

IGSN 2040 Technical Steering Committee Meeting Report

Canberra 13-14 May 2019

Table of Contents

I. EXECUTIVE SUMMARY	3
II. DAY ONE SUMMARY	5
A. ACKNOWLEDGMENT OF TRADITIONAL OWNERS	5
B. INTRODUCTION TO THE IGSN 2040 GRANT	5
1. BACKGROUND	5
C. WORKSHOP GOALS AND OBJECTIVES	6
D. INTRODUCTIONS AROUND THE ROOM	6
E. INTRODUCTION TO THE GOVERNANCE, BUSINESS MODELS AND ARCHITECTURE OF IGSN	6
1. DISCUSSION	7
F. DATACITE PRESENTATION	8
G. GBIF PRESENTATION	10
1. DISCUSSION	11
H. DISCUSSION OF THE SCOPE OF IGSN	11
I. AFTERNOON SESSIONS	13
1. STATUS QUO OF IGSN AGENTS AND USERS	13
2. CORE ELEMENTS	13
3. DISCUSSION	13
J. OUTCOMES OF AFTERNOON SESSIONS	16
K. SERVICES AND AUTOMATION	17
1. DISCUSSION	18
III. DAY TWO SUMMARY	19
A. RECAP OF DAY ONE	19
1. DISCUSSION	19
B. BREAKOUT GROUPS	20
1. BREAKOUT GROUP ONE – TECHNICAL ARCHITECTURE	20
2. BREAKOUT GROUP TWO – FUTURE VISION AND SERVICES	23
C. CLOSING DISCUSSION	24
1. DISCUSSION ON THE SCOPE OF THE IGSN E.V.	24

2.	TECHNICAL ARCHITECTURE SUMMARY	26
3.	DISCUSSION ON NEXT STEPS	26
4.	NEXT STEPS AND ACTIONS FOR THE ORGANIZATIONAL STEERING COMMITTEE	28
5.	ITEMS FOR THE ORGANIZATIONAL STEERING COMMITTEE	28
IV.	<u>APPENDIX</u>	<u>29</u>
A.	PARTICIPANTS	29
B.	AGENDA	30
1.	DAY 1 (13 MAY 2019)	30
2.	DAY 2 (14 MAY 2019)	31

I. Executive Summary

Outcomes

- The updated IGSN architecture will be able to scale to billions of samples and will enable the scope of IGSN to encompass any type of physical sample.
- Keep separation of registration and description of samples to keep the registration process lean and allow for extensible and domain specific sample descriptions.
- Transition from XML-based metadata to web architecture based on sitemaps (e.g. schema.org) to support the upscaling of the system from millions to billions of samples.
- Definition of the core services provided by the IGSN Registry and IGSN Agents, respectively.
- Keep services unbundled to uncouple the operation of the service from disruptions caused by maintenance and technology transitions.
- Automate services as much as possible to increase efficiency of processes beyond sample registration.

Next Steps

- Brief IGSN 2040 Organizational Steering Committee on workshop outcomes and their implications for the scope and business model of IGSN.
- Develop outline of the IGSN 2040 web architecture
- Develop blueprint of IGSN core services (IGSN Registry, IGSN Agency)
- Develop process for the development and curation of community vocabularies and description schemas
- Work with IGSN Allocating Agents to implement a global map and catalogue of IGSN registrations

The IGSN 2040 Technical Steering Committee (TSC) met in Canberra, Australia, on May 13 and 14, 2019, for its first workshop. IGSN 2040 is a project funded by the Alfred P. Sloan Foundations to conduct a strategic planning effort for the IGSN e.V., the implementation organization of the International Geo Sample Number¹, with the goal to “*achieve a trustworthy, stable, and adaptable architecture for the IGSN as a persistent unique identifier for material samples, both technically and organizationally.*”

During the 2-day workshop, the Principle Investigators (PI), committee members, and local observers reviewed the current IGSN architecture and implementations, learned about challenges and solutions at other organizations with concerns related to the IGSN (DataCite, the Global Biodiversity Information Facility GBIF, and DiSSCo, the Distributed Information System for Scientific Collections), and discussed various aspects of a technological strategy that would take advantage of modern internet technology such as cloud-based services and Structured Data on

¹ Following the workshop, the IGSN has been rebranded as the IGSN Global Samples Number to reflect the use of IGSN beyond the geosciences.

the Web/schema.org to achieve stable and trustworthy services of the IGSN to scale to the rapidly growing demands of its user community.

The technical design principles of the IGSN are more than ten years old and the shape of the IGSN system architecture had not changed since 2011. The TSC recognized that the world of internet technology has moved on: dedicated servers have been replaced by cloud-based services, XML as a way to structure information has been superseded by JSON and its dialects. It was therefore decided to transition from XML-based metadata to web architecture based on sitemaps (e.g. schema.org) to enable extensible sample descriptions and accommodate many use cases.

Part of the technological strategy towards sustainable IGSN operations is the decision to keep services unbundled to uncouple the operation as much as possible from disruptions caused by maintenance and technology transitions. From a business perspective, services should also be automated as much as possible.

Based on these high-level objectives, the workshop participants defined a number of design principles for IGSN services. With these design principles in mind, a breakout group analyzed three example workflows to identify the personas interacting with the IGSN system. From these interactions the group defined services that constitute the IGSN ecosystem and from these services defined the minimum viable product for the IGSN Registrar and for an IGSN Agent (IGSN in a Box). These minimum viable products define the services that the IGSN Central Registry and the IGSN Agents, respectively, have to provide to enable the principal IGSN workflows.

Since the number of objects used in research exceeds the number of datasets and publications at least by an order of magnitude, the scalability of the system to such large numbers is going to be a challenge. Different disciplines and organizations may come to different policies on what needs to be registered with IGSN and when in the lifecycle of the sample it needs to be registered. The IGSN 2040 project decided to investigate how the IGSN services can be scaled well beyond the current number of objects and transactions by leveraging modern web architectures and cloud-based services.

The outcomes of this workshop will inform the discussions of the IGSN 2040 Organizational Steering Committee at their workshop in Tacoma, WA, in July 2019.

A follow-up teleconference was held on 19 June 2019 to review the outcomes of the technical workshop. Two ad hoc working groups were initiated: (1) development of the IGSN 2040 web architecture for registration and description of IGSN objects, and (2) development and curation of community vocabularies and description schemas for IGSN objects.

The following text summarizes the discussions at the TSC workshop in a chronological order.

II. Day One Summary

A. Acknowledgment of Traditional Owners

Jens Klump opened the meeting by acknowledging the Traditional Owners of the land:

“I would like to begin by acknowledging the Ngunnawal and Ngambri people as the Traditional Owners of the land that we’re meeting on today, and pay my respect to their Elders past and present.”

B. Introduction to the IGSN 2040 Grant

This meeting is the first workshop of the Technical Steering Committee for the IGSN 2040 project. The project was funded by the Alfred P. Sloan foundation, and started in August of 2018. The IGSN 2040’s primary mission is to define the future of the IGSN as a global, persistent identifier for material samples. The outcomes of this project will be used to develop a plan for sustainability, suitable to the IGSN and its organizational structure.

The IGSN Grant proposal can be viewed on the project Google Drive.

1. Background

The IGSN is seeing growing adoption, from diverse domains, with diverse use cases. This is leading to increasing registration numbers and a diversifying of samples. From this need, the grant proposal was developed.

The IGSN 2040 grant has three work packages:

- WP 1 -- Organizational maturity
- WP 2 -- Architectural redesign
- WP 3 -- Management of the project

Overview of timelines for delivery

- Year One – Information Gathering and Strategic Planning
 - Technical Workshop in Canberra, Australia (*This event*)
 - Organizational Workshop in Tacoma, USA (*July 2019*)
 - This workshop will include an Open Forum at the ESIP Summer Meeting
- Year Two – Road mapping
 - Recommendations for building out the system
 - Tentative workshop for both steering committee groups together
 - Dissemination of results and publication of outcomes

Anticipated Outcomes

Architecture outputs

- Design of technical architecture
- Implementation plan and roadmap
- Outputs approved by IGSN e.V.

All documentation will be made available on the website for (new) members and kept updated.

The overview of the IGSN 2040 grant closed with a discussion on the intersection of this committee (the Technical Committee) with the Organizational Committee.

C. Workshop goals and objectives

- Meet and greet of the project technical steering committee
- Getting everyone on the same page
- Defining the requirements
- Sharing an overview of systems and architectures
- Mapping of different technical architectures as a basis for evaluation of solutions
- Exploration of technical architecture pain points and potential solution options
- Note where technical challenges overlap with governance and business model decisions which are the responsibility of the Organizational Steering Committee

D. Introductions around the room

Everyone introduced themselves with their organization and shared their relevant connections to the IGSN 2040 project.

E. Introduction to the governance, business models and architecture of IGSN

Jens Klump provided an introduction to the current IGSN Architecture. This included a graphic of the registration workflow.

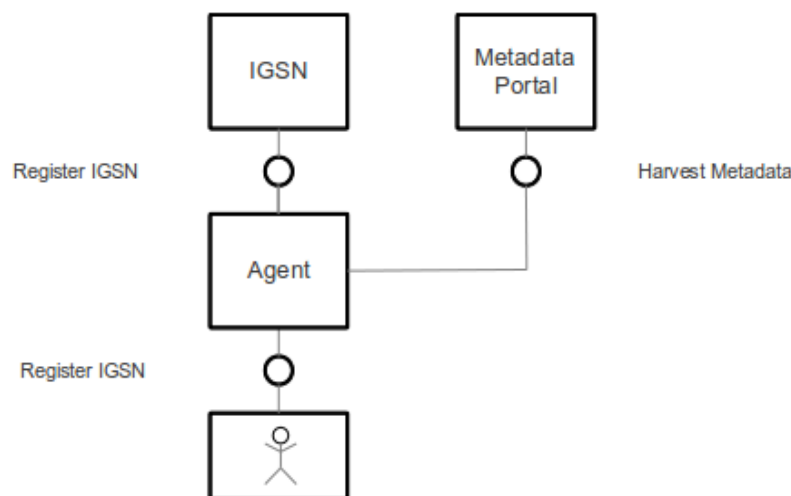


Figure 1: Schematic overview of the IGSN system architecture. Clients register their IGSN with an IGSN Agent. The IGSN Agent registers the identifier in the central IGSN registry (REST, IGSN registration metadata) and makes the catalogue of registered samples available for harvesting (OAI-PMH, IGSN description metadata).

- There was a discussion on each stage of the workflow
- The system requires users to submit 'birth certificate' but we are considering the term 'data passport'
- It uses XML to encode metadata. The OAI-PMH interface also offers Dublin Core (DC) encoded metadata because DC is the default encoding all OAI harvesters should be able to process. However, OAI-PMH is now outdated and its use should be reviewed.
- It has been working/stable since 2008 but it is dated and needs modernising because the methods used for encoding metadata and authenticating access to the system interfaces have become outdated. Furthermore, using cloud-based services would allow much better scalability of the system than the current use of dedicated and self-hosted servers.
- We can be fundamentally different -- how we implement this is part of this conversation this afternoon and tomorrow

1. Discussion

- Should consider using schema.org instead of XML. It is similar in scope, and can help with metadata distributed in several places. Move away from OAI-PMH for metadata syndication.
- Metadata portal. There are many, and many agents. There is no requirement for a metadata portal. There is a prototype but the team realizes that showing the number of samples is impressive and there is value for a centralized portal. There are still questions though -- once we grow in scope, what do we present and in what way? It was also noted that having redundancy would be good for the future.
- Background of the potential switch from 'birth certificate' to 'passport'. One applies for a passport whilst people are given a birth certificate automatically. Likewise, in the field or in the laboratory, anyone collecting/developing a sample gives it a label that is used within a local governance system (e.g., within the laboratory/activity/project to create a unique label (c.f., birth certificates are unique within a country). But to make the sample identity globally unique, an internationally recognised unique identifier system needs to be used. Note: passport is actually the terminology used by the FAO when referring to standard descriptors for genetic materials ([FAO document](#), [Multi-crop Passport Descriptors](#)).
- Process for each agent for IGSN getting their namespace: There is a process through GitHub -- put in central location for agents to use. This is one of the services that should be automated. Currently it is a manual process. It is also for top level only. Allocating agents might have a workflow for how they manage their own sub-namespace allocations.
- How the IGSN system connects to other PIDs. This is something that still needs to be worked out with the other PID providers. DataCite allows IGSN as related identifier type, and vice versa.
- What happens when you query the IGSN? There is a relation to other objects but bare minimum and it might be better to put that connection elsewhere in the architecture.

More discussion is needed on what goes to the central registry, and how rich the metadata would be there, or if one would have to go through the harvester to the originator.

- Timing on return for requests to the IGSN system. That depends on the allocating agents' architecture. When bulk implementing, there is a problem with the Handle system which could be overcome through load balancing. Slow in this case is defined as a considerable amount of micro seconds.
- Bulk uploads -- when a new organization joins and is dealing with their legacy, that is punctuated. However, there are often people coming back from the field registering thousands in a day not year.
- Speed will be an issue. Depending on the size and scope, that will affect workload on the central node. This led to a discussion on bottlenecks and scaling.
- If you have a handle service running, you have to send requests out to go back to your system. You can put in a back door so you do not have to go out. This was why DataCite moved their Handle server to be hosted by CNRI. The performance issue is between top-level registration service and Handle.net. The performance issue between agent service and top-level registration service has been resolved.
- Diversity of use cases is also an issue along with the scale of parallel registration and workflow of registration. These topics will be discussed in the afternoon breakouts.
- Registration of 1 million samples is rare, but if we (IGSN) grow, the redirection needs to be updated. If we are already changing infrastructure, may need to change/update this process. This delay will increase with number of samples and it is important to scale.

F. DataCite Presentation

In advance of the workshop, Martin Fenner, Technical Director of DataCite, provided a written assessment of the IGSN needs: Service-Oriented and Sustainable Infrastructure for IGSN: DataCite Perspective. The document showed that DataCite's service is complex and thus expensive to maintain and develop, which is balanced against the focused use case of registering data publications and describing them with bibliographic metadata, plus offering added value services based on these metadata. This distinguishes DataCite services from IGSN services: In the case of IGSN, the objects that will potentially be registered can be any kind of physical sample that originated from anywhere in the universe. Describing these objects comprehensively with one common set of metadata is not possible and will always require community extensions to the metadata. A key decision was therefore to keep the separation of registration and description of samples.

Critical facts from this assessment:

- Main costs factor is staff
 - Technical -- development and maintenance
 - Non-technical -- management, governance, community outreach, etc.
- Cost of services you want to provide vs revenue from services
- Can't have something that cages us in and gives us independence/stability that we currently do not have

Martin virtually joined the workshop and provided a presentation and led an in depth discussion on this document.

Cost-Scale-Complexity

- Cost -- has been THE limiting factor for IGSN. The Organizational Steering Committee also needs to discuss the business models related to this. We learned from other PID systems that annual budgets are around Euro 1.2 million (10% infrastructure, 90% staff)
- Scale -- no one knows what the scale is or will be. Percentage of adoption, how much of the community do you reach, want to reach, what does that do to the number of identifiers to be created ... services consumed.
- Complexity -- Number of metadata elements, common vs specific elements; integration with other PID services; what services are centralized vs not.
- DataCite Model -- central registry. More complex and more expensive to implement. ...
- Points to similarity to a paper on the data curation continuum by Treloar et al ([updated version of figure](#)). Do we need to register everything that is collected or just those samples that support publications?
- Review of levels of effort -- housekeeping for internal work. But also need to do work to share outwards. Once published, context is lost and must be explicit about these connections.

Discussion

- What part of the sample we use and when? Do we need IGSNs at the start? Or only what is used?
- Collections vs individual user perspective. Collection has a different perspective on where you would want to register the sample. Someone needs to pay for that convenience, so a fee per minting? Mint just in case has a cost on the central infrastructure and need a way to balance that.
- Suggestion that we present various cost models and options for various perspectives and let people decide.
- This will have impacts on the scaling model as well, if everyone wants to be on one end. Effects the architecture.
- Adoption. We want to augment adoption. What makes or encourages adoption? If it is a burden to implement for every sample, you collect in the field.
- We have greater metadata; where need bare minimum for publication but as we go down the path more specific metadata is needed. Assume this might happen with samples as well. Minimum metadata with context elsewhere.
- We have done minting in the field. From the user experience, it did not matter that it was an IGSN, but that the metadata was captured and didn't have to spend the night transcribing metadata. Just need a kit which can sustain the environment.
- When researchers came in from the field, there were concerns with duplicate data
- Different perspective of researchers that have been doing things the way they have been doing. Some use paper worksheets
- How and how many to register is a minor thing in defining workflows for updating metadata.

- Subsampling -- in a template, can add a suffix from a handle to refer to a subset. DOIs for video is one example. Video has a DOI and the individual timestamps gets a suffix. Want short strings but other fields do work with these strings, this technology exists which would lower the load on the resolver.
- This is a brittle technology implemented in the browser not server. Limited use case for navigation, not as PID.

Break

G. GBIF Presentation

Donald Hobern from GBIF gave a brief presentation.

- Brief overview of GBIF and their types of data. Showed an example of a specimen. Uses Darwin Core for metadata standard. Gets data as a CSV sometimes xml.
- Showed diagram related to parallel efforts from other groups
- Discussing a shared pipeline with other groups (switches at a national level). Wants to move away from each of these groups having their own identifier which has metadata which is processed differently. Reviewing stable persistent identifiers for all these things. How something is downloaded from one thing and then used in publication, can track back to previously contributing organizations which might have shared the data.
- Shared GUIs is something they want to deal with.
- 1.3 billion things in GBIF. Along with other new groups which will provide new streams. Adding one group added 2.5 million records at one time.
- Some records at best have an integer that is unique within a given dataset as an identifier. Review of some of the variety of identifiers. The European history groups have been trying to unify this and are optimistic about how it might work.
- GBIF assigns a unique identifier, but would like a more stable system for this.
- No single resolver for all specimen, leads to some issues.
- Donald reviewed what IGSN looks like -- *see diagram on slides*.
- Is there a different metadata record outside of what is submitted to IGSN?
- IGSN features of note
 - Standard syntax
 - Standard metadata
 - Handle based resolution
 - Stable metadata catalog (even if backend of source is rudimentary)
 - Option for large institutions to resolve their own
- Applicability of IGSN model (*see table on slides*)
 - What things from IGSN map to what they have in the biodiversity realm.
- Discussion on piggy backing or connecting to the IGSN system.
- Does IGSN go beyond physical object? 1) collections or aggregations of objects 2) being able to identify a location or locality (referred to by different objects). A site -- location, time, actor, sample.

- Bringing these in might not be good for the governance, but what if it is an umbrella for what GBIF is doing. Replicating tools and software (two different identifiers) but benefit from configuration.

1. Discussion

- Similar feelings from DataOne as what Donald discussed from GBIF. Similar technical and sociological issues. Good place for IGSN to fill a portion of what is missing from DOIs and ORCIDs.
- 1:1 relationship with a specimen and an identifier. Nature of the evidence is different (assertion, image, page in a book). Lots of bird observations, growing cases of samples of a many to one relationship to sample (seawater, soil with millions of bacteria inside it). Not different from a treatment to a rock core. There was a sample there. Likes that there is not a separate identifier for the rock from the metadata.

Alex Hardisty and Dimitris Koureas who were unable to attend sent a document outlining the requirements needed for DiSSCo. This was brought up during the discussion but not reviewed.

H. Discussion of the Scope of IGSN

- Big or small IGSN? Which domains should be included? Who is the target audience? Who are the clients?
- IGSN members will need to approve before any decisions are finalized -- we need to give them information to make an informed decision.
- General purpose identifiers for use in science to describe an entity. Material sciences (vs experiments in a lab). They want something to help uniquely identify stuff. Anything associated with a branch chain afterwards. For that general, metadata requirements must be simple.
- Problem with multiple registries depending on field (when multi domain). What about virtual samples? Where is the line?
- DataCite suggested we become a member of DataCite, but that becomes blurred.
- DataCite has objects registered but feel they do not have the expertise related to physical objects, community expectations, workflows. Managing physical samples.
- Do not want multiple implementations for the various samples they manage.
- With an extensible model, if you get an identifier with some information about where to go next, works. Fewer processes the better.
- Nervous about over generalizing. Might get forensics, social sciences, and other things with samples. Need to look at the core metadata, keeping it simple, and anything that doesn't fit in there doesn't fit scope.
- Sample is sampling a thing or a type of thing and that is an important part of the description.
- Should look at the invitation to DataCite. There are other parts that these big service providers are better suited to take care of. A best of both worlds arrangement would be

useful. That is for the IGSN 2040 Organizational Steering Committee. For the Technical Steering Committee, we have the handle system, at the most vanilla, you can get a handle, and want to identify something. A DOI is an identifier and some information. Author, title, model of the thing being identified and a service. What is the identifier plus metadata combination which is useful?

- Use case for some samples which are analyzed years later for a different purpose than originally collected. Need to account for that in depth of metadata that provides context.
- Context -- things that are identified as natural materials, when getting to museum collections which include cultural items, might have been archaeological or expedition or purchased. Where do we draw a line? On bottles of valuable wine?
- Is an important aspect in structuring the connection but also the different entities of the physical sample and the data generated? Yes, it can happen years later for a different purpose. But the linkage is the linkage between two identifiers. SESAR and EarthChem does this. On EarthChem it is in the metadata. You don't need to worry that the sample can be studied for a different reason. But SESAR includes sample procedure because it affects reuse. For example, Paleomag samples must have orientation, but the actual sample can be reused for other purposes.
- We need to understand why we don't just use handles for everything. A more community-specific identifier can help us with territory marking -- expect a certain kind of thing when you get it, would want to put more effort into standardizing at the handle level. The thing which sits above that is the community which shares the common practices and interests and wants to be able to rapidly retrieve. Some people want things that are more fine grained, others want broader classifications.
- IGSN has a goal around a common practice and common lens. IGSN up to this point has been a focused lens. A geological, petrological community of interest. Given that you are looking at boundary, what you are defining the commonality of interest. Like a metadata portal which would show all samples that fell into that community.
- Research is becoming more interdisciplinary. Example -- Some are on earth science, but do take leaves, genomic samples, and do not want to go to five different places to register a sample. Can't put the boundaries because it limits adoption.
- Could have management layers in the IGSN system, which support the different communities which use a shared architecture -- where you have your community specific with the allocating agents underneath this. Maybe parallel systems rather than do everything in one.
- Not surprised everyone wants to use IGSNs, well governed and persistent. We want to answer the same questions, but you retain something tight for IGSNs and use this as a model for replicating stack and governance for different classes of things. Different use cases.
- From a large research organization perspective (e.g., CSIRO), it is complex to have to run different identifier systems for different sample types.
- Define ways for a community to optimize for their community.
- Pattern matching, if it includes x, then it is this community of interest.
- Adding to bare bones metadata schema that has a field that identifies which type of samples it is, and being very good with maintenance of that... In 2011, talked about

object types but there are different classes, and being consistent in the class of objects. (physical appearance vs, method of collection)

- If the system allowed multiple classifiers would that help? This would be a nightmare, hard enough to be consistent.
- When you assign a classifier know where it is coming from and allow more than one, as long as you know which set it comes from. Usability issue with that though.

I. Afternoon Sessions

Kerstin recapped discussion and presented a modified agenda for the afternoon to redirect the conversation based on the outcomes from the morning.

1. Status Quo of IGSN Agents and Users

- Preliminary Survey results (Sarah)
- Currently active IGSN Agents (fill out spreadsheet)

2. Core Elements

Jens reviewed the core metadata required by IGSN -- this can be found on the IGSN GitHub account (<https://github.com/IGSN/metadata/wiki/IGSN-Registration-Metadata-Version-1.0>). In the service discussion we might talk about making the workflows better.

There are four core elements:

1. Sample number
2. Registrant
3. Related resource identifier
4. Log

All other metadata was left to the agent, that might be used for discovery. That might be harvested and be made searchable.

We do not have documentation that captures processes such as: How do you request a namespace? How do you transfer custodianship of a sample?

We have a registration services, pilots of catalogs, and manual services (accounting and billing, and namespace registration). Push authentication and accessibility/embargoes to the agent level.

3. Discussion

- Privacy controls and restricting access to metadata
 - When a sample is registered with an IGSN, it may be made private. These are not made available in a metadata search.
 - If using a related resource identifier, must be concerned with metadata becoming out of step

- Use of access control -- would need shared authentication system. Authentication is currently the responsibility of the allocating agent
 - Discovery is a different service in the IGSN infrastructure
 - Currently type of schema and the relationships type are required for registration
- Public portal for metadata is an issue for the Organizational Steering Committee as it impacts governance.
- A registry of all OAI-PMH endpoints would be another service needed.
 - This is 10-year-old technology.
 - Would like to move away from it.
- Suggested solution -- Profiles, bio, rock, with a common core -- like JSON context. That whole environment has an RDF data model. That has benefits over XML. You don't have to crawl because we publish a site map, so when harvest, what you need to look for is discoverable and you can get it. Exposed on a landing page, so that you have a type, and an evolution of a type. Advantage is the allocating agents use web architecture patterns, pipelines and workflows instead of OAI-PMH.
 - The attractive part of the above is it is used in large scale by google and others. Instead of pushing to central authority or constant crawl, only crawl what changes. And is simple for a provider to implement.
 - This model doesn't require local or decentralized identifiers. Have an ability to maintain a central.
- The IGSN Bull's Eye Diagram: We have an agreed very small common kernel that suits multiple domains and beyond that, you get what you need out of other governed systems within the disciplines. You can also have constraints based on these rings. (See *Figure 2 below*).
 - If you use this agreed profile for the common generic kernel, it can add individual governed vocabularies, ontologies and specifications from multiple disciplines, mixing and matching whatever is required within a small community to build a profile that suits their business requirements and use cases.
 - Points raised in response:
 - 1) registration and description element. Requires less from providers. Is this IGSN? The harvesting step needs to answer that question, and interact with the handling service.
 - 2) what ends up in the index is an entry for web pages, a prerequisite is that you have a webpage for every sample -- landing page for each sample. Plus, the site map. Would the handle part -- it is not quite part of this story?

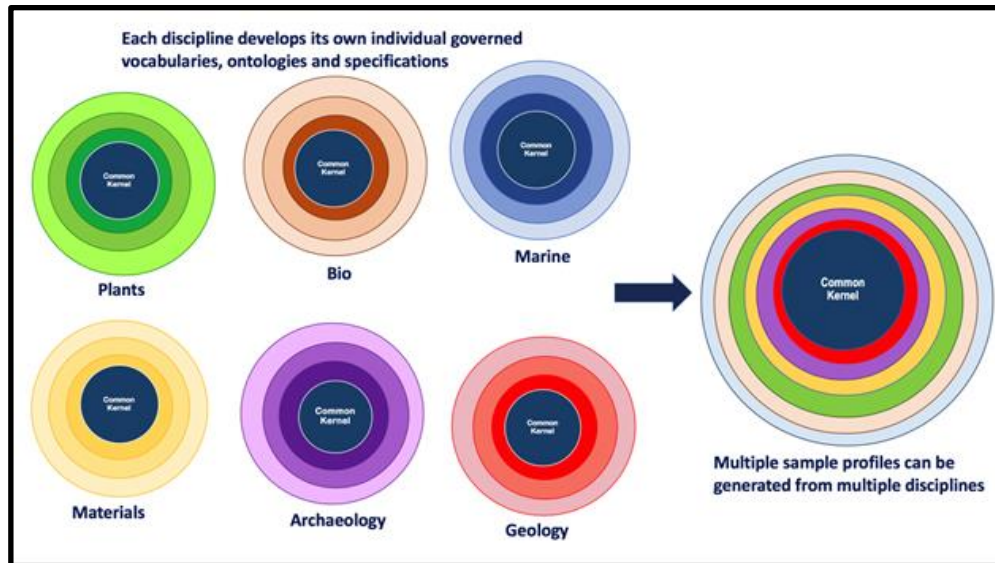


Figure 2: schematic diagram how multiple profiles can be built for more specific use cases, based around a core kernel that is common to all samples from all domains. For the descriptive metadata in the outer shells to be useful and persist over time, attributes from each shell should come from a governed vocabulary in persistent name space. (Source of Figure: L. Wyborn).

- Schema.org gives us the right hand side of the diagram, getting all the information. But it doesn't answer how do I refer to the sample in an unambiguous way?
- You can have a lightweight registration method, and as you add pages into the map, it adds an element for the schema for each. But it is error prone. And you can validate that in a central area.
- Keeping registration explicit, makes accounting and billing easier.
- This method suggested keeps the discoverability separate. Which is part of the current model anyways. 'I have the journal article with IGSN' and want to provide the reference information. Can't ping a centralized service. Aside of the top stuff. Unless your landing page provides that information. Which can be done if collected on the individual allocating agent.
- Linked data API. Can ask for different formats, but on a sample by sample basis, and they did it, it is not at other allocating agents, and is something that can be developed.
- If responsibilities are separated, there is concern with the implementation. The IGSN e.V. is responsible for this part, and your organization is responsible for the rest. But how do you ensure quality?
- Unless someone maintains a cache, it is heterogeneous. GBIF was just harvesting, and the only public version was what we indexed, and is no longer in the repository. Needed a policy, linked open data that we keep the identifier that is shared and common.
- The persistence of the landing pages of the allocation agents is important.
- With DataOne, we find some organizations are already using this to make data discoverable, but there are no standards in how to represent a data set in schema.org.
- Need for best practices in this regard.

- Suggestion to design this as evolution, where people can upgrade alongside what they already have. Might be more painful along the way doing both, but will be better for the community.
- Question about if there is a need to bring together the folks from the allocating agents with the Technical Steering Committee to work on next steps.
- In regard to the concept of persistence, will have to integrate with core trust seal to see that an allocating agent meets certain standards.
 - GBIF is going to seek CoreTrustSeal certification for the central data index, individual contributing datasets should also be managed in best practice repositories (CoreTrustSeal).
 - CoreTrustSeal is based on OAIS Reference Model. So is transferable.
- Question -- when a sample is under an embargo, identifier is ok but sample is not?
 - Can think about cases where do not want people to know that samples exist.
 - One allocating agent waits until the user says they are ready to publish, before registering the IGNSs

J. Outcomes of afternoon sessions

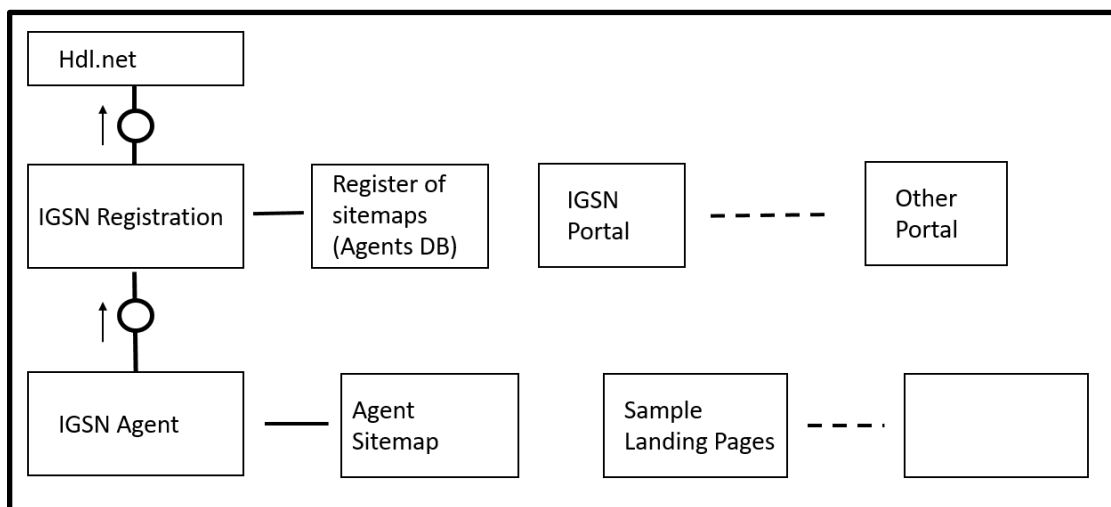


Figure 3: Relationship between the IGSN agent and the IGSN registration

Left side is mostly still the same. Then the metadata comes from the agent level and goes to a landing page or site map that can be accessed by other people. Need a registry of site maps or use the allocating agents to create that (database of the agents can be home to site map registration). Can run a portal on top of that -- (but need a type to be able to filter that). The client is not required to hold harvestable metadata, but the Agent is required to provide a catalog of all IGSN registered by them.

Martin joined the call. Mentioned identifiers.org. Provided some updates on schema.org relationships. It still provides challenges to searching across repositories.

Break

- Martin provided an update on DataCite's work with JSON LD. Talked about use of schema.org. Suggested against using microdata format.
- Discussion of a BOF at RDA -- it will be converting to a working group at RDA to coordinate between different disciplines of schema.org to develop common metadata. Should help with related activities.
- There was a discussion on schema.org being controlled by Google and Microsoft. But there was disagreement with the concern -- that while driven by commercial interests would prefer schema.org over Dublin Core or creating something yourself.
- Question about moving away from OAI-PMH, if it should be incremental.
 - if a demand by users you might have to use it. But it is older technology. It is still used widely in the library community. Would recommend supporting but not using it. Not as a centerpiece of the architecture. Continue providing support but move to schema.org.
 - What work is required from the organization and what from the allocating agents to implement this new approach? Need an action list of priorities. How to do this without resources. Had a point to have a Reference implementation list -- a test implementation that works in the way it is designed.
- The EarthCube project has a gleaner that can harvest from schema.org or JSON LD, so if an allocating agent goes through this effort -- there are tools that can harvest and can make some simple work to validate, it is part of what they are already doing in EarthCube.
- Needs
 - Need a tool to read in existing XML and 'spit out' schema.org
 - Schema.org does not enforce anything: need a validation tool that assures you that certain bits are there.
 - Need to define shape constraints. This is what we expect for this type of profile.
 - We need to flesh out the use cases for discovery, registry
- With this process you can control the vocabs - they are built in. Can specify something things as compulsory and others as optional.
- DataOne is putting out recommendations on what they can harvest from schema.org.
- Schema.org has not solved the community specific needs, and the extensions are not used that way. It has standard metadata, and when going to subdisciplines, keep your own metadata schemas and supplement with this.

K. Services and Automation

- Managed services
- Validation services
- Metadata completeness
- Cloud based and multi region (rather than having to round trip to Germany if you need access)
- Persistence of metadata (responsibility of the central repository -- whole lot or just top level?)
- Outsource handle service (or buffer service)

- IGSN de-duplication
- Transfer of governance of sample
- Allocate/transfer namespace
- Authentication (maybe into metadata you are not allowed to see)
- Deletion/Errata

1. Discussion

- There is an issue with latency but the current problem is between IGSN and the handle system, what can we do in Potsdam to make this better?
- Uniqueness -- when minting an identifier, how do you see if there is a preexisting identifier for the sample? There may be an IGSN and they don't know it. And you need to check before minting a new one. So may know the sample number, the rock type, the collector, but it is not a letter for letter matching. It requires fuzzy logic that recommends there is a match.
 - highly relevant. When entering analytical data and link to the same sample (currently requires the curator to check).
- If you have two versions of the metadata, do you retain both or are you editing the same record?
- Samples you have are different from namespace; some journals want IGSNs, and the journal issues the namespace, the identifier doesn't mean anything about who is in possession of the sample.
- There is an issue with legacy systems, operators of repositories that do not want to introduce another identifier but to use an existing prefix. But other examples where the random string works. Jens pointed us to [this page on wiki](#). This discussion is for the Organizational Steering Committee.
- If we want to provide a service that provides deduplication, how will the top level service be supported? It needs to be drilled down into the agent.
- Single identifier per object is the goal, but lazier solution, as things show up in portals, forcing people to assist with identify the match. Either two paths to the same thing, don't accidentally lose one that way.
- Discussion on authoritative sources and trust related to annotations on samples.
- We need to look at which are generic issues, which are PID issues, which are IGSN, and which are agent issues. Perhaps we can sort them on this and rank them on urgency.

End of Day One

III. Day Two Summary

A. Recap of Day One

Kerstin provided a recap of the first day of the workshop. Discussed that perhaps we should not have separated the Technical and Organizational Steering Committees at the start, but the work we are doing here and in Tacoma in July can work towards that shared goal.

Perhaps we can use scenario planning -- envisioning positive, negative, etc. to see what influences your success and where you should adjust to drivers.

This morning we are going to split into two groups to talk about the technology drivers and the user drivers. We did a lot of backwards discussion yesterday, and today we should be focused on the future. What will the technology be in 10 years, how does the IGSN place itself in that? What about the research ecosystem? How will that change?

1. Discussion

- Looking at IGSNs, John Morrissey went through questions as to why an enterprise like CSIRO would want to adopt and risk profile with that? Sustainability and things tend to be blockers. His final assessment was that it was risky, but the pay off with all their collections, was they needed something that works for them all. Going forward, have base service like registration and ID, but what is the value add services that other research organizations need to implement service? They need support and that is something you can charge for. If you have a standard deployment, uptake and growth will be faster. Need a reference package and federated architecture.
- **This is where the Organizational and Technical Steering Committee crossover.** Will not make money off of selling the individual ids. Have discovery portals, world views, are high value which people will pay a subscription for. Also willing to pay for a helpdesk/chat feature.
- Useful to separate out the use cases, and with the perspective towards the future, what are the goals of the use cases, and how IGSN can meet those goals.
- The question of who we want to be? Can't just focus on individual investigators and have the organization support itself. Would be reluctant to give up on that.
- If they only target national libraries, there is a limited number of targeted users that could ever reach and that would not support them funding wise. So went broader, and smaller fee and service fees, and need 400 to make even. Have a tiered fee system. The question for IGSN how many members can we ever envision.
- **Usage is growing, but members are shrinking**, because the current structure says all should join and have one person join.

B. Breakout Groups

Following the morning discussions, the attendees split into two breakout groups.

- The first group was tasked with discussing the system architecture of IGSN.
- The second group was tasked with discussing the scope of the IGSN.

1. Breakout Group One – Technical Architecture

a) Summary of outcomes

Group one discussed:

- Architectural design principles for the IGSN system
- Key workflows in IGSN
- Interfaces between the system components and actors involved in these workflows
- Minimum viable product for the IGSN Registry and for IGSN Agent services (IGSN in a Box).

b) Architectural Design Principles

We started by outlining the design principles for the future IGSN system architecture.

- Data is becoming more mobile
- IGSN as part of the Research Graph
 - Persistence of samples and related metadata
- Connect physical samples to digital context, provide context
 - Commodity platform
 - Usability in field and lab
 - Findability
 - Schema generator
- Usability in the field and lab applications, supporting user communities – users and best practices
- Readability of the identifier
- Cope with bulk registration
- Offline maintaining before registration
- SLAs IGSN – Agent – Client
- Reference to community schemas
- Simple and Stable API

c) Roles and Example Workflows

As the next step, we identified the personas in the IGSN ecosystem. We then outlined three example workflows and identified the personas involved in these workflows. The example workflows were:

- From sample to publication
- Obtaining a sample for reuse
- A community of practice identifies the need for a new sample description

Personas involved in IGSN processes:

- A. Sample Provider (individual)
- B. Sample Information User
- C. Sample Information Aggregator as a Consumer
- D. Collection Manager
- E. Sample Aggregator as a Provider
- F. Schema Developers
- G. Communities of Practice
- H. IGSN Agent
- I. IGSN Registrar (IGSN e.V.)
- J. International Standards Bodies (e.g. RDA)
- K. Publishers and Editors
- L. Manuscript Author
- M. Application Developers
- N. Lab Managers
- O. Data Repository

Example Workflow: Sample to Publication

1. A collects sample
2. A registers sample with D, H, N
3. A, N makes observations on sample, derives samples (optional loop-back to 1)
4. B, C perform data analysis and use external data
5. L write paper and submits manuscript containing IGSN to K
6. A, D, N, O deposit data
7. B, I, K verify identity of samples and data

Example Workflow: Obtain Sample for Reuse

1. B has search criteria
2. M enables search
3. B evaluates search results
4. B selects samples
5. B evaluates data provided by O
6. B selects samples for reuse
7. B requests sample from D
8. B, D1, D2, H1, H2 update custodian
9. Link to “sample to publication” workflow
10. H update sample information at I

Example Workflow: Community of Practice Requires Development of a Sample Description Standard

1. A, B identify need for a new sample description standard
2. G recognises need for a new sample description standard
3. F, G, H develop draft

4. F, G, H, I, J, K go into a black hole
5. I adopts draft
6. C, H, M, O build reference implementation and cool gadgets
7. A collects metadata in new description schema

d) Services and Minimum Viable Product

Following on from the workflows and involved personas, we extracted the services needed to support the workflows and then mapped these services to personas that use or provide these services.

Table 1: Services and Minimum Viable Product
H is the IGSN Agent. I is the IGSN Registrar (IGSN e.V.)

Service	Sample to Publication	Obtain Sample for Reuse	MVP IGSN Registry	MVP IGSN in a box
Sample Doc App	A, (M, N, D)			
Sample Metadata upload	A, H, D, L			required
Sample Registration	A, H, I		needs update	required
Sample Discovery Service	A, B, C, N, O	B, M, C		
IGSN Resolver	I		available	
Data Submission	A, L, O			
IGSN Validator	K, I, B			
Federated Data Search		B, H, M, O		
Sitemap Registry		H, I, M	missing	publish to registry
Schema Registry		H, I, M		required
Sample Request		B, D, F		
IGSN Ownership Transfer		D, H, I	missing	required
Authentication Service		H, I	needs update	required
Namespace Allocation		H, I	needs update	

In the last step, we identified the services required for a minimum viable product (MVP) for the IGSN central registry and for an IGSN in a Box services deployment for an IGSN Agent.

2. Breakout Group Two – Future Vision and Services

a) Summary of outcomes

Group Two was tasked with discussing the questions:

1. What is the scope of IGSN?
2. How multi-disciplinary should it be?

To address these questions, the group considered the fundamental goals of the IGSN. These concepts were approached from the view of the technical aspects of the IGSN and should be reviewed by the Organizational Steering Committee.

b) IGSN - What we do

- IGSN e.V. is a non-profit organization that operates infrastructure for a global sample registration system to enable transparent and traceable connections between samples, collections, data, publications, people, and organizations.
- IGSN promotes responsible management, sharing, reuse, tracking and attribution of samples. Organizations join IGSN e.V. as members to support this mission as well as to assign globally unique persistent identifiers (IGSNs) to physical samples. This way, their physical samples become digitally identified, linked and made discoverable to the community.
- IGSN e.V. actively promotes best practice for the management and description of physical samples. These activities enable our members to connect and share their physical samples with the broader research ecosystem.

c) Vision

IGSN's vision is a world where physical samples are valued, uniquely identified and linked into the scholarly ecosystem to enhance their impact and to support transparent and reproducible research across disciplines, borders, and time.

d) Mission

IGSN is part of the wider digital infrastructure needed by researchers and others to share information on a global scale. We enable transparent and traceable connections between samples, collections, data, publications, people, and organizations by providing a globally unique persistent identifier system of choice for physical samples.

e) Principles

(T) indicates items relevant to the Technical Steering Committee.

1. IGSN e.V. promotes the responsible sharing, reuse, tracking and attribution of samples. (T: discoverability, privacy of sensible metadata, capture of provenance)

2. IGSN will work to support a permanent, clear, and unambiguous identifier of physical samples.
3. IGSN will transcend discipline, geographic, national, and institutional boundaries. (T)
4. Use of the IGSN is open to any person or organization who has an interest in the identification of physical samples
5. Access to IGSN e.V. services will be based on transparent and non-discriminatory terms posted on the IGSN website.
6. All data contributed to IGSN e.V. will be available in standard formats for free download. (T)
7. All software developed by IGSN e.V. will be publicly released under an Open Source Software license approved by the Open Source Initiative. For the software it adopts, IGSN e.V. will prefer Open Source. (T)
8. IGSN e.V. will be governed by its members as defined in its statutes.
9. IGSN e.V. promotes best practice for digital management and description of physical samples. (T)
10. IGSN e.V. promotes the documentation of relevant ethical and legal issues where necessary.
11. Each IGSN identifier will be supported by a landing page (T)

C. Closing Discussion

The final discussion focused on what points still needed clarification. The following questions

- Are we all in agreement about the scope of the IGSN e.V.
- What are the next steps? How do we use what we discussed and achieved during the meeting?

1. Discussion on the Scope of the IGSN e.V.

- IGSN could be a platform for sample identifiers, community specific. IGSN like framework used by museums, etc. Or IGSN provides identifiers across a number of domains.
- The idea of this being a multi domain identifier. Is it a single schema, and if so what are the costs, implications, and use cases for discovery for all samples. Or is it a platform which allows each group tailored to physical samples do their own thing. That encourages new communities to come to this the way geo did, and allows them to do the things that fit their needs?
- There are technical questions with this approach. Do they own their own namespace? Do they own their own infrastructure (cloned) or is it shared? The minting services, technical support? Could be a commodity.
- There is a risk with a framework approach. User confusion. Mission is less clear. For data sets, the identifiers do not work together. There is a risk of punting the problem and then you have 10 different technologies that do not integrate. **Suggestion - IGSN says this is how we do identifiers for samples and sets up a path.** Worries about difficulties coordinating if different systems.
- If it is an all-encompassing system, could you still support individual communities to work within their interests? And build their own best practices? That it all works in the same

infrastructure, but different recommendations for how each type of sample (rock or specimen) is managed.

- If base infrastructure, where can plug in a vocab, as a community deciding what top five things they want to search on. Discovery could be on a field from a birth certificate - picked up in an area, by this person, and is this type. In a domain might have more, faceted searches. Science is becoming more complex. Being able to look at an area and doing an environmental impact study -- what samples do we have in this area? Multidisciplinary by design. The difficult questions involve solving problems cross domains.
- What does success look like in 10 years in this multi sample type area? Maybe there is a shared infrastructure which enables different sample type communities? Manage and govern parts for discovery? Have not gotten down to why is it common? And what is different you want to manage in each. But we don't have the way are they on the same diagram. The success of the IGSN is the social and community framework for samples in a number of domains where they can use the common identifier infrastructure.
- Examples of services that are generic and used in different ways
 - Max Planck
 - Pangea
 - CSIRO data.csiro.au is multi-disciplinary by design.
 - Common science data model and underneath can do Darwin Core, Dublin Core, and they do a cross walk of those things. The same portal, do a slightly different search, without leaving the portal look at Darwin Core metadata.
- Would need a vocabulary server.
- **Is this a model that we can agree on? It is a governance decision.** But there are technical implications. If we want to support the geoscience use cases, the more costs there are in metadata filtering and how we harvest and index to make this possible. So there is a trade off in expanding contributors and providing the specificity they need. Implications for governance.
- How do you build something in two years with 10 years in mind that still works for today. Do not lock your path forward. Not losing those who are here now.
- **We need to define a set of questions that we can carry forward to the other committee to talk about those governance issues. Are we at a place where the technical questions can move forward without these answers?**
- Yes, the broad technical issues, we can rely on existing. To build a functional prototype to see what really comes out of it. Could but straight forward and provide a template or narrow framework that provides reliable identifiers for a community where they develop their own schema and scope. From a governance perspective, we can focus on physical samples with this infrastructure pattern which is generally applicable, but don't try to expand to astronomy.

2. Technical Architecture Summary

- The group looked at guiding principles for system architecture and identified actors. And three workflows: 1) Sample to publication, 2) obtaining a sample for reuse, and 3) coming up with a new description for something (e.g. adding a new community). They then Identified who was involved in which step.
- Next, they look at these workflows, what services do we need for these workflows, and who is touched by these workflows. Then they looked at where the IGSN central repository is concerned, and what is needed to develop this 'IGSN in a box' (i.e. the instantiation for IGSN Agents).

a) Discussion

- Many groups are using handle services for identifiers. And perhaps discussing upgrading infrastructure and having someone else doing some of this for IGSN. and focus on syntax for identifiers.
- The community does not see moving away from handles, but other namespaces might be how do we bring other groups on board. **Do we build on what we have - or have other handle namespaces for other communities? I would leave that to the governance on how to move forward.**
- Could see a discussion on how handles are operated coming out of this discussion. Some concrete work on how samples are supported by schema.org.
- Instead of relating to schema.org, it is a more general schema. We know schema.org does not describe samples.
- In relation to handle systems, that is a good low hanging fruit that would benefit IGSN with an improved or optimized infrastructure there. Did not come up in our discussion here. However, when we discussed what is IGSN as an organization, it is a place where you can mint an identifier. Minting identifiers is just the beginning. That is a core function. Which maps to the first aspiration -- to globally identify and help to manage identifiers.
- Are there other things we can do better? Or say that needs to be discussed. Not just about registering identifiers but also about metadata. What else should be part of going forward? Aside from where this should be in 10 years.
- The values. What values can we offer, and what complexity do they raise in the demands.

3. Discussion on Next Steps

Considerations -- What comes next? From what we discussed. We have a rich source of considerations discussed here. Where do we want to get to with this? What are the tasks we want to get to next for this group? Ongoing discussion? What is the concrete outcome from this? Do we just write this up or is there work that can be achieved?

- **Suggestion -- generate a better diagram for the other committee where the vision is going.**

- **Would like to see some of the things we discussed in the other group codified.** Technical decisions were not made yet. Something easily digestible. **Also need some of those early workflows developed.** Where we want to put development into based on these workflows (site map identifiers because enables these workflows). Can make governance decisions based on that.
- Let's us identify who does what, internal and external. And where is the potential for revenue.
- Where is the first deployment and how quickly, so we can find out if this is working.
- One allocating agent mentioned they had a discussion on doing system development, will be developing a system -- design and system requirements, hope to do open source together. And some of those around what you expect a registration agent to do, some kind of metadata, of a certain domain type -- we can put that in. We will be doing it anyway.
- Need for communication and coordination. For matters of convergence, and learning efficiency, how do we support that type of exchange? Is that this group? During regular calls? Is that something a stream in the PID forum? Something that is not completely open as development stage. Not ready for public consumption but more internal IGSN group?
- GitHub as a communication tool? Would say within the context of the committee, that is a fine point to discuss who is doing what and who wants to collaborate, but the development work will be off line and there will be a shared GitHub with communications by the technical group not the IGSN committee.
- Another allocating agent mentioned they put up a platform for code contributions and would like to participate in discussion on shared infrastructure but do not always have the connections to identify where this can be done.
- Use of PID Forum for internal discussions. For high level, a discussion forum is better than GitHub. Different kind of communication.
- Question - OAI-PMH discussion from yesterday. Are you going away from that?
- One piece of the technology stack which is important, is the discovery part. Specify out how the current system could look like, to get feedback, if want to search for 500 million samples and metadata, and how do you do distributed architecture. What could be initial steps. Recommended connecting with someone from Pangea.
- How you would do discovery in a distributed environment? We defined a persona for sample identifier which might not be IGSN central. The question is who provides that service? It is an ecosystem, that not all the value is provided by IGSN.
- 1) organization IGSN looks to promote reuse and discoverability of samples and will mint identifiers, so will help with tracking, but not sure if mission statement would say provides or supports discovery. 2) how do you allow a community to own the information model for different needs.
- Comes with drivers that made GBIF build an index for millions of materials. And not sure the same drivers are there or needs for the same faceted approach. Might need to focus on the foundation and hooks that others can build what they want.
- If there is a gap, maybe the organization can take that on.

- Stresses the search index is an important piece, is central to DataCite. The resolver solves this. Do not see how this scales with just using databases. It is central and enables nice things. Needs a clear statement on how IGSN approaches this for many use cases. Handle system is not a good place for storing metadata.
- In outlining the responsibilities, and minimal product, we saw sample discovery as an essential service, but not necessarily something IGSN e.V. provides, but if no one does, we need to step up.
- Leaving out discovery would be a mistake. If people can't find stuff they are not going to use it. There is efficient technology for using indices that are turnkey/off the shelf. But is not part of the core infrastructure of IGSN, can be detached without breaking IGSN.
- Having someone else doing it, would we have to be more attentive to issues like restricted access? If we do it ourselves, is it easier to contain? Group consensus was no.

4. Next Steps and Actions for the Organizational Steering Committee

- A forum will be set up on the [PIDForum](#) to support continued discussion between workshops.
- The meeting notes from this workshop will be written up as a report and shared with both the Technical and Organizational Steering Committees.
- *TENTATIVE* -- The next TSC workshop will be a joint meeting with the OSC and will be held in conjunction with the [PIDapalooza](#) event in January 2020.

5. Items for the Organizational Steering Committee

- Review and finalize the Mission/Vision statement developed by this committee
- Questions
 - Base infrastructure, shared across different domains or different infrastructure for different domains?
 - Multiple handle prefixes or one?
 - Is the model presented here something we can agree on from a governance perspective?
 - Do we build on what we have - or have other handle namespaces for other communities?
 - Would like to see some of the things discussed here codified by the OSC.
 - Need early workflows developed.

End of Day Two

IV.APPENDIX

A. Participants

Project Steering Committee

- | | |
|--------------------|---|
| 1. Kerstin Lehnert | LDEO, lehnert@ldeo.columbia.edu |
| 2. Lesley Wyborn | ARDC / ANU, lesley.wyborn@anu.edu.au |
| 3. Jens Klump | CSIRO, jens.klump@csiro.au |
| 4. Sarah Ramdeen | Ronin Institute, sarah.ramdeen@gmail.com |

IGSN 2040 Technical Steering Committee

- | | |
|-----------------------|---|
| 5. Simon Cox | CSIRO, simon.cox@csiro.au |
| 6. Anusuriya Devaraju | MARUM, adevaraju@marum.de |
| 7. Doug Fils | Ocean Leadership, dfils@oceanleadership.org |
| 8. Jess Robertson | Unearthed Solutions, jess@unearthed.solutions |
| 9. Natasha Simons | ARDC, natasha.simons@ardc.edu.au |
| 10. Dirk Fleischer | University Kiel, dfleischer@kms.uni-kiel.de |
| 11. Donald Hobern | GBIF, dhobern@gbif.org |
| 12. Kirsten Elger | GFZ Potsdam, kirsten.elger@gfz-potsdam.de |

Guests

- | | |
|--------------------|---|
| 13. Dave Vieglaiss | DataONE, dave.vieglais@gmail.com |
| 14. Adrian Burton | ARDC, Organizational SC, adrian.burton@ardc.edu.au |

Observers

- | | |
|---------------------|--|
| 15. David Lescinsky | Geoscience Australia, David.Lescinsky@ga.gov.au |
| 16. John Morrissey | CSIRO IM&T, john.morrissey@csiro.au |
| 17. Joel Benn | ARDC, joel.benn@ardc.edu.au |
| 18. Julia Martin | ARDC, julia.martin@ardc.edu.au |
| 19. Alex Ip | Geoscience Australia, alex.ip@ga.gov.au |

Dial in/Remote

- | | |
|-------------------|--|
| 20. Wim Hugo | SAEON, wim@saeon.ac.za (remote participant) |
| 21. Martin Fenner | DataCite, martin.fenner@datacite.org |

Apologies

- | | |
|-------------------|---|
| 22. Xiaogang Ma | University of Idaho, max@uidaho.edu |
| 23. Ramona Walls | CyVerse, rwalls@cyverse.org (Dave Vieglaiss is replacement) |
| 24. Alex Hardisty | Cardiff University (DiSSCo), HardistyAR@cardiff.ac.uk |

B. Agenda

1. Day 1 (13 May 2019)

<p>09.00 - 12.30</p> <p>Including morning tea from 10.30 to 10.45</p>	<p>Workshop Opening:</p> <ul style="list-style-type: none"> • Opening and Welcome to Country • Housekeeping and logistics <p>Getting everyone on the same page for the Workshop:</p> <ul style="list-style-type: none"> • Intro to technical steering committee - scope, members • Overview of timelines for delivery • Intersection with Organizational Steering Committee • Workshop goals and objectives <p>Introduction to the governance, business models and architecture of:</p> <ul style="list-style-type: none"> • IGSN • DataCite (contribution by Martin Fenner) • GBIF • ORCID <p>Status Quo of IGSN Agents and Users:</p> <ul style="list-style-type: none"> • Preliminary Survey results (Sarah) • Currently active IGSN Agents (fill out spreadsheet) • Current technical challenges of the IGSN <p>Discussion of the Scope of IGSN:</p> <ul style="list-style-type: none"> • Big or small IGSN? Which domains should be included? Who is the target audience? Who are the clients? (Document by Alex Hardisty wrt DISSCo requirements) • Elements specific to the IGSN central registry • Elements specific to IGSN Agents - e.g. data models, APIs <p>Synthesize morning and decide structure for afternoon</p>
<p>12.30 - 13.30</p>	<p>Lunch</p>
<p>13.30 - 17.00</p> <p>Including afternoon tea from 3.00 until 3.15</p>	<p>Discussion of the:</p> <ul style="list-style-type: none"> • Unspecific elements (undifferentiated heavy lifting - what can we rely on an existing services platform to provide - e.g. data storage, networking, auth, deployments...) • Technology alternatives <p>Develop summary of challenges</p>
<p>18:00 - 21.00</p>	<p>Dinner</p>

2. Day 2 (14 May 2019)

09.00 - 12.30 Including morning tea from 10.30 to 10.45	Recap of Day one Breakout Groups: <ul style="list-style-type: none">• Technical Architecture<ul style="list-style-type: none">○ Architectural design principles○ Key workflows in IGSN○ Interfaces between the system components and actors involved○ Minimum viable product for the IGSN Registry/IGSN Agent services• Future Vision and Services<ul style="list-style-type: none">○ IGSN “What we do”○ Vision○ Mission○ Principles
12.30 - 13.30	Lunch
13.30 - 17.00 Including afternoon tea from 3 until 3.15	Continued work in Breakout Groups <ul style="list-style-type: none">• Report out from groups• Discussion Closing Discussion