PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

11–13 June, Prague, Czechia

From Collection to Citation: The Diverse Actors and PIDs Needed for Comprehensive Material Sample Management

Jens Klump
CSIRO, Mineral Resources
IGSN e.V.

Rorie Edmunds
DataCite

Andrea K. Thomer
University of Arizona, School of Information

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiativ

ARDC
Australian Research Data Commons

Agenda

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

- Introduction to session
- ESIP sample citation guidelines: Identifying obstacles to PID use and adoption
- Panel discussion:
  - *What are the missing obstacles, actors, and institutions in this diagram?*
  - *what are the problems?*
  - *What are the different “domains of responsibility” and what metadata is contextual/tacit vs formal?*
  - *Can shared metadata be used to enrich the context vs dumbing it down?*
  - *Do the APIs enrich metadata or does it reduce it to lowest common denominator?*
- Wrap up: Next steps? Takeaways?

Introduction to session

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

What are the missing links, obstacles, actors, and institutions?

Who are the actors and institutions?

What are the problems?

What are the domains of responsibility (Treloar & Klump 2019) and what metadata is contextual/tacit vs formal

Can shared metadata be used to enrich the context vs dumbing it down?

Do the APIs enrich metadata or does it reduce it to lowest common denominator?

Wrap up: Next steps? Takeaways?

Background: ESIP Physical Sample Citation Cluster

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

- ESIP: Earth Science Information Partners
- Cluster goals: to better support the curation, reuse and sharing of physical sample collections
- Me: Andrea Thomer, Associate Professor of Information @ U Arizona, SESAR product manager, Cluster co-chair
  - Fellow co-chairs: Joan Damerow, Val Stanley
  - Community fellow: Natalie Raia

The challenge: finding sample-related resources and tracking sample use

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

- Researchers don't usually reference/cite samples in papers
  - They also often leave sample PIDs out of datasets!
- When they do, they often don't use PIDs…
  - Instead, they use their own IDs or non-unique catalog numbers
- …and sample references are buried in inaccessible places
  - Supplementary materials
  - Acknowledgements
  - Figures that can't be scraped
- Some fields use 100s-1000s of samples per paper, which = a lot of work to reference
- Even when PIDs are used, there's limited infrastructure to compile citations

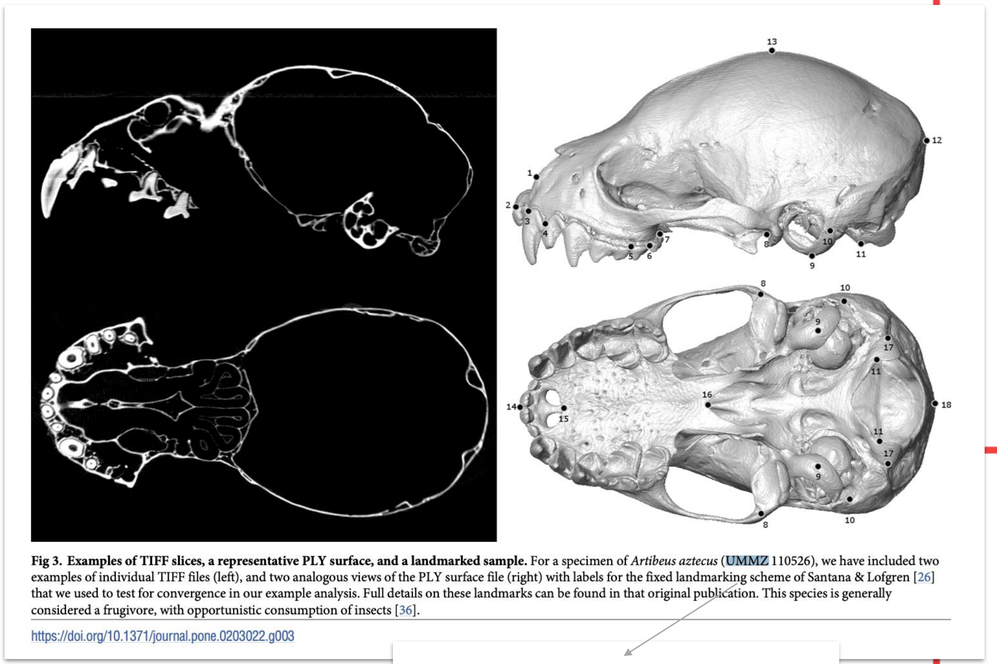


Fig 3. Examples of TIFF slices, a representative PLY surface, and a landmarked sample. For a specimen of *Artibeus aztecus* (UMMZ 110526), we have included two examples of individual TIFF files (left), and two analogous views of the PLY surface file (right) with labels for the fixed landmarking scheme of Santana & Lofgren [26] that we used to test for convergence in our example analysis. Full details on these landmarks can be found in that original publication. This species is generally considered a frugivore, with opportunistic consumption of insects [36].

https://doi.org/10.1371/journal.pone.0203022.g003

us aztecus (UMMZ 110526), we have in
d landmarking scheme of Santana & Lo
ginal publication. This species is general

PID

The challenge: finding sample-related resources and tracking sample use

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

- Decreases reproducibility of research and reusability of samples.
- We can’t consistently link sample metadata records to the papers, datasets, and other research products that are created using analysis of samples
- We can’t give collections, collectors, funders, and others involved in sample curation and funding the credit they need for their work
- Samples are not truly FAIR – which means it’s more challenging (if not impossible) to apply CARE guidelines

ESIP Physical Sample Citation Guidelines

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

- Partial solution: guidelines to encourage standardized use of sample PIDs in papers
- Outputs:
  - **Flyer**: 4 Steps to Publish and Cite Open Earth Science Samples https://doi.org/10.6084/m9.figshare.24291148.v1
  - **Checklist**: A Scientific Author Guide for Publishing Open Research Using Physical Samples https://doi.org/10.6084/m9.figshare.24669057.v1
  - **Paper**: Opening Doors to Physical Sample Data Discovery, Integration, and Credit
  - https://eartharxiv.org/repository/view/7194/

4 STEPS TO PUBLISH OPEN EARTH SCIENCE SAMPLES

Increase the impact of your research! Citing samples in publications and sharing sample metadata enables scientific discovery, ensures credit is given to sample curators, and enhances sample data reuse and reproducibility. Take these four steps now to support open science for sample-based research and streamline your publication workflow:

1. DESCRIBE SAMPLES WITH RICH METADATA

Metadata is "data about data" – it's the information needed to find and reuse a sample. Create a spreadsheet that describes basic characteristics of your samples (e.g., material, geographic coordinates). The System for Earth Sample Registration (SESAR) helps you create a template for geoscience samples. Take time up front to maximize the information you provide: it will save you time when reusing or sharing samples in the future. Remember that information that might be trivial or obvious to you (e.g., sample was collected with a hammer) may be important to a future researcher.

2. ASSIGN OR USE IDENTIFIERS FOR SAMPLES

Persistent identifiers (PIDs) are unique numbers that can be assigned to a sample and then used to store and retrieve information about that sample.

If you collect your own samples: Upload your sample metadata to SESAR to register them for persistent identifiers (PIDs), specifically International Generic Sample Numbers (IGSN IDs). Each registered sample will have an associated, openly accessible landing page containing its metadata; SESAR provides printable sample labels with a QR code linking to this page. You can then start building your personal sample catalog, and make it easier to quickly share or update metadata about your samples (e.g., add related links over time). Even if a sample has already been published, you can still register it with SESAR now!

If you are using samples from an existing collection: Ask the sample owner or repository manager how you should acknowledge the collection and what PIDs or other identifiers to use to reference samples in your paper and datasets. Be aware that sub-samples may require a different PID from the "parent." This makes it possible for repositories to show the impact of their collections and enables future researchers to locate samples.

3. PUBLISH AND CITE DATASETS WITH SAMPLE PIDS

Publish data associated with your samples, ideally in a domain or institution-specific data archive. To ensure samples are linked to datasets, include sample PIDs in your dataset metadata and designate a column in each file/table for your sample PIDs, even if you use a related sample name. Then cite the dataset in the reference section of your paper.

4. REFERENCE SAMPLES IN YOUR PAPERS USING CONSISTENT FORMATTING

When referencing PIDs in text, files/tables, and data availability statements, include a prefix identifying what kind of PID you are using before writing the number, and a hyperlink to the sample landing page – for instance, igsn:10.58052/IENHR0000. This will make your PID findable by both humans and computers.

Author Guidelines for Publishing Open Research Using Physical Samples

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

Step 1. Describe Samples with Rich Metadata

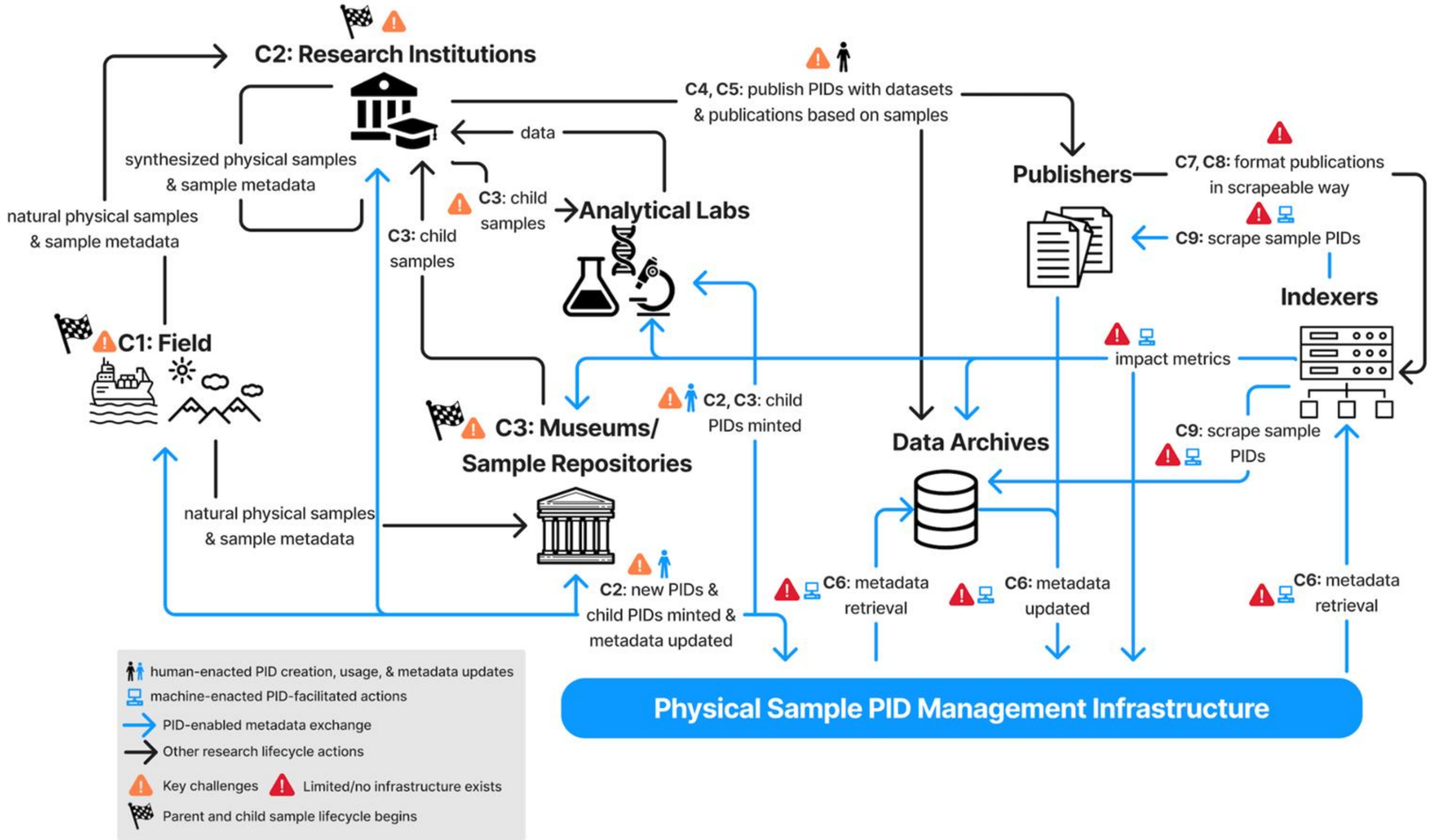
Step 2. Assign and/or Use Identifiers for Samples

Step 3. Publish and Cite Samples in Datasets

Step 4. Reference Sample Identifiers in Paper

https://doi.org/10.6084/m9.figshare.24669057.v1

Identifying obstacles to PID use and adoption



Panel discussion

Jens Klump

Samples and physical objects are at the heart of many research disciplines.

Samples are anchors for knowledge graphs into reality.

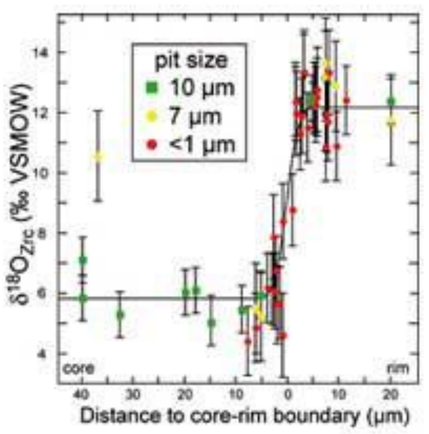
“The really exciting part is not about putting labels on things, but about what you can do when you put machine learning to work on the labelled data.”

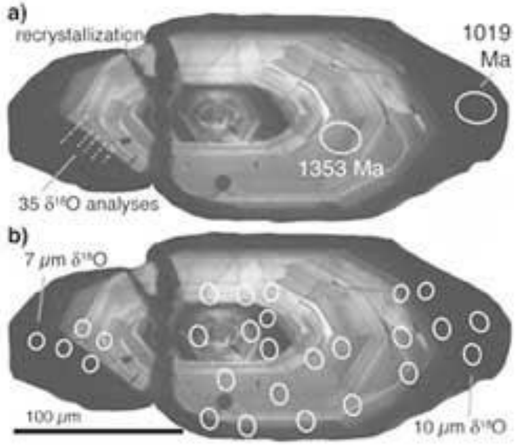
Jens Klump

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

The scenario:

- Samples get collected in the field
- Sample preparation
- Analysis (perhaps in another lab at another institution)
- Data get processed
- Data get incorporated into a larger dataset
- The dataset is interpreted in a paper in the scientific literature





Cody Ross

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

DataCite

CONNECTING RESEARCH, ADVANCING KNOWLEDGE

Application Support Engineer

Representing the **Material Sample PID Management Infrastructure**

DataCite provides the **PID infrastructure** and **metadata catalog** for **IGSN (International Generic Sample Number) IDs**.

IGSN IDs are a globally unique and persistent identifier for material samples, aggregations of samples, features-of-interest or collection sites, and destroyed, discarded, or ephemeral samples.

Cody Ross

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

14,602,617+

DataCite DOIs have been registered for Physical Objects

12,601,269+

are IGSN IDs registered in DataCite systems and the DOI infrastructure

23.5%

of the DataCite metadata catalog is Physical Object identifiers

IGSN ID registration services were migrated to DataCite systems as part of the **IGSN–DataCite Partnership** established in 2021, embedding material sample PIDs into the DataCite PID Graph and metadata catalog.

DataCite and the IGSN e.V. continue to collaborate on **sample community engagement** to foster the global adoption, implementation, and use of physical sample identifiers.

These activities include the facilitation of Communities of Practice, as well as providing guidance on sample PID metadata, the role of PIDs in sample workflows, and the representation of sample identifiers in academic publishing.

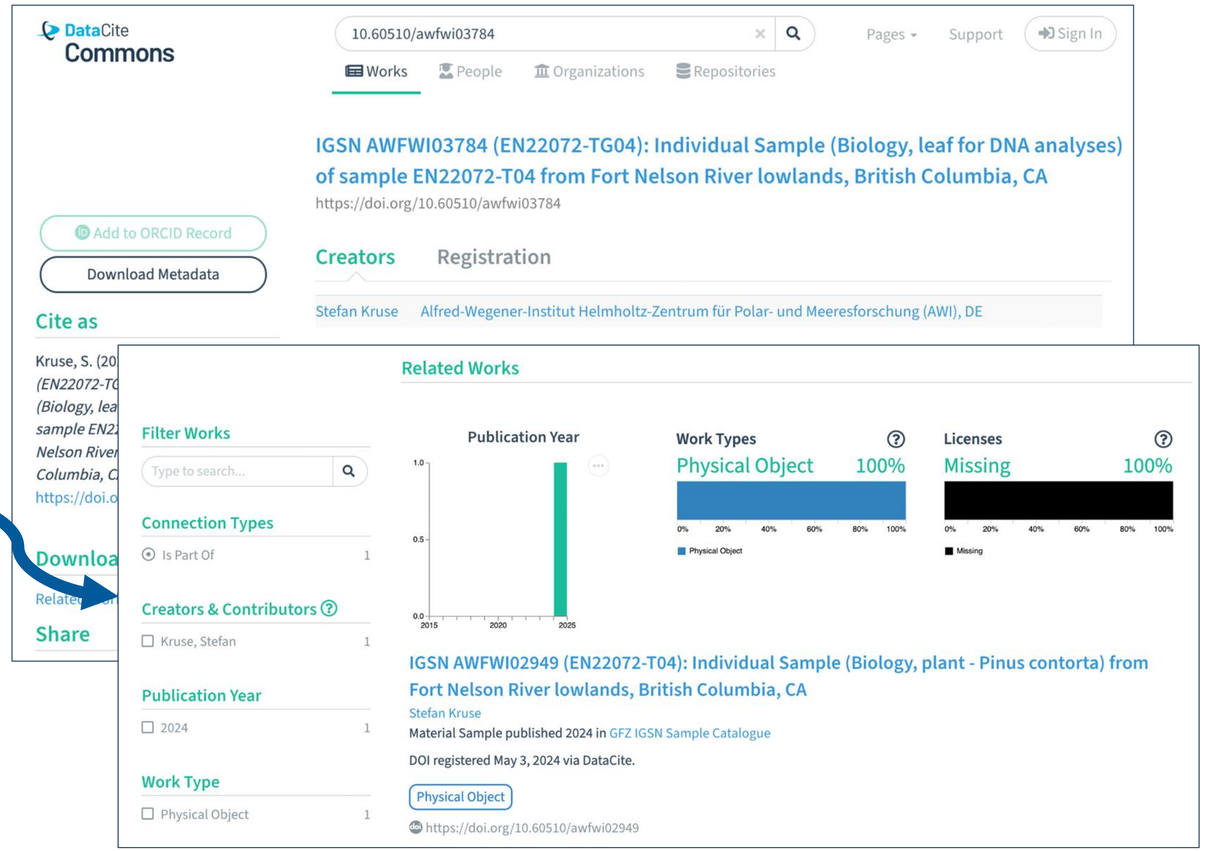
Cody Ross

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

IGSN IDs and their relationships to other PIDs are represented in **DataCite Commons**—DataCite's discovery service—as well as in DataCite's **REST** and **GraphQL APIs**.

DataCite relationship metadata

<relatedIdentifiers>
<relatedIdentifier relatedIdentifierType="DOI" relationType="IsPartOf">10.60510/AWFWI02949</relatedIdentifier>
</relatedIdentifiers>



Simone Frenzel

IGSN@ GFZ Helmholtz Centre Potsdam

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

TEAM GFZ Data Services
Geosciences Data Publisher

contact: igsn@gfz-potsdam.de

- GFZ is a founding member of IGSN e.V. (2011)
- One of ten allocating agents since 2016
- Responsible for the central Handle server until December 2022
- No central database at GFZ: from the long tail to databases different approaches for IGSN registration

ICDP = International Continental Scientific Drilling Programme

mdis = mobile drilling information system

Medusa: Online sample management system for DFG Prio-rity Programme 1803

Make stones out of the nature
EarthShape
Earth Surface Shaping by Biota
A German-Chilean Research Initiative

Sample description tables from **MS Sonne 273 Cruise**

SO273 Cruise

Type	Dredge/ROV (Dive)	#	Sample Name	IGSN	Website	Gesamtheit Proben
Dredge	11	8	SO273-11D-8	GFSO273C29	http://igsn.org/GFSO273C29	120
Dredge	11	9	SO273-11D-9	GFSO273C30	http://igsn.org/GFSO273C30	121
Dredge	11	10	SO273-11D-10	GFSO273C31	http://igsn.org/GFSO273C31	122
Dredge	11	11	SO273-11D-11	GFSO273C32	http://igsn.org/GFSO273C32	123
Dredge	11	12	SO273-11D-12	GFSO273C33	http://igsn.org/GFSO273C33	124
Dredge	11	13	SO273-11D-13	GFSO273C34	http://igsn.org/GFSO273C34	125

IGSN LANDING PAGE

Recent Activities @ GFZ

FAIR WISH

HMC HELMHOLTZ Metadata Collaboration

HMC Project Cohorte 2020

FAIR Workflows to establish IGSN for Samples in the Helmholtz Association

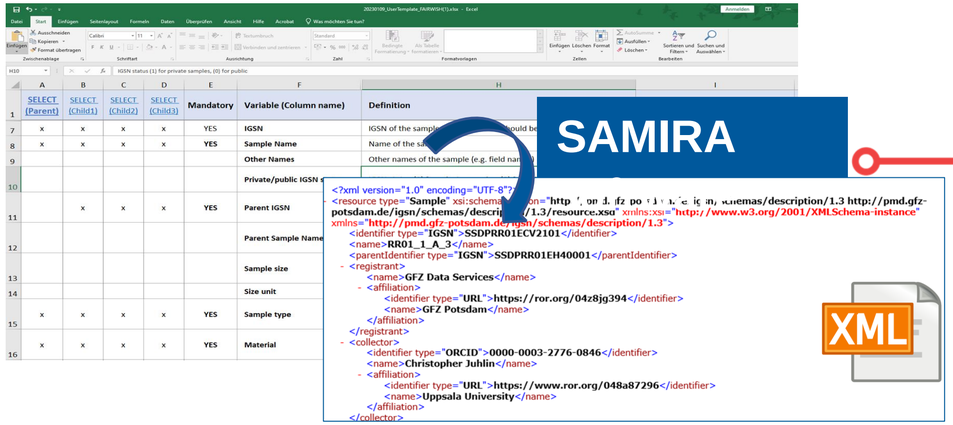
FAIR WISH Results:

- Full **documentation** of IGSN Metadata Schema
- **Metadata profiles** for additional sample types in the Geosciences (soil, water, vegetation, rocks) complementing core IGSN schema
- Recommendation of **controlled vocabularies**
- Workflows to **generate machine-readable IGSN metadata** from different states of digitisation and **automatically register IGSNs** (SAMIRA)

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

FAIR SAMPLES Template

The **FAIR Samples Template** is a customisable template to provide standardsised sample descriptions (XML). **Users can identify metadata elements from the full schema** that align with their samples.



Template: https://doi.org/10.5281/zenodo.7520015
Documentation: https://doi.org/10.5281/zenodo.7147531
Video Tutorial: https://www.youtube.com/watch?v=SfIbTUtknNo
SAMIRA: https://doi.org/10.5880/GFZ.LIS.2023.001

SAMIRA Sample IGSN Registration Automation

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

INPUT

- **FAIRWISH Sample Template (csv)**
- **XML Request form GFZ Lab systems**
- **XML from Parnters**
- **FAIRWISH Sample Template (Spreadsheet)**

SAMIRA:
- reads various Inputs
- create IGSN Objects
- writes XMLs IGSN/Datacite
- communicate via REST

Process

- **IGSN Registration Metadata Object**
- **IGSN Descriptive Metadata Object**
- **IGSN Specific Metadata Object**
- **Datacite Metadata Object**

SAMIRA:
- simplifies XML Generation in PHP
- bridges the IGSN LEGACY System
- serve as foundation GFZ IGSN Core

Output

- **XML**
- **Database entries**
- **json**

SAMIRA:
- will create various output
- will be better on next version (end of May)

Legend

- **implemented**
- **future task**
- **in progress**

Communication

- **DataCite**
- **IGSN Landing Page**
- **IGSN Sample Catalog**
- **IGSN OAI-PMH**

Frenzel, Simone (2023): FAIR WISH Software Tool: SAMIRA: FAIR SAMPLES Template Processing. V. 1.0. GFZ Data Services. https://doi.org/10.5880/GFZ.LIS.2023.001

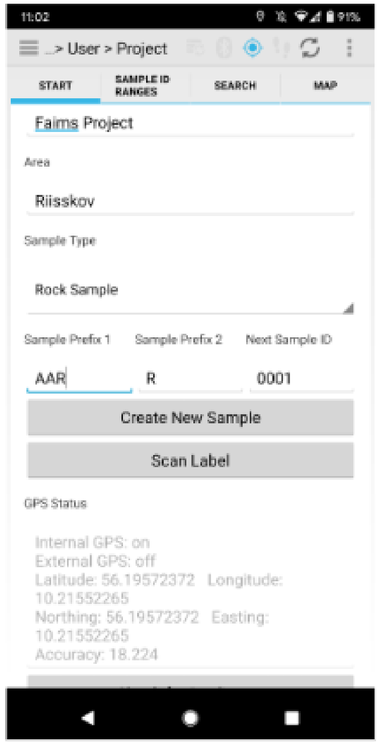
Shawn Ross

Co-Director FAIMS Project; field archaeologist

IGSN in FAIMS Mobile

Field data collection mobile applications

- FAIMS intends to make field data more FAIR, including in-field PID allocation
- IGSN implemented in CSIRO Geochemistry workflow in 2016
- FAIMS feature set was expanded to include:
  - **QR label scanning** and loading of records, and
  - label printing via a **portable printer**
- FAIMS3:
  - **QR label scanning**
- Opportunities for implementation in geochemistry and other workflows in Fieldmark notebooks



Original FAIMS interface for scanning pre-printed IGSN QR codes. Source: https://github.com/FAIMS/CSIRO-Geochemistry-Sampling

Shawn Ross

Product Manager, RAiD; Manager, NII Products

https://raid.org/10.1234/j6k28

Title (primary) Lorem Ipsum Project
Title (acronym) LIP
Start date 2023-01-30
Description (short) Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas vitae condimentum nisl, eget ornare felis.

Principle investigator orcid.org/0000-0002-3843-0000
Role https://credit.niso.org/contributor-roles/conceptualization/
Co-investigator orcid.org/0000-0002-3639-2080
Role https://credit.niso.org/contributor-roles/data-curation/

Lead organisation ror.org/00rqy9422
Partner organisation ror.org/03b94tp07

Grant doi.org/10.8948/908234D93EAF
Dataset doi.org/10.1594/PANGAEA.726855
Article doi.org/10.1038/nphys1170
Instrument doi.org/10.1337/jdlc-tima
Sample XXAB000IH

Local storage uq.edu.au/114/32
Cloud storage 79.152.127.243

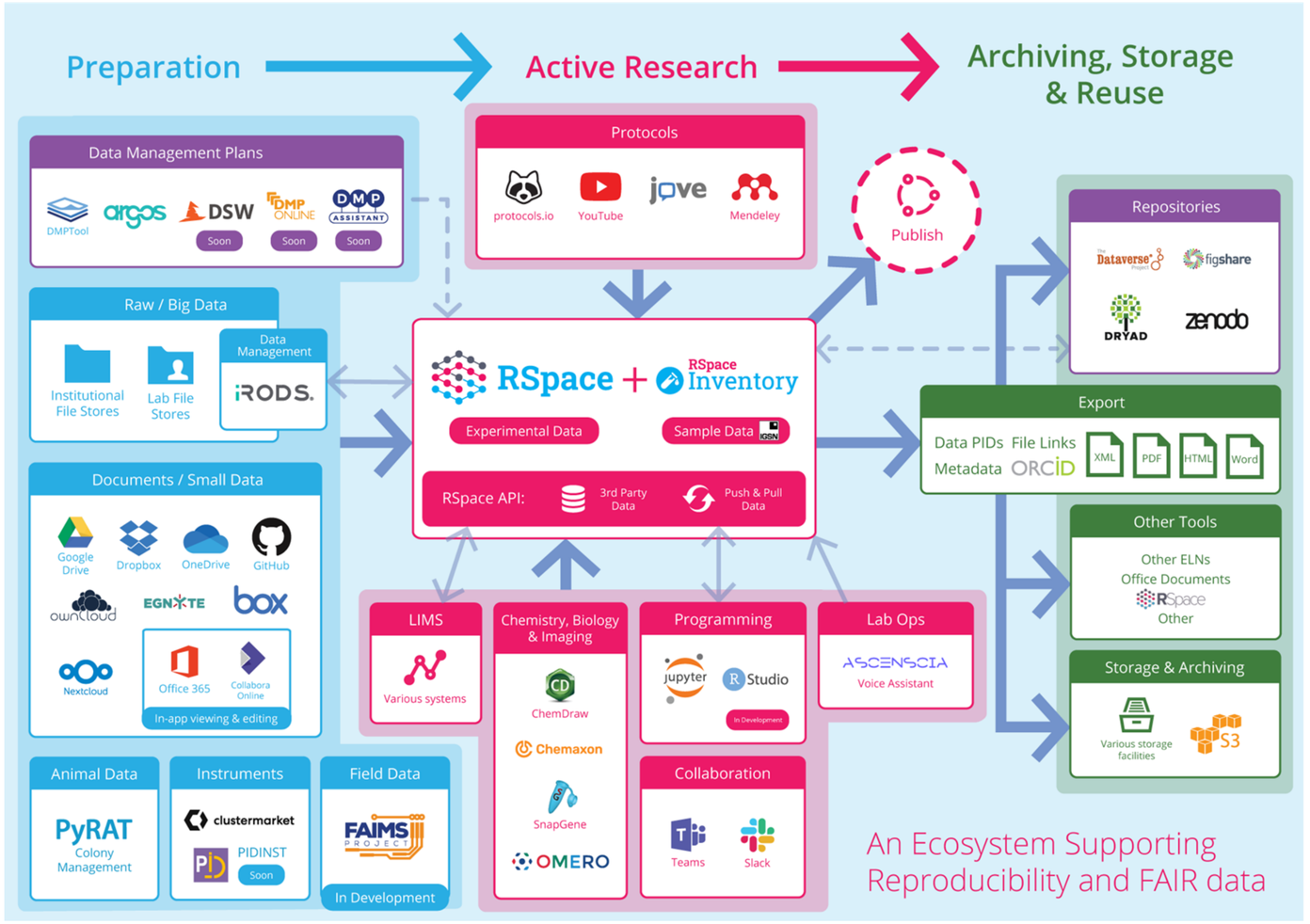
Sub-project https://raid.org/10.3010/401XQPOI

Tilo Mathes

Product Manager and Open Source Lead

RSpace

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024



https://www.researchspace.com | https://github.com/rspace-os | https://www.linkedin.com/in/tilo-mathes/ | https://orcid.org/0000-0002-9694-748X

Tilo Mathes

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

Recent work

RSpace researchspace.com

Project Groups

ROR

Customer Requests

Interoperability Guidelines

doi.org/10.5281/zenodo.8284206

OMERO

IGSN

DataCite FIND, ACCESS, AND REUSE DATA

Sample Workflows with IGSN IDs

Controlled Vocabularies

Performance

Outreach

Usability

RDM

Tilo Mathes

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

Upcoming work

RSpace researchspace.com

PIDINST

FieldMark

iRODS RDM front-end

API

RAiD Projects

June

Open Source Chemistry

Open Sourcing RSpace

UI & UX

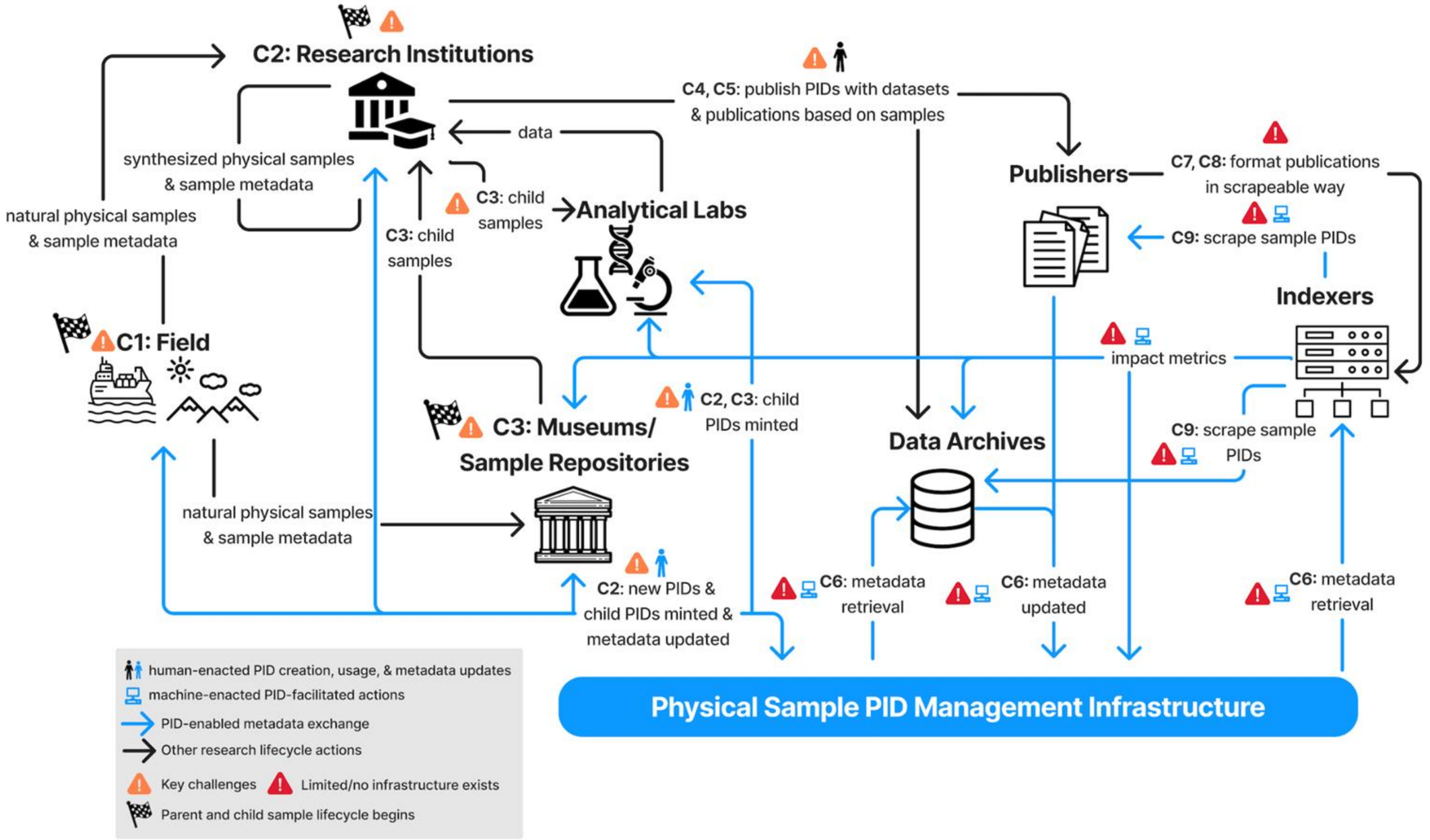
Enabling rich metadata

Collaborations

FDOs

Community

Dataverse


C2: Research Institutions
C4, C5: publish PIDs with datasets & publications based on samples
data
synthesized physical samples & sample metadata
natural physical samples & sample metadata
C3: child samples
C3: child samples
Analytical Labs
Publishers
C7, C8: format publications in scrapeable way
C9: scrape sample PIDs
Indexers
C1: Field
impact metrics
C2, C3: child PIDs minted
C3: Museums/ Sample Repositories
Data Archives
C9: scrape sample PIDs
natural physical samples & sample metadata
C2: new PIDs & child PIDs minted & metadata updated
C6: metadata retrieval
C6: metadata updated
C6: metadata retrieval
human-enacted PID creation, usage, & metadata updates
machine-enacted PID-facilitated actions
PID-enabled metadata exchange
Other research lifecycle actions
Key challenges
Limited/no infrastructure exists
Parent and child sample lifecycle begins
Physical Sample PID Management Infrastructure