P14 Helsinki Breakout Session
Physical Samples and Collections in the Research Data Ecosystem:

Meeting title: Designing a pathway towards Implementing a Transdisciplinary Data Infrastructure for Physical Samples

Group Page:

https://www.rd-alliance.org/groups/physical-samples-and-collections-research-data-ecosystem-ig

Session Page:

https://rd-alliance.org/ig-physical-samples-and-collections-research-data-ecosystem-rda-13-plenary

Session Presentations:

Meeting agenda:

1. 0-10 minutes (10 minutes): Introduction and tour de table, including a survey of which sample types and communities are present in the audience.
2. 10-15 (5 minutes): Time for audience participation to present on new projects for sample metadata schemes.
3. 15-30 minutes minutes (15 minutes): Review of the current status of the PhySC IG. Summary of the PhySC IG session at p13 (https://docs.google.com/document/d/1R_SpuKQ-rcMI81K66WqtculmQQrplgkay-UfD8uPZ-g/edit)
4. 30 - 40 minutes (10 minutes) Community updates from the room
5. 40-70 minutes (20 minutes): Concurrent Breakout sessions on:
    a. Business models for sustaining identifier systems for samples - what are the key considerations?

https://docs.google.com/document/d/1i1BcDh2LtwuYNufryPYQUWl_M4ZdZqlW8DDLcxf3T7I/edit?usp=sharing

b. Review of the current core metadata schema - is it appropriate for your domain of interest? https://docs.google.com/document/d/1l4QyaoduPezUlNKJYhaw0Yip4VdsUuzO2vBD1R2yVFM/edit?usp=sharing

6. 70-75 minutes (5 minutes): report back
7. 75-85 (10 minutes) New leaders/co-chairs
8. 85-90 minutes (5 minutes): Wrap-up and next steps.

Session Notes:

Please add your contact information to the attendees list at the bottom of this page.

Introductions

Introductions by Co-chairs. Kerstin Lehnert - moved from lab to work with digital aspects. Want to bring people together to talk about physical samples in the digital data ecosystem.
Lesley Wyborn - Interest in being able to identify physical samples and link them to literature and analysis.

Introductions around the room (*see table below for details on attendees*).

Brief change to the agenda, to combine the audience participation with the breakout session, to focus more on metadata than perhaps the business model.

At the end of the meeting, we will take some time to discuss a change in leadership. Looking for someone from Europe, or other regions outside of the US/AUS and other sample types besides Geo samples.

Overview/current status

Review of the large number of institutions and examples of various samples, specimens, and artifacts that get used in research. Wide range of topics related to samples.

“Internet of Samples” - connecting physical samples to the digital research data ecosystem. Need a digital representation, and need to link it to the object. Persistent identifiers are part of that system. Also need to capture and link essential information about these objects. What are the essential data/metadata we need to capture and make available? THe identifiers also need to be integrated into the larger ecosystem like publications, other identifiers like for lab equipment etc. Also want to integrate all the information that is created over time about the samples, to be able to do data analytics.

This fits into FAIR -- making relevance entities findable, making metadata reusable, making samples interoperable.

History of the group. Endorsed in 2017. Had meetings at the past few plenaries. List of topics discussed. (See slides)

Drivers for the interest group -- originally linking data and publications. Glad to see a diverse range of disciplines gathered today. And would like to know what is happening in your world, is your community using identifiers? Can talk about earth science and the expansion of the IGSN beyond that community.

Diagram (see slides) showing linkage between identifiers to build a system or infrastructure to allow you to locate, access, link, and cite samples.

Introduction to IGSN -- founded in 2007, modeled after DataCite, has a handle system. Has a custom profile for physical samples, and incorporates workflows on physical samples. 8 allocating agents world wide. Not the only identifier used for physical samples. Working to coordinate and align with other identifier systems to see how IGSN can serve other communities that do not have an identifier system.

IGSN 2040 -- current effort funded by the Sloan Foundation to create an organizational structure for the IGSN to gain long term sustainability and to update the technical architecture to be more scalable.

Summary of P13 session
Increasing diversity of science domains represented, and diverse stakeholders (See slides for list).

Questions that came up in discussion last time -- concerns and questions regarding best practices. Procedures, policies, how do I implement IGSNs, DOIs, etc. and make them work for my specific situation. What catalogs are out there. How is Sample provenance and its evolution tracked over time? How are samples cited? How do we link to other identifiers? What about certification (CoreTrustSeal, etc.) for sample catalogs?

Priorities identified -- to gather use cases. Not sure what is out there in regards to workflows, handling, field work, lab work, scenarios of preservation. One of our challenges is the sample metadata. Need a core set of metadata for all samples so we can exchange information about samples. Increasingly finding that samples cross domains.

Diagrams by Lesley (see slides) on the common kernel. For Geo, it is location, but for other fields that is not the common important fields. We want a slim, sleek core of metadata. Each circle in the diagram is meant to identify another vocabulary. Marine want grain size, that is also

wanted in material or geological samples. Moving to a linked data world, with schemas for each one, if we can have a series of uniquely identified schemas, people can mix and match. Going forward, where linking samples together, being able to do it and use the power of linked data, having known vocabularies for grain size, color, name, etc. we can create a global network of identifiers.

What is happening in your neck of the woods? Are there standards in your community?
-- We created our own standards for bacteria samples. We have analysis, results, and need to use right away. There was a huge problem with how government orgs shared data. Worked with EPA to make it easier to share mass amounts of data.
-- Standards do not exist in archeology, not shared, dysfunctional, everyone makes their own. Do you see a move towards a shared? People say they would be interested but not much movement. Archeologists have samples that come from other fields, soil, etc. and the easy path might be to adopt another than use their own. Samples for paleoecology, if it has been done, we would like to adopt it. *Availability of tools and how they might evolve if there is a standard underneath.*
-- Define what you mean by standards? Darwin core is one standard. Communities have their own conventions that may not use standards, but common vocabulary. It is important to define what you are talking about. And maintain adherence to a standard at the basic level, and flexibility for community needs.
-- There are plenty of domain specific vocabulary, would be useful to use registries, which have governance. (RDF, machine readable that link to the definitions, follow SKOS). INstead of creating your own vocabulary.
-- Looking at how other groups develop standards, looking at crystallography, this is all time controlled, and you can look at the data and the time stamp to see what vocabulary and standard was used at that time. Need other groups to come on board with these types of methods. As vocabularies are dynamic, as more words are needed to describe.
--Ontologies and metadata are governed by scientific unions, but samples are a cross domain topic. Who is out there who can manage those aspects?
-- Need to know what others are doing. IGSN is doing that but might need to do more, that is the driver. Need to know the particular domain, but it is hard to sell it for adoption. As there are so many flavors, but used sparingly because it is a lot of work. Not an easy task, and how it can be more manageable.
-- Some groups have it very well defined other groups only have 5 words and do not like a controlled vocabulary. But when you need to do machine learning and data mining, you have problems with the different ways of spelling things. Can we have a set up with version controlled vocabularies we can point people to and ask if people are putting them here?
-- Asked about material sciences, if samples are identified? Physics had some questions about new compounds, and if it is safe -- health and safety concerns. Someone else is new to material sciences, and asked about standards in their community, but they do not have them, where do they start? Available standards to use? IGSN was a way to have good documentation of the sample, and connect it with a good identifier that could be linked in other areas. The unique identifier gets tracked as it is being worked with, and linked to data, samples, people,

etc. That has led to other communities being interested in working with IGSNs. Worth investigating if it can be scaled for other communities.
-- When researchers get samples from industry, they buy them. They do not have identifiers. They do not know which forest the samples are coming from. RRIDs, research resource IDs used in the medical sciences. It seems to be used for chemicals used in experiments in the lab, something to look into. Maybe next RDA we get someone from that community. It is life science, neuro science, antibodies, that are expensive, and mice strains. So really need to be able to identify the thing. Close to samples, related in some was.
-- Making internal nomenclature, but will map it to what the repository requires, and if different repositories, might have to map it to different ontologies multiple times. Will have many different forms.

Research Vocabularies Australia
https://vocabs.ands.org.au/
Can search on concept or word, allows for discovery. Allows people to search by concept, and what vocabulary it is in.

-- Common vocabularies is not just about samples. No one has found a solution, but that is not just about samples. How do you make them work together, where do you list them, etc. The hopefully is some group somewhere talking about this.
-- The idea to enrich metadata by domain vocabularies would be helpful for data sets. It is often more work to look for the vocabulary to use than inventing one's own. Using registered vocabulaires that link in each field is a key step in interoperability and machine reading. Registered vocabularies in RDF is one way to do this.

Final business
Biggest concerns with respect to samples? With so many new faces, need to think about how best to run the sessions as opposed to a continuum.
-- Evolution to a working group? We tried in Berlin, but than the IGSN 2040 project came along, and we did not have the bandwidth to do both. But if people in this room have the willpower and bandwidth, we can have working groups. Working groups must be complete in 18 months max which is a challenge.

Next plenary - half day workshop? For discussions and developments? Also is there anything this group can do between the plenaries? How can we bring continutinety into the discussions of curation of physical samples?

IGSN 2040 project
Defining the future of the IGSN. Would like to engage this community in participation. We can run webinars if needed. Looking to see if IGSN can scale for other communities. But we do not currently have the capacities to extend the services drastically. We have two steering committees, there are many in the room from them. One technical and one organisational. We have had two meetings, have plans for two more. The driver is that the IGSN system has

grown. There are talks in communities about it getting from millions of samples to billions. Mentioned DIsSCO as an example partner. How do we grow the system to accommodate this scale.

Sample registration often comes in bursts, field campaigns, collections being registered at once. tHere are more appropriate systems for managing this behavior. Reviewing prototype of an architecture that can better support this.

Other aspect, current IGSN architecture is to have allocating agents who use OAI-PMH to expose metadata, but that does not scale well. So we are talking about approaches with schema.org and site maps to allow better performance.

iSamples -- Internet of Samples (iSamples) - Towards interdisciplinary cyber infrastructure for material samples. Bio, archeologist (open context), and earth sciences to develop a system which is better scalable to provide services to smaller communities and take into account different entities may use different identifiers for samples (See diagram on slides). To have a registry that works with any of these identifiers to support connections and common metadata profiles and schemas. The technology would be to have the iSamples in a box, which lets reappen across domains and in a shared registry.

Who would like to volunteer to be actively involved in this interest group? Please send us an email to let us know. Would like to have more representation from Europe and Asia, and material sciences, bio, agriculture, etc. diverse a group as possible.

Breakout two would be an excellent working group. Probably many local groups working on this in domain specific areas.

What else to do -- use cases, collecting them is something that can be done in the spaces between sessions, to see which communities see the most need to structure the work along that. This group has not formally started collecting use cases. iSamples from EarthCube had started a collection of use cases, this could be expanded. This is something where the infrastructure for IGs in RDA lacks a way of gathering use cases. Maybe we can set up a way to capture them.

We also had webinars where different groups could talk about their systems and challenges. Perhaps that is another thing we can start doing.

Action items
Vacancy for co-chair
Sign up for group to get on listserv
Develop method for capturing use cases (google form?)

Pitch for unconference session on capturing information about field data workflows.

Attendees:

(about 40+ in the room)

	Name	Object of interest	Organization	Email address
1	Lesley Wyborn	Granites specifically, but any rock and mineral samples	ARDC, AuScope, IGSN	lesley.wyborn@anu.edu.au
2	Kerstin Lehnert	Physical samples for research (geo, bio, arch, etc.)	IGSN/SESAR	lehnert@ldeo.columbia.edu
3	Sarah Ramdeen	Physical samples for research (background in geo - cores/cuttings)	SESAR	sramdeen@ldeo.columbia.edu
4	Parvaneh Abbaspour	Physical samples for research (materials chem, animal tissue)	Lewis & Clark College	parvaneh@lclark.edu
5	Tina Dohna	Physical samples for research (bio, geo, cryo)	PANGAEA MARUM/AWI	tdohnaqmarum.de
6	Zachary Trautt	Materials Science and Engineering Samples	NIST	zachary.trautt@nist.gov
7	Ana Slavec	Samples of wood and other renewable materials (biopolymers)	InnoRenew CoE	ana.slavec@innorenew.eu
8	Alberto Stella	Insects / Soil	Rothamsted Research	alberto.stella@rothamsted.ac.uk
9	Jill Benn		University of Western Australia	jill.benn@uwa.edu.au
10	Steve Collins	Materials Science and Engineering Samples plus Bio	Diamond Light Source	steve.collins@diamond.ac.uk
11	Gabrielle Parent-Doliner	Recreational water quality - indicator bacteria - bio	Swim Drink Fish Canada	gabrielle@swimdrinkfish.ca
12	Tobias Richter	Materials Science and Engineering Samples plus Bio	European Spallation Source	tobias.richter@esss.se

13	Charles VardemanAdela Sobotkova	Archaeological samples (bio, geo, everything imaginable)	University of Notre DameAarhus University | FAIMS	cvardema@nd.eduadela@cas.au.dk
14	Cynthia Love	Soils archive, viruses, insects, wildlife, herbarium, fish, rockstore, air	CSIRO	Cynthia.Love@csiro.au
15	Rolf Krahl	Divers, whatever our users bring in	HZB	rolf.krahl@helmholtz-berlin.de
16	Anne S. Bergsaker	Geophysics, IT for research, data analysis, data collection management	University of Oslo	a.s.bergsaker@usit.uio.no
17	Shawn Ross	Archaeological samples; field data collection software	Macquarie University, Sydney | FAIMS	shawn.ross@mq.edu.au
18	Lindsay Powers	Geological and biological	USGS	lpowers@usgs.gov
19	Esther Plomp	Archaeological artefacts, rocks, animal/human tissues, chemical/biotechnical	TU Delft	e.plomp@tudelft.nl
20	Minna Harjuniemi	IT services view: interested in which service we need to offer to researchers	University of Helsinki / IT center	minna.harjuniemi@helsinki.fi
21	Yasuhiro Murayama	Advocacy and coordination of Japanese research institutions	NICT, Japan	murayama@nict.go.jp ymurayama2003@gmail.com
22	Erin Robinson		ESIP	erinrobinson@esipfed.org
23	Damian Ulbricht	biological, liquids, rock, sediment, soil	GFZ Potsdam	ulbricht@gfz-potsdam.de
24	Adam Leadbetter	Marine and freshwater biological samples	Marine Institute, Ireland	adam.leadbetter@marine.ie
25	Tovo Rabemanantsoa	Soil and plants samples	INRA, France	tovo.rabemanantsoa@inra.fr
26	Kirsten Elger	biological, liquids, rock,	GFZ	kelger@gfz-potsdam.de

		sediment, soil	Potsdam	
27	Otto Lange	Rock, library artifacts (letters, annotated print, wood blocks), medical	Utrecht University Library	o.a.lange@uu.nl
28	Jessica Rex	Materials Science, chemical/biotechnical samples	Technical University Ilmenau	j.rex@tu-ilmenau.de
29	Doug Fils (online)			
30	Vasily Bunakov (online)			
31	Yunato Kao (online)			