 2018. Bioschemas & Schema.org: a Lightweight Semantic Layer for Life Sciences Websites. *Biodiversity Information Science and Standards*, 2: e25836. DOI: https://doi.org/10.3897/biss.2.25836

Neumann, J and **Brase, J.** 2014. DataCite and DOI names for research data. *Journal of Computer-Aided Molecular Design*, 28(10): 1035–1041. DOI: https://doi.org/10.1007/s10822-014-9776-5

Noy, N and **Brickley, D.** 2017. Facilitating the discovery of public datasets. *Google AI Blog*. Available at http://ai.googleblog.com/2017/01/facilitating-discovery-of-public.html [Last accessed 3 March 2020].

Parsons, MA, Duerr, R and **Godøy, Ø.** 2022. The evolution of a geoscience standard: An instructive tale of science keyword development and adoption. *Geoscience Frontiers*, in press: 101400. DOI: https://doi.org/10.1016/j.gsf.2022.101400

Plomp, E. 2020. Going digital: Persistent identifiers for research samples, resources and instruments. *Data Science Journal*, 19(46): 8. DOI: https://doi.org/10.5334/dsj-2020-046

Robertson, JC, Fils, D, Devaraju, A, Song, L, Ramdeen, S and **Klump, J.** 2020. IGSN/igsn-json: Test schema repo for IGSN 2040 Architecture sprint. Available at https://github.com/IGSN/igsn-json [Last accessed 10 November 2022].

Ross, S, Ballsun-Stanton, B, Cassidy, S, Cook, P, Sobotkova, A and **Klump, J.** 2020. FAIMS 3.0: Electronic Field Notebooks. In: *CAAA Digital Archaeology Conference 2020. Presented at the CAA Australasia 2020*. Online: Computer Applications and Quantitative Methods in Archaeology. Available at https://au.caa-international.org/2020-conference-abstracts/.

Schindler, U and **Devaraju, A.** 2020. MARUM DIS IGSN landing page mockup implementation. Available at https://github.com/pangaea-data-publisher/marum-dis-igsn [Last accessed 10 November 2022].

Schwardmann, U, Fenner, M, Hellström, M, Koers, H, L'Hours, H, Matthews, B, Ritz, R, Valle, M, van de Sanden, M and **Zamani, T.** 2021. PID architecture for the EOSC: report from the EOSC Executive Board Working Group (WG) Architecture PID Task Force (TF) (No. KI-03-20-757-EN-N). Luxembourg, L: Directorate-General for Research and Innovation (European Commission). [Last accessed 19 October 2021]. DOI: https://doi.org/10.2777/525581

Servilla, MS, Brunt, J, Costa, D, Gries, C, Grossman-Clarke, S, Hanson, PC, O'Brien, M, Smith, C, Vanderbilt, K and **Waide, R.** 2018. Facilitating data discovery on the internet using sitemaps.org and schema.org dataset metadata through the Environmental Data Initiative Data Portal. In: *AGU Fall Meeting 2018. Presented at the AGU Fall Meeting 2018*. Washington, DC: AGU. pp. IN31B-20. Available at https://agu.confex.com/agu/fm18/meetingapp.cgi/Paper/445657 [Last accessed 6 May 2022].

sitemaps.org. 2006. *What are Sitemaps?* Available at https://www.sitemaps.org/ [Last accessed 12 July 2021].

Suresh, J. 2014. Bird's eye view on "big data management." In: *2014 Conference on IT in Business, Industry and Government (CSIBIG). Presented at the 2014 Conference on IT in Business, Industry and Government (CSIBIG)*. Indore, India: IEEE. pp. 1–5. DOI: https://doi.org/10.1109/CSIBIG.2014.7056930

Thessen, AE, Poelen, JH, Collins, M and **Hammock, J.** 2018. 20 GB in 10 minutes: A case for linking major biodiversity databases using an open socio-technical infrastructure and a pragmatic, cross-institutional collaboration. *PeerJ Computer Science*, 4: e164. DOI: https://doi.org/10.7717/peerj-cs.164

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

Van de Sompel, H, Nelson, ML, Lagoze, C and **Warner, S.** 2004. Resource harvesting within the OAI-PMH framework. *D-Lib Magazine*, 10(12): 18. DOI: https://doi.org/10.1045/december2004-vandesompel

Wilkinson, MD, Dumontier, M, Packer, AL, Gray, AJG, Mons, A, Gonzalez-Beltran, A, Waagmeester, A, Baak, A, Brookes, AJ, Evelo, CT, Mons, B, Persson, B, Goble, C, Schultes, E, van Mulligen, E, Aalbersberg, IjJ, Appleton, G, Boiten, J-W, Dillo, I, Grethe, JS, Heringa, J, Strawn, G, Velterop, J, Bouwman, J, van der Lei, J, Kok, J, Zhao, J, Wolstencroft, K, da Santos, LB, Roos, M, Thompson, M, Martone, ME, Crosas, M, Swertz, MA, Axton, M, Blomberg, N, Dumon, O, Groth, P, 't Hoen, PAC, Wittenburg, P, Bourne, PE, Rocca-Serra, P, van Schaik, R, Finkers, R, Hooft, R, Kok, R, Edmunds, S, Lusher, SJ, Sansone, S-A, Slater, T, Sengstag, T, Clark, T and **Kuhn, T.** 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. DOI: https://doi.org/10.1038/sdata.2016.18

TO CITE THIS ARTICLE:
Klump, J, Fils, D, Devaraju, A, Ramdeen, S Robertson, J, Wyborn, L and Lehnert, K. 2023. Scaling Identifiers and their Metadata to Gigascale: An Architecture to Tackle the Challenges of Volume and Variety. *Data Science Journal*, 22: 5, pp. 1–17. DOI: https://doi.org/10.5334/dsj-2023-005

Submitted: 08 September 2022
Accepted: 20 December 2022
Published: 01 March 2023

COPYRIGHT:
© 2023 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

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