An Astoundingly Brief Primer on Persistent Identifier Context, Orgs, Libraries, and Open Infrastructure

Mike Nason
Scholarly Communications and Publishing Librarian, UNB Libraries | Crossref and Metadata Liaison, PKP

An ~~Astoundingly~~ ~~Brief Primer~~ on Persistent Identifier Context, Orgs, Libraries, and Open Infrastructure

Mike Nason
Scholarly Communications and Publishing Librarian, UNB Libraries | Crossref and Metadata Liaison, PKP

Taking Our Time With Persistent Identifier Context, Orgs, Libraries, and Open Infrastructure

Mike Nason
Scholarly Communications and Publishing Librarian, UNB Libraries | Crossref and Metadata Liaison, PKP

Before we dig in, I need to make one thing super clear…

PIDs are in the drinking water of scholarly publishing.

Let’s review some things we know.

PIDs are unique IDs that we assign to an increasing number of things:

- Institutions
- Datasets
- People
- Organizations
- Articles
- Monographs
- Serials

PIDs are unique IDs that we assign to a growing number of things:

- Institutions
- Datasets
- People
- Organizations
- Articles
- Monographs
- Serials

It might occur to you that libraries are *increasingly* tied into all these things.

- Digital Publishing
- Scholar Profiles
- Research Data
- CRIS Systems
- Repositories
- Bibliometrics/Collections
- Open Scholarship

PIDs should/can make locating and tracking materials/research easier.

PIDs are great for disambiguation and consistent metadata, because:

- Names aren’t unique.
- Names don’t follow rules.
- URLs change.
- Places, people, institutions… etc. are identified in myriad ways.

Judicious application of PIDs (and ubiquitous uptake) could save a lot of time .

PIDs are tied to registration agencies who collect & distribute metadata publicly.

Let’s review some lesser -known things.

“Persistence is purely a matter of service.”

- J. Kunze, 2013

Persistent ≠Permanent

Minting a DOI is different than registering a DOI.

"You see, you know how to [*mint*] the [DOI], you just don't know how to [*register*] the [DOI]. And that's really the most important part of the [DOI]: the [*registration*]. Anybody can just [mint] them."

- Jerry Seinfeld, 1991

PIDs aren't meant to be human-readable, custom URLs.

(DOIs ≠ fancy bit.ly)

PIDs aren't meant to be human-readable, custom URLs.

10.1234/ 097813rhujrho7
10.1234/ journal.24.1.0001

These do the same thing!
No one reads suffixes!

Imagine having to care about a typo in a DOI and the amount of work it takes to fix one.

PIDs as only as useful as their registered metadata .

Garbage in. Garbage Out .

Registration agencies use different and variably compatible metadata schema.

...

PIDs don't have to be assigned to literally everything, nor should they be .

We need to *relax* .

But ! PIDs can be assigned to many things that aren't journal articles and datasets!

There are many registration organizations and types of PIDs.

ROR
GRID
ISNI

Institutions

ORCID (ISNI)
ScopusID
WoS
ResearcherID

Researchers

Crossref / DOI

Articles
Proceedings
Monographs
*Datasets
Funding Agencies
Grants
Reports
Standards
Preprints

Crossref / DOI

Articles
Proceedings
Monographs
*Datasets
Funding Agencies
Grants
Reports
Standards
Preprints

Datacite / DOI

Software
Datasets
Collections
Audio/Visual
Events
Models

Datacite / DOI

Software
Datasets
Collections
Audio/Visual
Events
Models

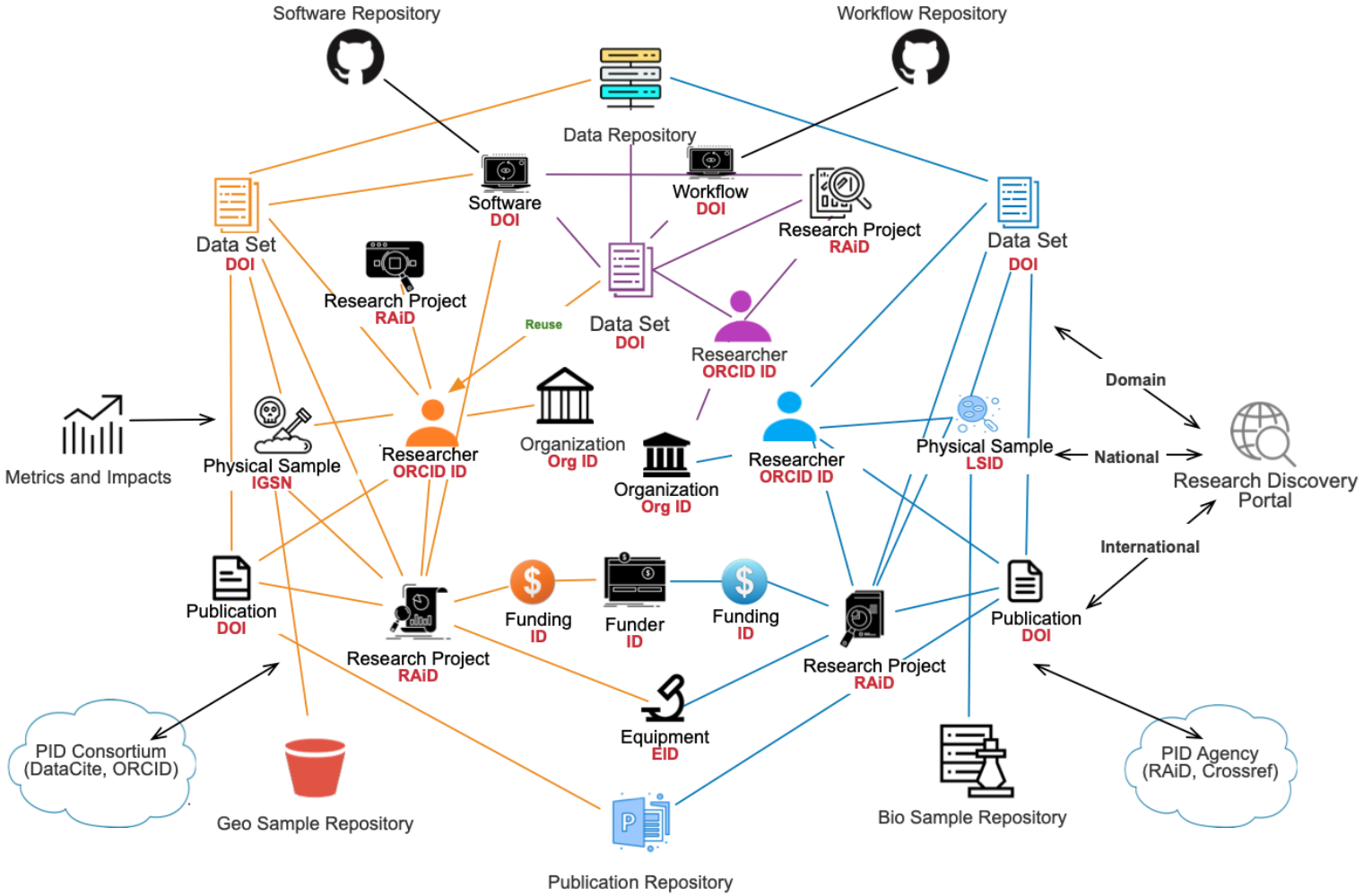
RaID

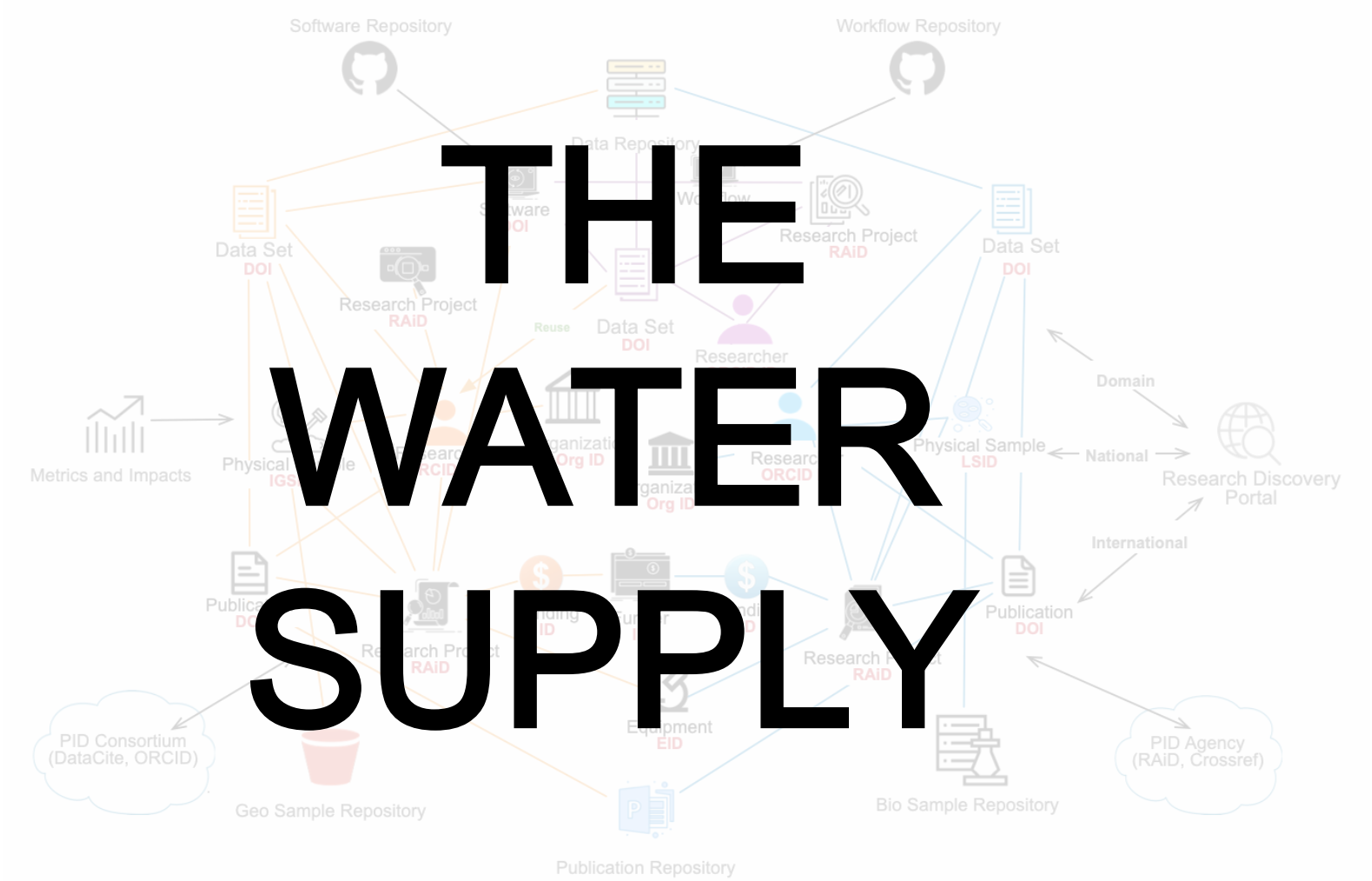
PIDs

All of these platforms either pull data from, or push data to, an open pipeline of metadata.
(aka the water supply)

The “API”

Application Programming Interface


Software Repository
Workflow Repository
Data Repository
Software
DOI
Workflow
DOI
Data Set
DOI
Research Project
RAiD
Data Set
DOI
Research Project
RAiD
Reuse
Data Set
DOI
Researcher
ORCID ID
Domain
Metrics and Impacts
Physical Sample
IGSN
Researcher
ORCID ID
Organization
Org ID
Organization
Org ID
Researcher
ORCID ID
Physical Sample
LSID
National
Research Discovery
Portal
International
Publication
DOI
Funding
ID
Funder
ID
Funding
ID
Publication
DOI
Research Project
RAiD
Research Project
RAiD
Equipment
EID
PID Consortium
(DataCite, ORCID)
Geo Sample Repository
Bio Sample Repository
PID Agency
(RAiD, Crossref)
Publication Repository



THE WATER SUPPLY

Most of these organizations are not -for -profit (obviously not Scopus or WoS).

S'allright? Let's discuss some ways the water flows!

Let’s try a very basic example.

Let’s pretend…

I am setting up my ORCID account.

I want to add my publications!

Within ORCID, I can check against the Crossref and Datacite APIs for any publications matching my name

Most publications assign DOIs.

It will take me a while to do this the first time, and it'll only work if my articles have DOIs.

But…

For all my publications I know are mine, that have DOIs, the metadata is automatically pulled into my ORCID account.

And…!

Now that I have an ORCID, that metadata (ideally) is included when I publish, which means systems will know who I am.

author metadata (unformatted) 

publisher metadata (jats or similar) 

crossref (crossref xml schema) 

orcid (orcid schema… dublin core -ish, "bibtesque")

Each schema is a little different!

Mike, I know how ORCID works.

Fine, let's pretend…

I've applied for funding from an agency that has an ORCID account or integration.

That agency can push new data to my ORCID account.

Funding ID
Grant ID
Datasets
Articles

And ideally...

The next time I apply for funding, I just push my ORCID to the agency and they can pull my works without me filling out the same form again.

If someone from the tri -agency is looking at this, please know it's all I need you to retain. It's the *one thing* .

The next time I apply for funding, I just push my ORCID to the agency and they can pull my works without me filling out the same form again.

Let's try a *more complicated example.*

This time, *we'll crank up the "libraries" knob* .

Let’s pretend…

My institution is using Unsub to get a grasp on where my faculty publishes and how this matches our collections

Unsub is created and maintained by only *two people* .

The software takes affiliation information from the Microsoft Academic Graph which scrapes publications and uses NLP pattern matching.

Open infrastructure does the heavy lifting…

It then takes that affiliation data and checks against the Crossref API for ISSN and publications, your provided collection, GRID or ROR institutional IDs

Unsub then uses that data…

… to tell you where your scholars are publishing, if it's OA (checks against DOAJ and scrapes for policies) and if journals you subscribe to are being published in.

Without the Crossref API, this whole process disappears.

Publications that aren’t using DOIs are, essentially, “off the grid.”

Publications that aren’t using DOIs are, essentially, “off the grid.”

This results in a lot of folks entering the same metadata into systems, by hand. Or hiring graduate students to do this for them. That's an excellent use of everyone's time, definitely.

Persistent identifiers allow us to see the big picture through all of these connections and interactions.

When we talk about support for PIDs we're talking about supporting *open infrastructure* and *free exchange of metadata* .

But what about all these other objects?

Right! Yes. There's three (3) general rules.

Almost every major location a researcher puts their work these days will incorporate PIDs more or less automatically.

1.

Odds are, you'll never really have to worry too much about institutional PIDs or attaching DOIs to pre-prints.

You probably already have a ROR ID, and Arxiv handles DOIs automatically.

1.33

Most of the time, in the library space, PIDs will be happening *to/for you* .

1.66

If you're hosting content that doesn't live anywhere else, or that content is *primarily hosted* on a service you maintain, it is appropriate for you to mint (and register) a DOI for it!

2.

You know this stuff! It's very common in repos.

Grey Lit.
Reports
*Working Papers
Theses
Projects
Slide Decks

You shouldn't mint DOIs
for things that already
have them elsewhere.

You are actively making
things worse.

Stop it.

2.5

Who you register your PID
with *does, in fact, matter* .

3.

Let's take a look at
Crossref/Datacite again.

Crossref / DOI

Articles
Proceedings
Monographs
Datasets*
Funding Agencies
Grants
Reports
Standards
Preprints

Datacite / DOI

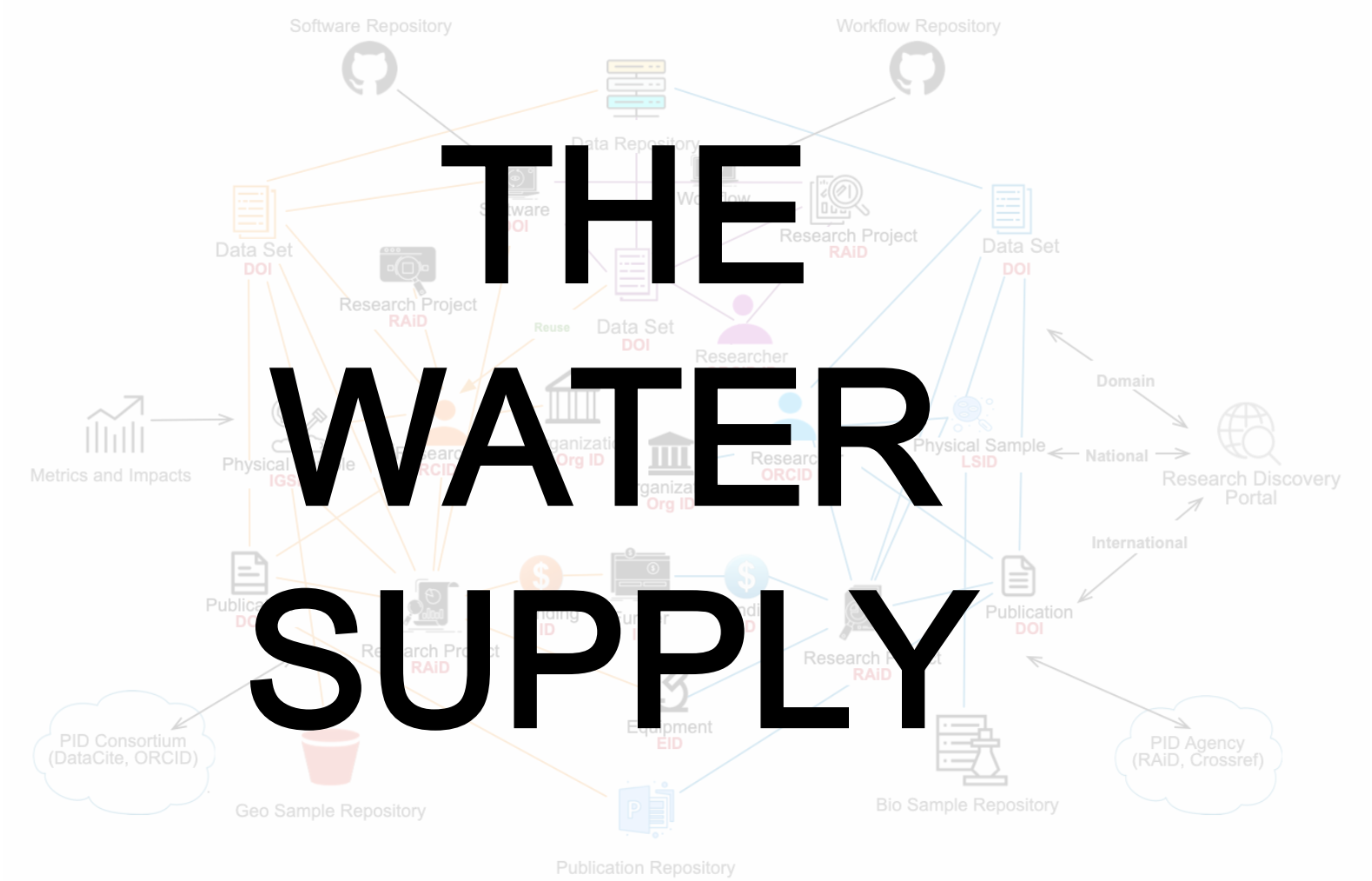
Software
Datasets
Collections
Audio/Visual
Events
Models

Because their schema are specifically designed to represent certain types of content, the fidelity of the metadata may suffer in translation from system - to -system.

Crossref and Datacite are *friends* for this reason.

3.75

Garbage in. Garbage out. 3.81



THE WATER SUPPLY

"I really want to be proactive about judicious use of PIDs!"

- You, incredibly, just now.

Great, do these:

Advocate for ORCID without being so pushy as to remind people that it's a little like being barcoded.

Emphasize researcher agency and privacy.

Promote metadata literacy where you can by helping researchers understand why good metadata will save them time later.

Ally first with a research office under the sales - pitch of metrics that may be abused in scholarly assessment .

Please, don't do these:

Ignore the concerns of faculty unions who (justifiably) aren't thrilled about being boiled down into digestible numbers.

Aren't fine line s fun?

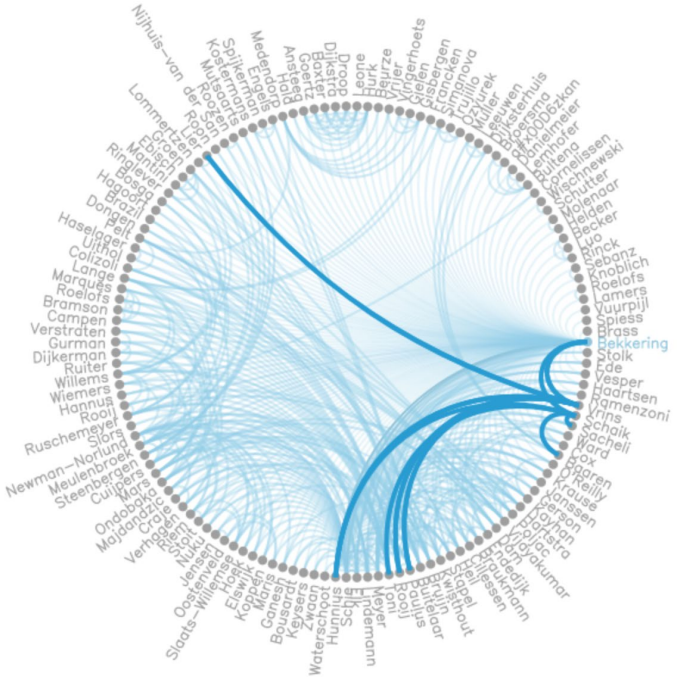
I'm sorry about how very, very fast that was.

I was **Mike Nason,**
UNB Libraries and the Public Knowledge Project.

Now, over to **Mark.**

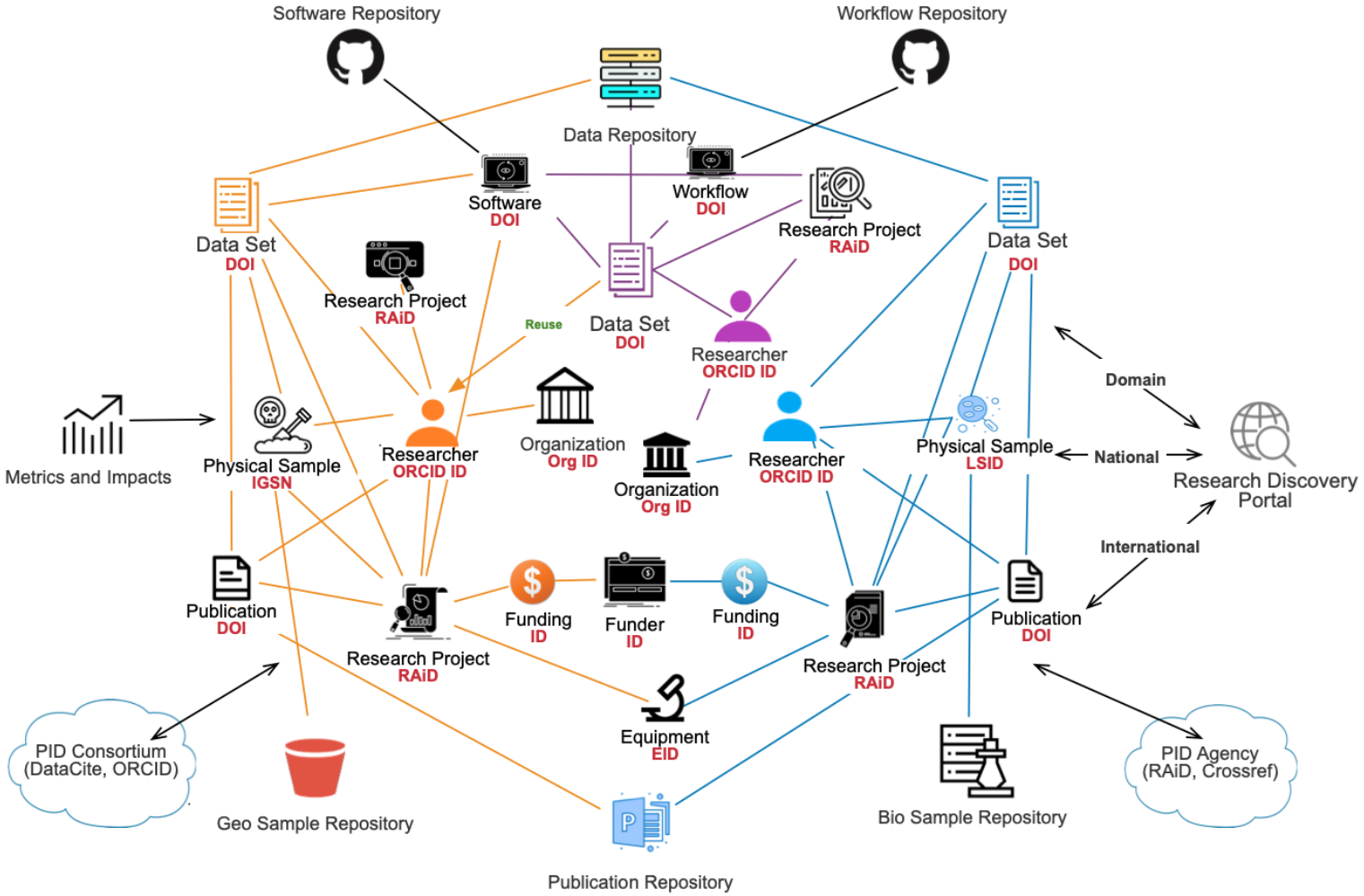
Bringing Object PIDs Together

1. PIDs by themselves are useful, but...
2. PIDs become most useful in the context of a rich PID ecosystem, or the PID Graph
3. Creating links between the full range of research outputs is where the real value lies.



https://www.narcis.nl/personcowo/RecordID/PRS1291410/uquery/bekkering/id/6/Language/EN.

https://www.youtube.com/watch?v=yWOqeyPIVRo


Software Repository
Workflow Repository
Data Repository
Software
DOI
Workflow
DOI
Data Set
DOI
Research Project
RAiD
Data Set
DOI
Research Project
RAiD
Reuse
Data Set
DOI
Researcher
ORCID ID
Domain
Metrics and Impacts
Physical Sample
IGSN
Researcher
ORCID ID
Organization
Org ID
Organization
Org ID
Researcher
ORCID ID
Physical Sample
LSID
National
Research Discovery
Portal
International
Publication
DOI
Funding
ID
Funder
ID
Funding
ID
Publication
DOI
Research Project
RAiD
Research Project
RAiD
Equipment
EID
PID Consortium
(DataCite, ORCID)
Geo Sample Repository
Bio Sample Repository
PID Agency
(RAiD, Crossref)
Publication Repository

Challenge 1: Agreement on PIDs

1. Agreeing on the **PID To Rule Them All** is not feasible…

2. But we can agree on best practice PIDs for specific objects
3. Which facilitates adoption and development of software

4. Ultimately creating a rich PID ecosystem that makes it easy to find any asset in the research ecosystem

The “Conductor” PID

1. Research Activity ID (RaID) is a unique type of PID, that acts as an aggregator of PIDs of all types, associated with a specific research project or defined activity
2. Having one PID that can be accessed in the same way, via a single API, provides an efficient and useful representation of the PID Graph
3. Examples in specific disciplines, e.g. BioProject, but RaID has the potential to be the DOI for all research projects

Display Settings: Send to:

Severe acute respiratory syndrome coronavirus 2

Accession: PRJNA686984 ID: 686984

SARS-CoV-2 Genome sequencing and assembly

NGS of PCR-tiled SARS-CoV-2

Accession	PRJNA686984
Data Type	Genome sequencing
Scope	Multiisolate
Organism	**Severe acute respiratory syndrome coronavirus 2** [Taxonomy ID: 2697049] Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome-related coronavirus; Severe acute respiratory syndrome coronavirus 2
Grants	"CK19-1904 Epidemiology and Laboratory Capacity for Infectious Diseases (ELC)" (Grant ID 6 NU50CK000552-02-01, Center for Disease Control and Prevention)
Submission	Registration date: 21-Dec-2020 **Colorado Department of Public Health and Environment**

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (Genomic RNA)	1962
SRA Experiments	1971
Protein Sequences	23526
OTHER DATASETS	
BioSample	1971

▾ SRA Data Details

Parameter	Value
Data volume, Gbases	138
Data volume, Mbytes	72794

See Genome Information for Severe acute respiratory syndrome-related coronavirus

NAVIGATE UP

This project is a component of the COVID-19 Outbreak

This project is a component of the INSDC SARS-CoV-2 Viral Sequencing Data

NAVIGATE ACROSS

29 additional projects are components of the COVID-19 Outbreak.

250 additional projects are components of the INSDC SARS-CoV-2 Viral Sequencing Data.

134 additional projects are related by organism.

Related information

- BioProject
- BioSample
- Genomic RNA
- Nucleotide
- Protein
- SRA
- Taxonomy
- Umbrella projects

Recent activity

Turn Off Clear

- Severe acute respiratory syndrome coronavirus 2 — BioProject
- 615625[top bioproject] NOT 718231[uid] (29) — BioProject
- Lichtheimia corymbifera strain:B63a — BioProject
- Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, c — Nucleotide
- covid-19 (71327) — Nucleotide

See more...

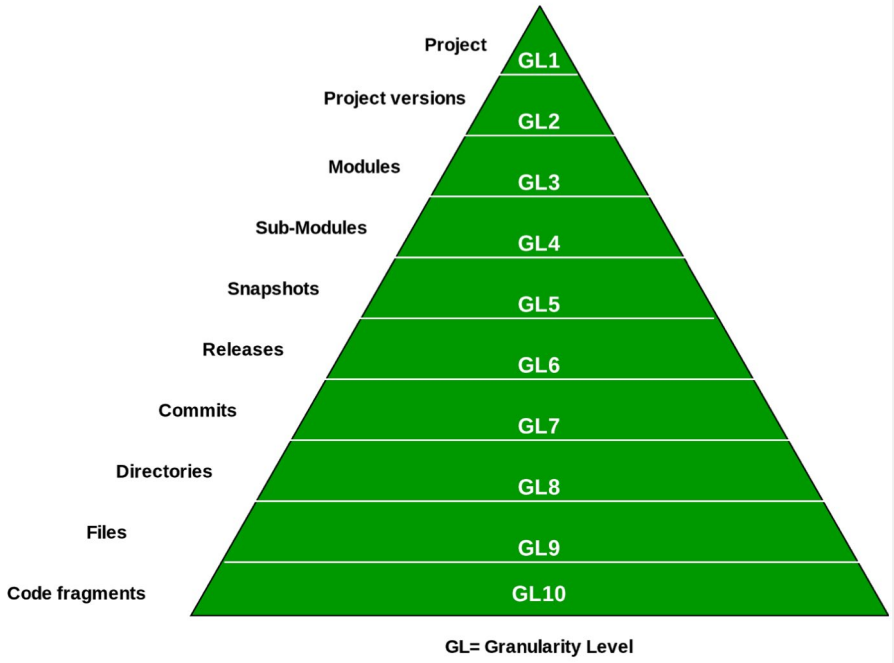
Research Activity Identifier (RaID)

1. Comes from work in Australia to create a “research management record”
2. Aggregates PIDs for all resources associated with a specific project
3. Undergoing ISO review/approval (completed this year)
4. RaIDs will be minted by regional partners and is a free service

RaID Video

Other Objects PIDs: Software

1. SW PIDs are complicated!
2. Intrinsic
   a. PIDs generated by the SW development environment, such as VCS (Git ID)
   b. Basis of SWHID
3. Extrinsic
   a. PIDs external to the SW context
   b. DOIs created by repositories like Zenodo



http://doi.org/10.15497/RDA00053

Other Objects PIDs: Equipment

1. “to interpret a digital dataset, much must be known about the hardware used to generate the data, whether sensor networks or laboratory machines”
2. RDA PIDINST Working Group created a 43 -element schema, such as ID, Owners, Manufacturer, Measured Variables, etc.
3. 2 examples
    a. DataCite DOIs
    b. ePIC Framework
4. Others, such as RRID use for equipment/facilities

http://doi.org/10.5334/dsj-2020-018

Other Object PIDs: Resources

1. RRIDs: Research Resource IDs
    a. cell lines, antibodies, plasmids, model organisms, facilities and equipment
2. Additional rigour/detail in describing associated reference resources, typically in *Materials*
3. Increasingly used by journal publishers

Many Others in Domain Contexts

MycoBank ID

InChi

DIN

EC Number

PDP ID

Future State?

1. Short-term goal would be to have wide adoption of core Best Practice PIDs, including RaID
2. W3Cs Decentralized Identifiers (DIDs)
    a. *“a new type of identifier that enables verifiable, decentralized digital identity. A DID identifies any subject (e.g., a person, organization, thing, data model, abstract entity, etc.) that the controller of the DID decides that it identifies.”*
3. Non-Fungible Tokens (NFTs)
    a. Role in Scholarly Communications and persistent identity? (see Scholarly Kitchen)