Metadata & Persistent identifiers

Espen Åberg
Data Steward
ELIXIR Norway/BioMedData

Your tasks

Documentation and metadata

How can you document data during the project?

Description

Data documentation could be defined as the clear description of everything that a new "data user" or "your future-self" would need to know in order to find, understand, reproduce and reuse your data, independently. Data documentation should clearly describe how you generated or used the data, why, and where to find the related files. It could be used also as onboarding documentation for new colleagues, even if the responsible researcher leaves the project.

Due to the large variety of experiments, techniques and collaborative studies that usually occur within the same project, it is challenging to keep good documentation. However, lack of good data documentation often leads to data loss, not reproducible results and therefore, waste of money and time for scientists. Here we provide best practices and guidelines to help you properly document your data.

Considerations

- Write the documentation in such a way that someone else who is known to the field cannot misinterpret any of the data.
- It is best practice to use one appropriate tool or an integration of multiple tools (also called tool assembly or ecosystem) for data documentation during a project. Suitable tools for data documentation are Electronic Lab Notebooks (ELNs), Electronic Data Capture (EDC) systems, Laboratory Information Management Systems (LIMS). Moreover, online platforms for collaborative research and file sharing services (such as OpenScienceFramework) could also be used as ELN or data management systems. Check with your institute or other relevant infrastructures to know what is offered.
- Independently of the tools you will use, data documentation is needed at two levels: documentation about the entire study or project and documentation about individual records, observations or data points.
  - Study-level documentation describes the project title and summary, study aims, authors, institutions involved, funds, methods, licence and identifier for each dataset, folders structure, file naming conventions, versioning system, relation between files or tables and other general information.
  - Data-level documentation provides information about individual records or data point, such as the

Your tasks

Identifiers

Which types of identifiers can you use during data collection?

Description

A lot of (meta)data is collected in the form of tables, representing quantitative or qualitative measurements (values in cells) of certain named properties (variables in columns) of a range of subjects or samples (records or observations in rows). It can help your research a lot if you make sure you can address each of these records, variables and values unambiguously, i.e. if each has a unique identifier. This is also true for (meta)data that is not in tabular format ("key":value format, unstructured data, etc.). Identifiers should be always used for metadata and data independently of the format.

If the research institute or group has a centralised and structured system (such as a central electronic database) in place to describe and store (meta)data, this process can be quite straight forward for the researcher. However, if there is no such system, often researchers have to set up an internal "database" to keep track of each record or observation in a study. This situation can be quite challenging for many reasons, one of which is assigning identifiers. The use of identifiers for records, variables and values will increase the reusability and interoperability of the data for you, your future-self and others.

Considerations

- At the beginning of your research project, check if your institute or research group has a centralised database where data must be entered during data collection. Usually, large and international research projects, industries, research institutes or hospitals have a centralised electronic database, an Electronic Data Capture (EDC) system, a Laboratory Information Management System (LIMS) or an Electronic Lab Notebook (ELN) with an user interface for data entry. More details about using ELNs are given by e.g. University of Cambridge - Electronic Research Notebook Products and Harvard Medical School - Electronic Lab Notebooks.
- If you can choose how to manage your data entry system, consider what's the level of exposure of the identifier for each record or observation in the dataset. Define the context in which the identifier should be used and is unique. This is a key aspect to define what kind of identifier for each individual record is appropriate in your case.
  - Should the identifier of a record or observation be unique within your spreadsheet, your entire research project files or across the whole institute? What's the reference system (or "target audience") of your identifier?
  - Will your reference system change in due time? If it will be opened up later, assigning globally unique identifiers from the beginning may be saving time.
  - Will the identifiers for individual records or observations be made openly accessible on the internet,

Link to RDMkit: https://rdmkit.elixir-europe.org/

Has a useful purpose

Can be acted upon and processed by humans and machines

"Metadata is constructed, constructive, and actionable."

Definition from Karen Coyle, Digital Librarian and Author of Coyle's InFormation

"information about something"

What is metadata?

"data about data"

"Data is content, and metadata is context"

"Metadata is a Love Note to the Future"

Metadata “all over the place”?

Link to RDMkit: https://rdmkit.elixir-europe.org/

This is why you care

Metadata **facilitates** organization, indexing, discovery, access, analysis, and use of data.

Metadata **presence and quality** (or the lack thereof) can significantly **help or hinder** time and money expenditures in research activities.

Metadata helps make data FAIR

Data should be **Findable**	F1. (meta)data are assigned a globally unique and persistent identifier (DOI) F2. data are described with rich metadata F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource
Data should be **Accessible**	A1. (meta)data are retrievable by their identifier using a standardized communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary A2. metadata are accessible, even when the data are no longer available
Data should be **Interoperable**	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles I3. (meta)data include qualified references to other (meta)data
Data should be **Reusable**	R1. meta(data) are richly described with a plurality of accurate and relevant attributes R1.1. (meta)data are released with a clear and accessible data usage license R1.2. (meta)data are associated with detailed provenance R1.3. (meta)data meet domain-relevant community standards

Medyckyj-Scott, David et al. (2016).

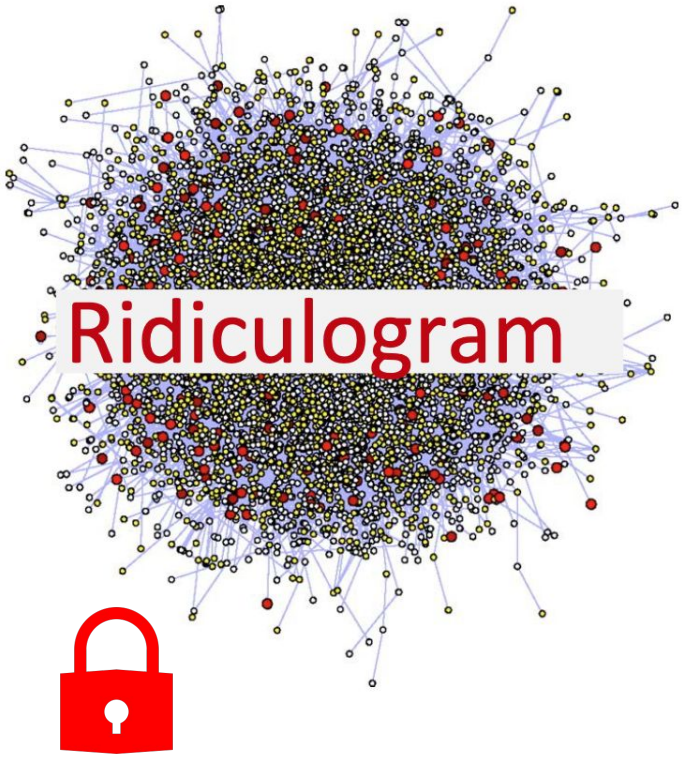
What is Metadata?

"Data"

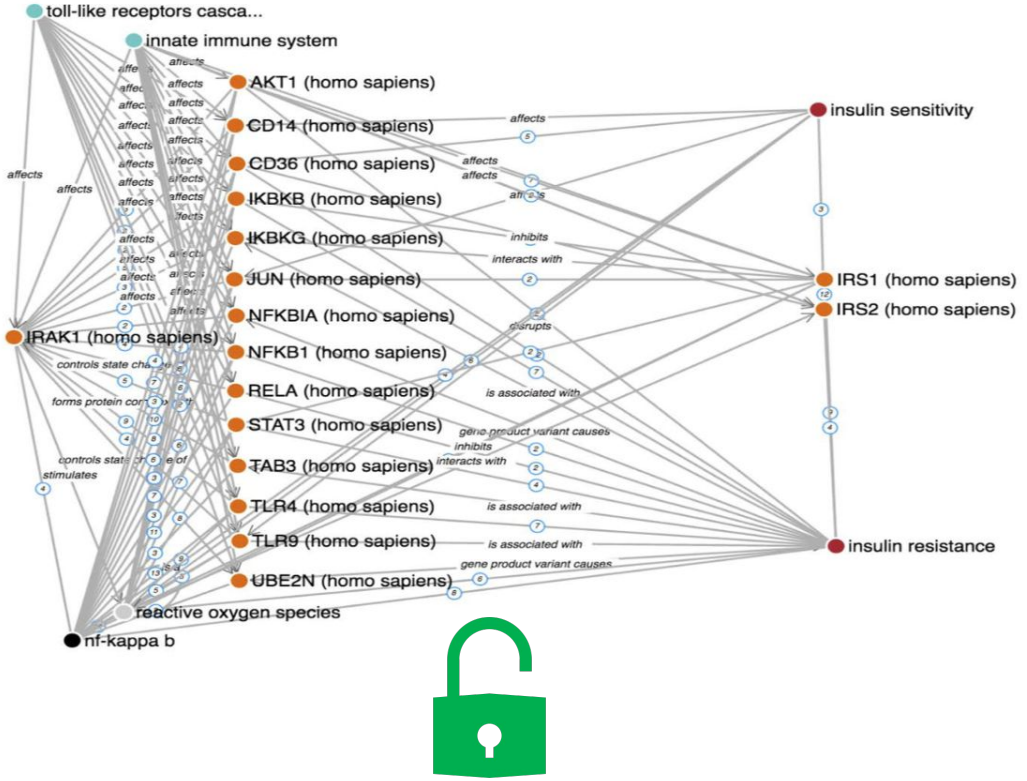
"Metadata"

Outcome =

Helps to gain insight



Metadata

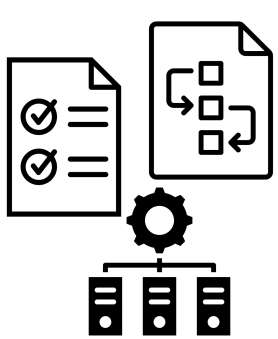


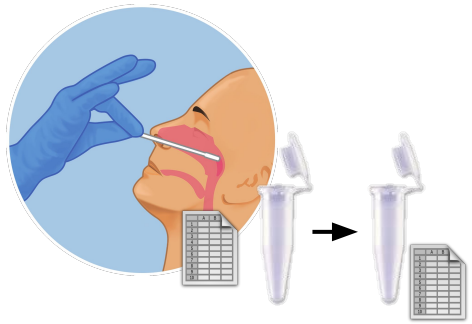
"If data is the new oil, metadata is the refinery"

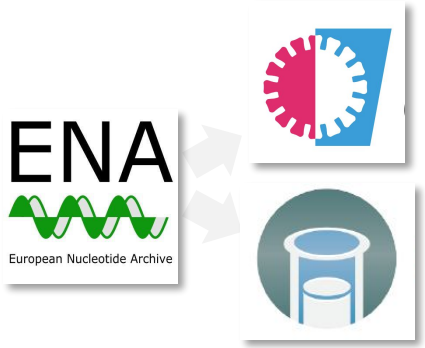
— Adam Rauh

(Meta)data generation and collection

"Protocol" & "project plan" icons by Justin Blake, and "infrastructure" icon by Eko Purnomo, from thenounproject.com







Study & data design	Sampling & specimen collection	Sample preparation	Sample analysis & data generation	Data processing to prepare inputs for analysis	Data analysis	Communicating results

Procedures

data protection,
ethics permit,
infrastructure,
standards,
protocols,
data dictionaries,
data access, ...

Biosamples and instruments

populations (statistical) and inclusion criteria,
physical processing steps,
working storage conditions,
long-term storage location,
sample quality assessment,
sample annotations,
reagents, instruments, kits, ...

Data and computational workflows

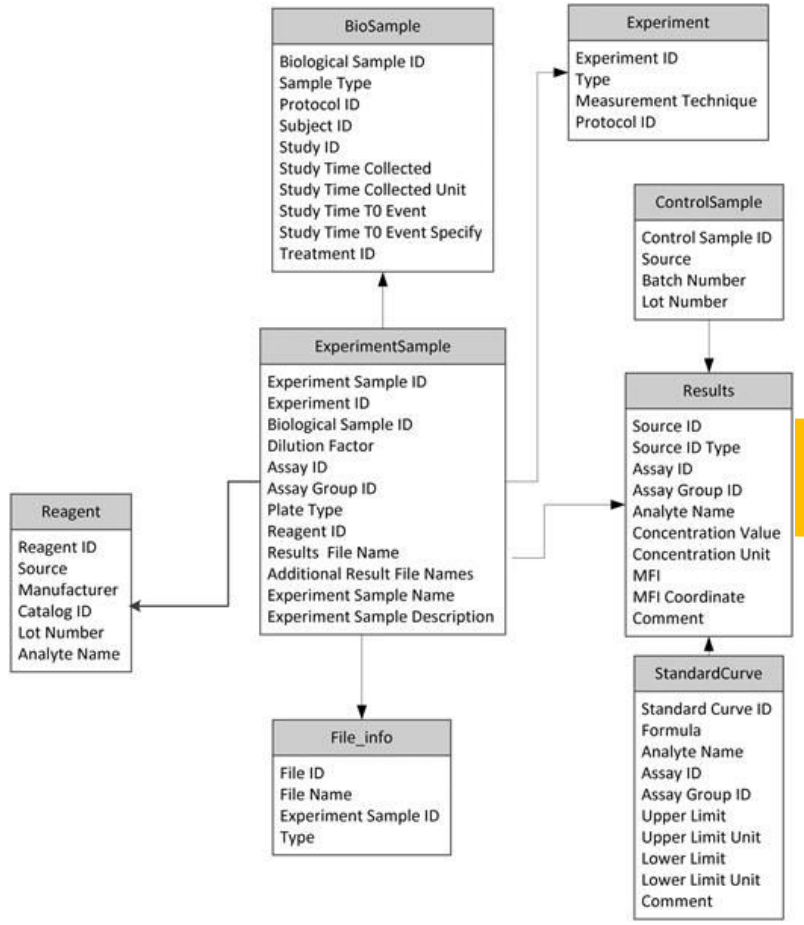
digital processing steps,
working storage conditions,
long-term storage location,
data quality assessment,
sample/data annotations,
reference data,
analysis method...

Outputs

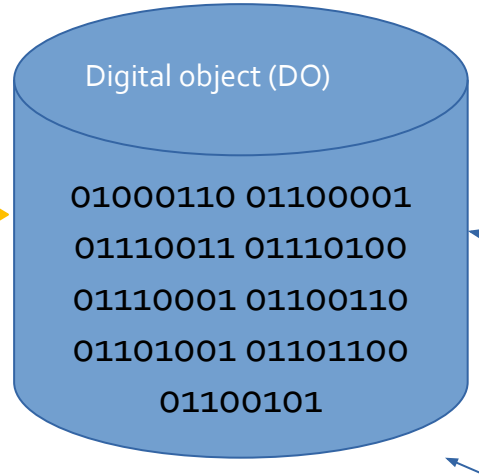
publications,
data,
tools,
workflows,
reports,
dashboards, ...

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN SciLifeLab elixir SWEDEN Slide by: **Wolmar** N. Åkerström

"Rich" Metadata



Metadata



Authorship:
Creator
Creator date
Modified date
Publisher
Owner
Contact

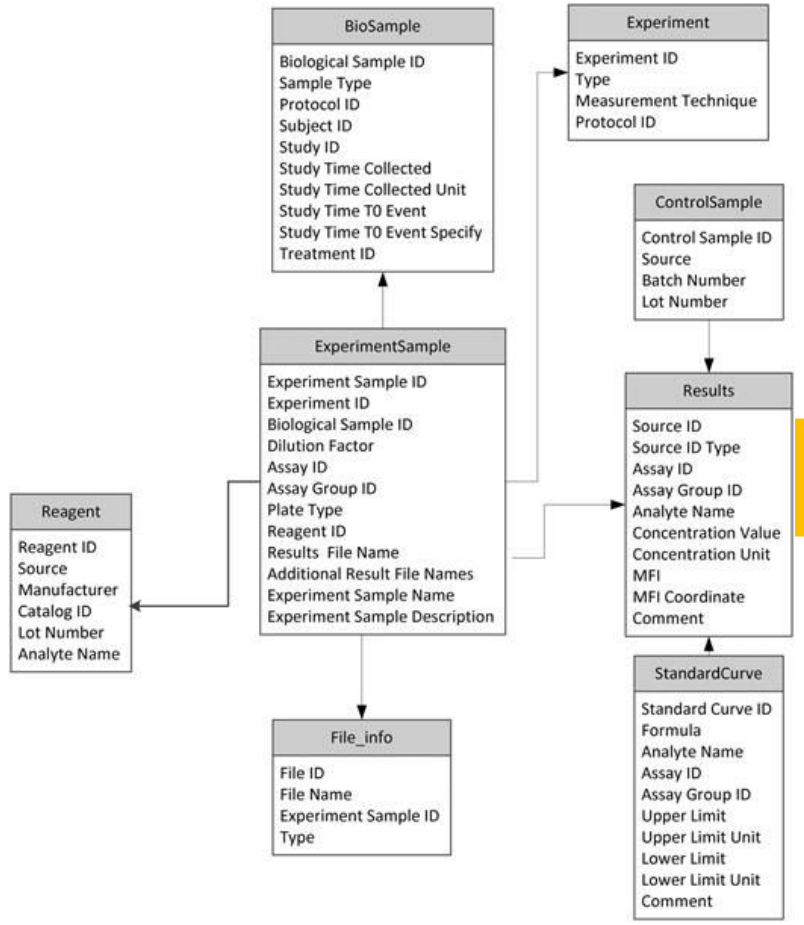
Descriptive:
Title
Description
Keywords
Theme
Main subject
Version
Review changes
Validation status
Size
Format
Language
Semantic model
Conforms to standard/schema

Context:
Project description
Temporal coverage
Spatial coverage
Creation process
Source description

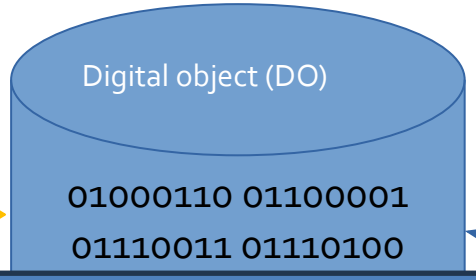
Quality:
Quality parameters

Accessibility conditions:
License
Access rights
Download/access URL

"Rich" Metadata



Metadata



Authorship:
Creator
Creator date
Modified date
Publisher
Owner
Contact

Descriptive:
Title
Description
Keywords
Theme
Main subject
Version
Review changes
Validation status
Size
Format

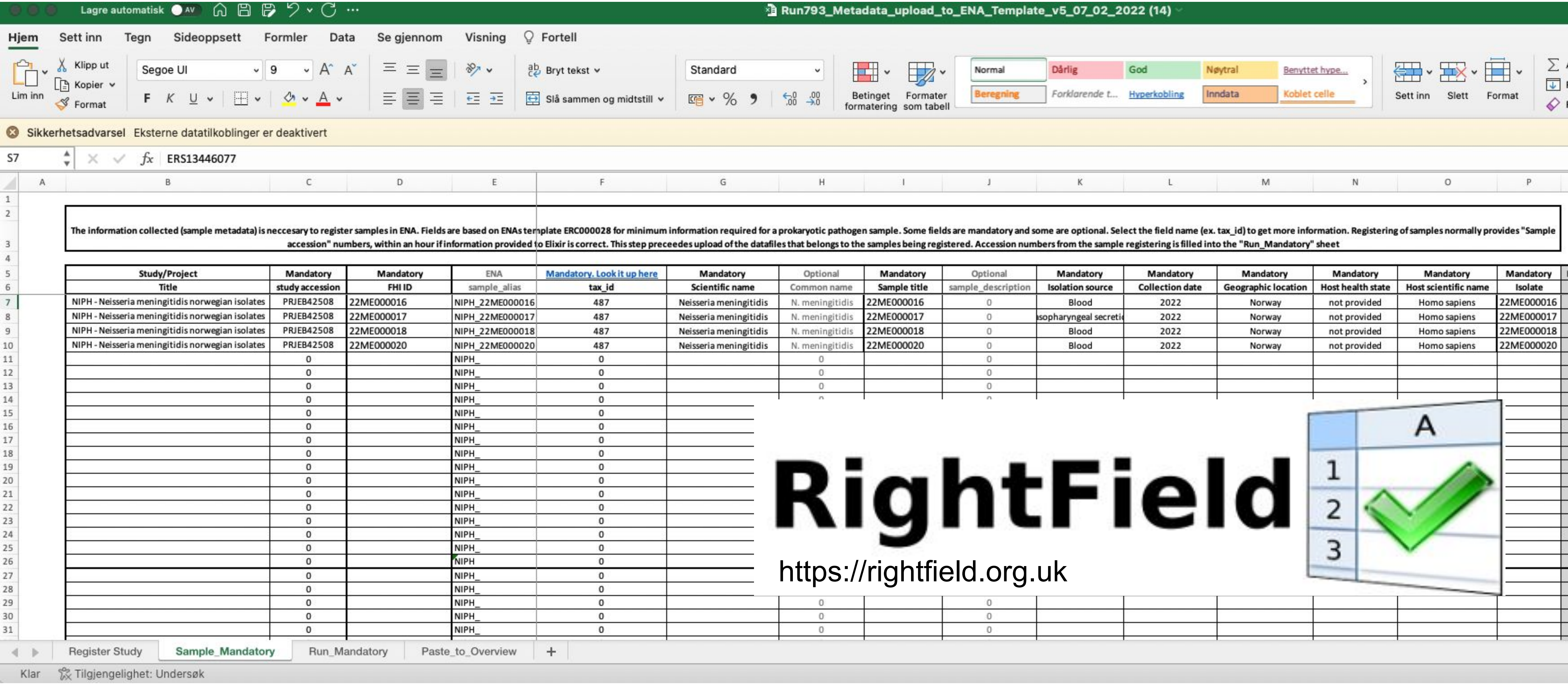
Context:
Project description

Download/access URL

A rule of thumb, you should never say 'this metadata isn't useful'; be generous and provide it anyway!»

Metadata collection (Simple solution)

Not optimal but can be fitted to a set of rules (schema) and allowed values (controlled vocabulary)



Study/Project Title	Mandatory study accession	Mandatory FHI ID	ENA sample_alias	Mandatory. Look it up here tax_id	Mandatory Scientific name	Optional Common name	Mandatory Sample title	Optional sample_description	Mandatory Isolation source	Mandatory Collection date	Mandatory Geographic location	Mandatory Host health state	Mandatory Host scientific name	Mandatory Isolate
NIPH - Neisseria meningitidis norwegian isolates	PRJEB42508	22ME000016	NIPH_22ME000016	487	Neisseria meningitidis	N. meningitidis	22ME000016	0	Blood	2022	Norway	not provided	Homo sapiens	22ME000016
NIPH - Neisseria meningitidis norwegian isolates	PRJEB42508	22ME000017	NIPH_22ME000017	487	Neisseria meningitidis	N. meningitidis	22ME000017	0	sopharyngeal secreti	2022	Norway	not provided	Homo sapiens	22ME000017
NIPH - Neisseria meningitidis norwegian isolates	PRJEB42508	22ME000018	NIPH_22ME000018	487	Neisseria meningitidis	N. meningitidis	22ME000018	0	Blood	2022	Norway	not provided	Homo sapiens	22ME000018
NIPH - Neisseria meningitidis norwegian isolates	PRJEB42508	22ME000020	NIPH_22ME000020	487	Neisseria meningitidis	N. meningitidis	22ME000020	0	Blood	2022	Norway	not provided	Homo sapiens	22ME000020
	0		NIPH_	0		0		0						
	0		NIPH_	0		0		0						
	0		NIPH_	0		0		0						
	0		NIPH_	0										
	0		NIPH_	0										
	0		NIPH_	0										
	0		NIPH_	0										
	0		NIPH_	0										
	0		NIPH_	0										
	0		NIPH_	0										
	0		NIPH_	0										
	0		NIPH_	0										
	0		NIPH_	0										
	0		NIPH_	0										
	0		NIPH_	0										
	0		NIPH	0										
	0		NIPH_	0										
	0		NIPH_	0										
	0		NIPH_	0		0		0						
	0		NIPH_	0		0		0						
	0		NIPH_	0		0		0						

Metadata templates/checklists

Sample Checklists

There is a minimum amount of information required during ENA sample registration and all samples must conform to a defined checklist of expected metadata values. The most suitable checklist for sample registration depends on the type of the sample.

These sample checklists have been developed to meet the needs of different research communities. Different communities have different requirements on the minimum metadata expected to describe biological samples.

Filter checklists...

Accession	Name	Description
ERC000012	GSC MIxS air	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000013	GSC MIxS host associated	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000014	GSC MIxS human associated	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000015	GSC MIxS human gut	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000016	GSC MIxS human oral	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000017	GSC MIxS human skin	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000018	GSC MIxS human vaginal	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...

Checklist: ERC000031

GSC MIxS built environment

Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.

Checklist Fields

Filter fields...

Filter by type:

- non-sample terms
- Collection event information
- internal environment
- sample collection
- host description
- local environment conditions
- concentration measurement
- organism characteristics
- local environment conditions imposed

Field Name	Field Format	(Field Restriction)	Requirement	(Units)
relative air humidity	restricted text	regular expression	mandatory	%
absolute air humidity	restricted text	regular expression	mandatory	kg
surface humidity	restricted text	regular expression	optional	%
air temperature	restricted text	regular expression	mandatory	°C
surface temperature	restricted text	regular expression	optional	°C
surface moisture	restricted text	regular expression	optional	options
surface moisture pH	restricted text	regular expression	optional	
dew point	restricted text	regular expression	optional	°C

https://www.ebi.ac.uk/ena/browser/checklists

Metadata Submission Workflow

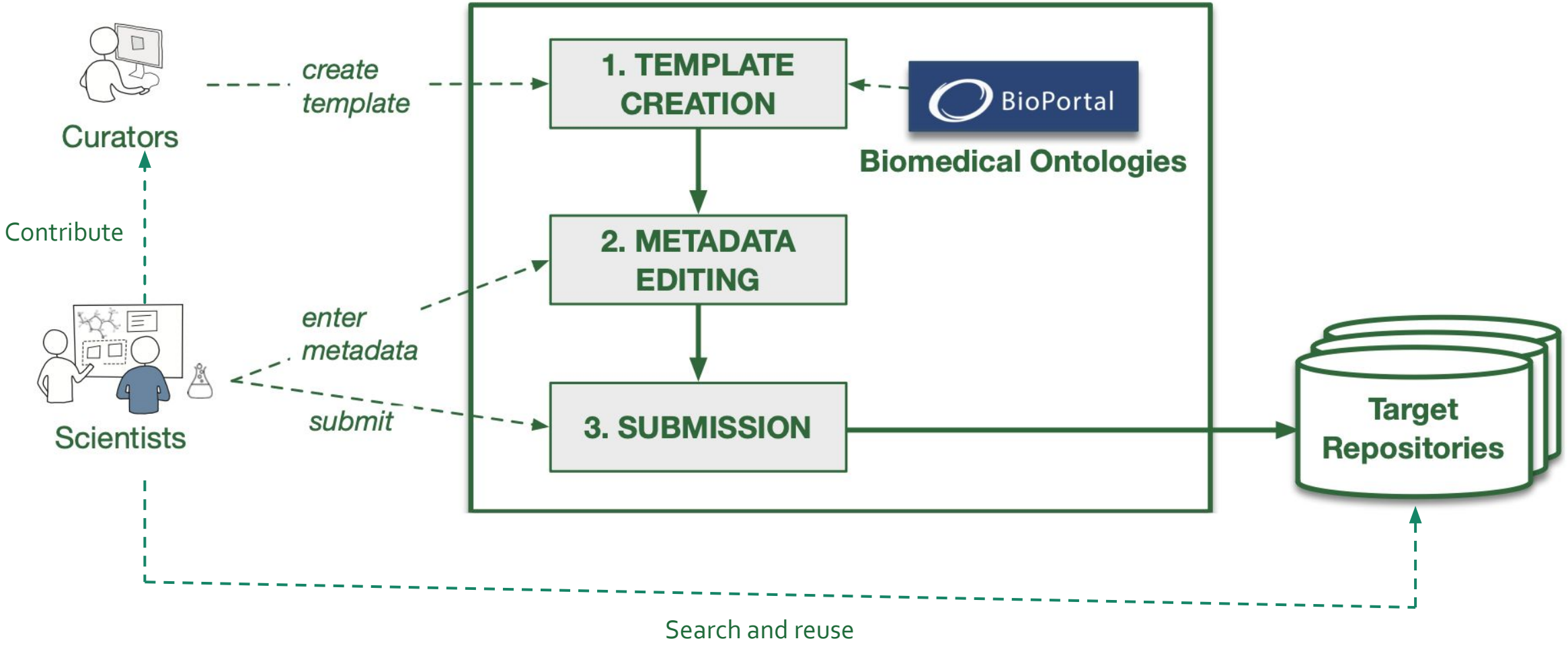


Figure 2 from: Using Semantic Technologies to Enhance Metadata Submissions to Public Repositories in Biomedicine

Different Persona

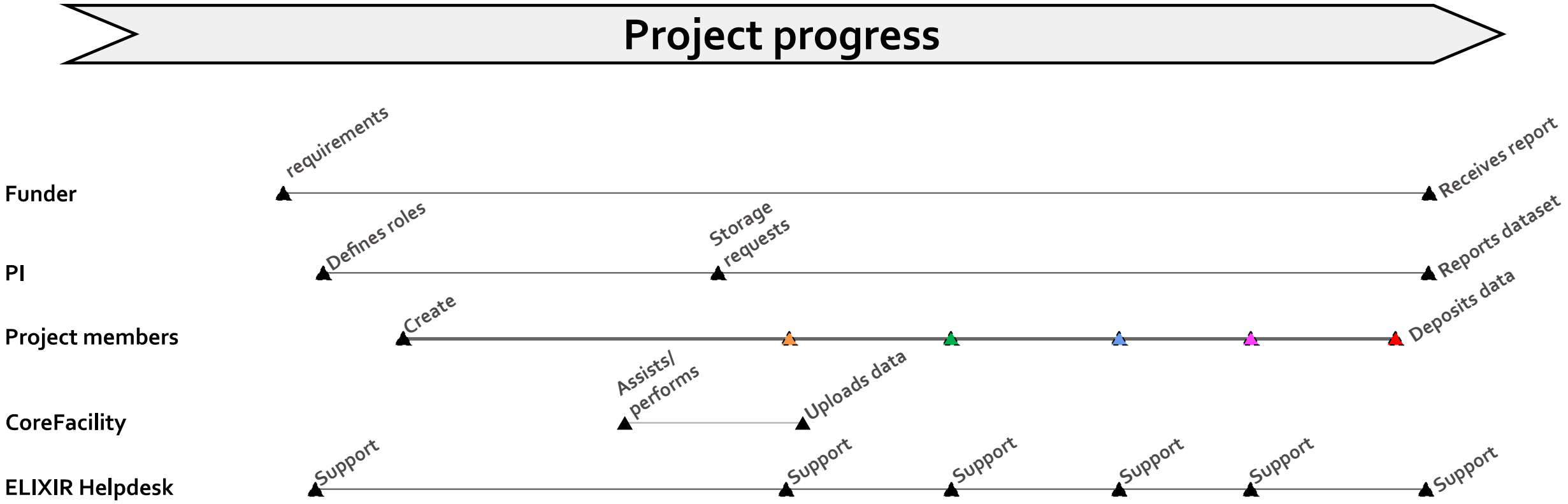


illustration: Korbinian Bösl

Infrastructure tools

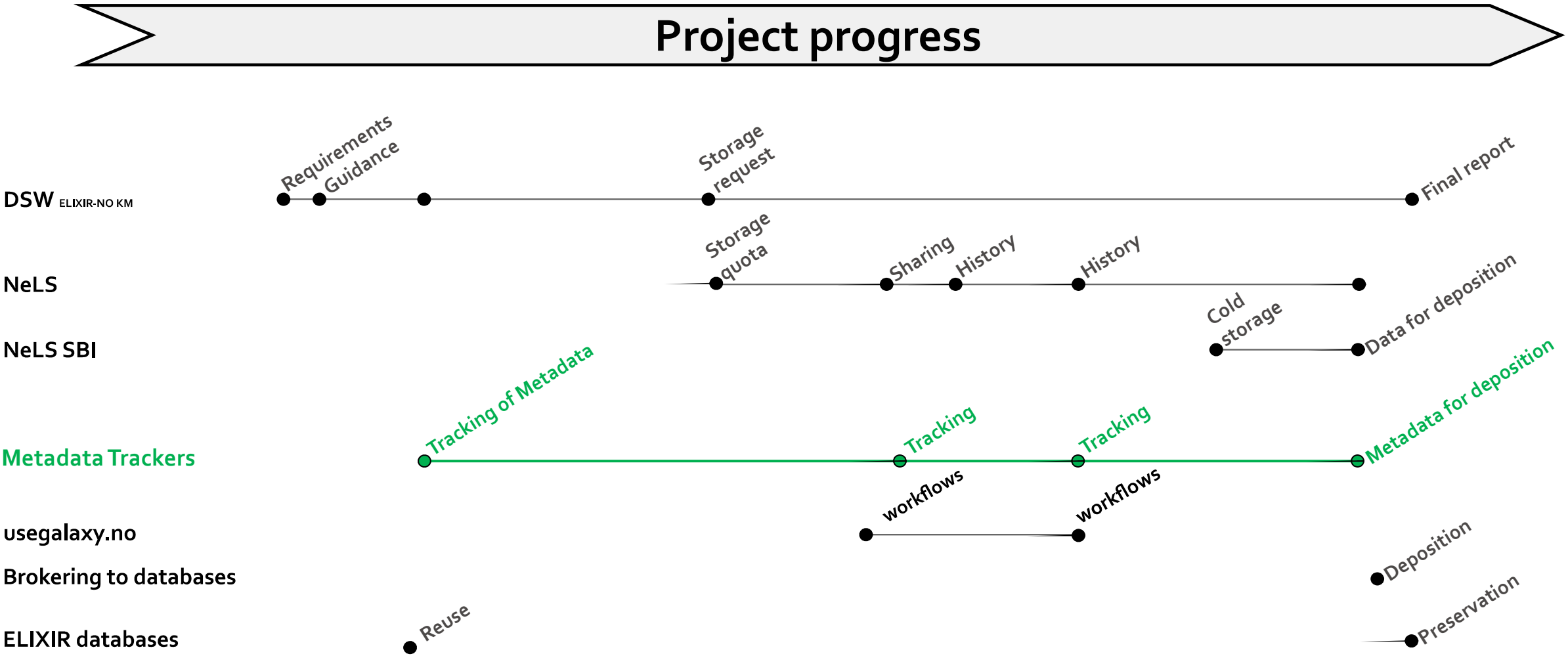


illustration: Korbinian Bösl

Metadata tracking platforms

Domain specific:

- COPO for plant sciences
- MOLGENIS for biobanking
- …

Adaptable (configuration requires domain knowledge):

- Proprietary ELNs/LIMS - often poor support for ontologies
- openBIS - open source ELN/LIMS
- SEEK

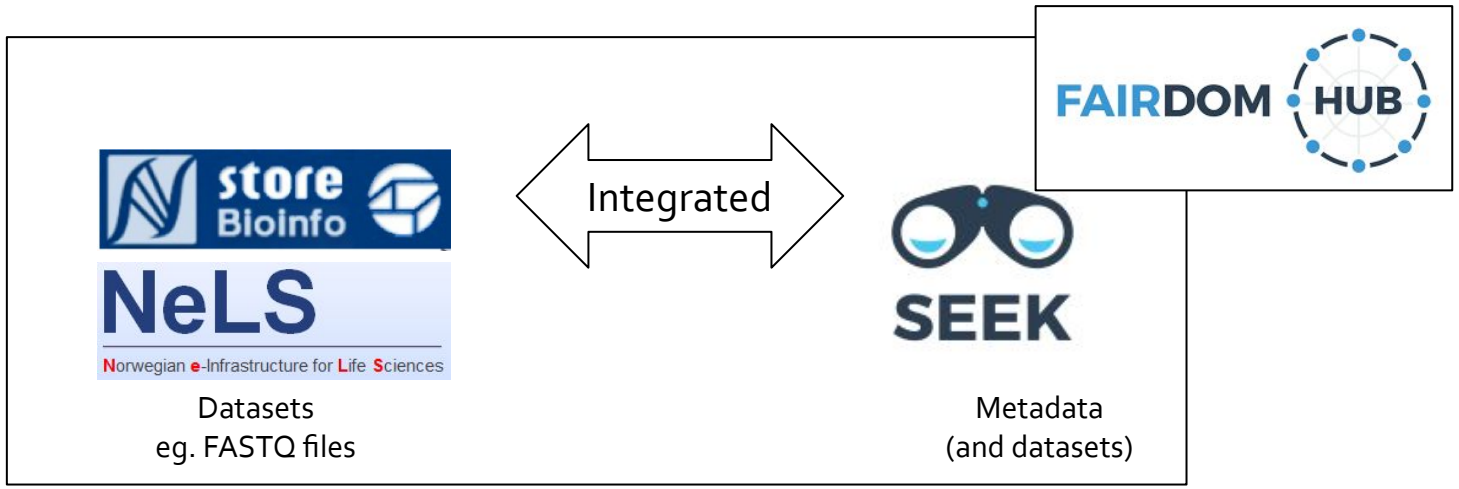
https://copo-project.org/ https://www.molgenis.org/
https://openbis.ch/ https://seek4science.org/

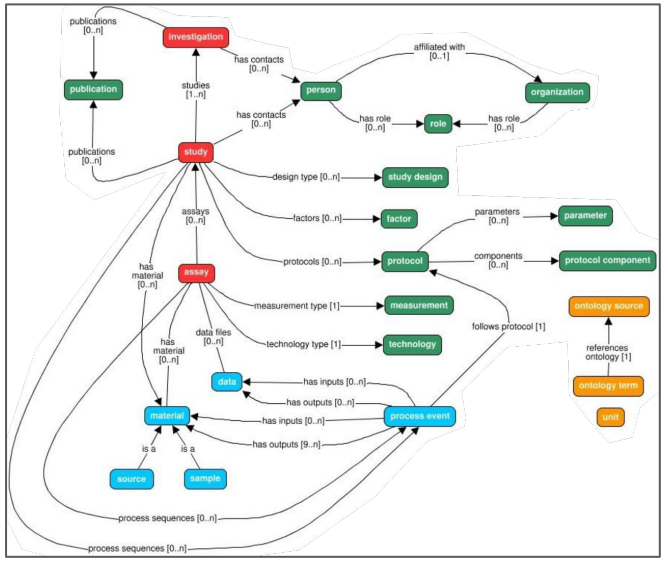
FAIRDOM SEEK

The SEEK platform is a web-based tool for organising and storing data, for sharing, exploring and annotating data

ISA model

- **I**nvestigation — "overall goals and means"
- **S**tudy — Subject/Specimen being studied
- **A**ssay — Experimental steps and outputs



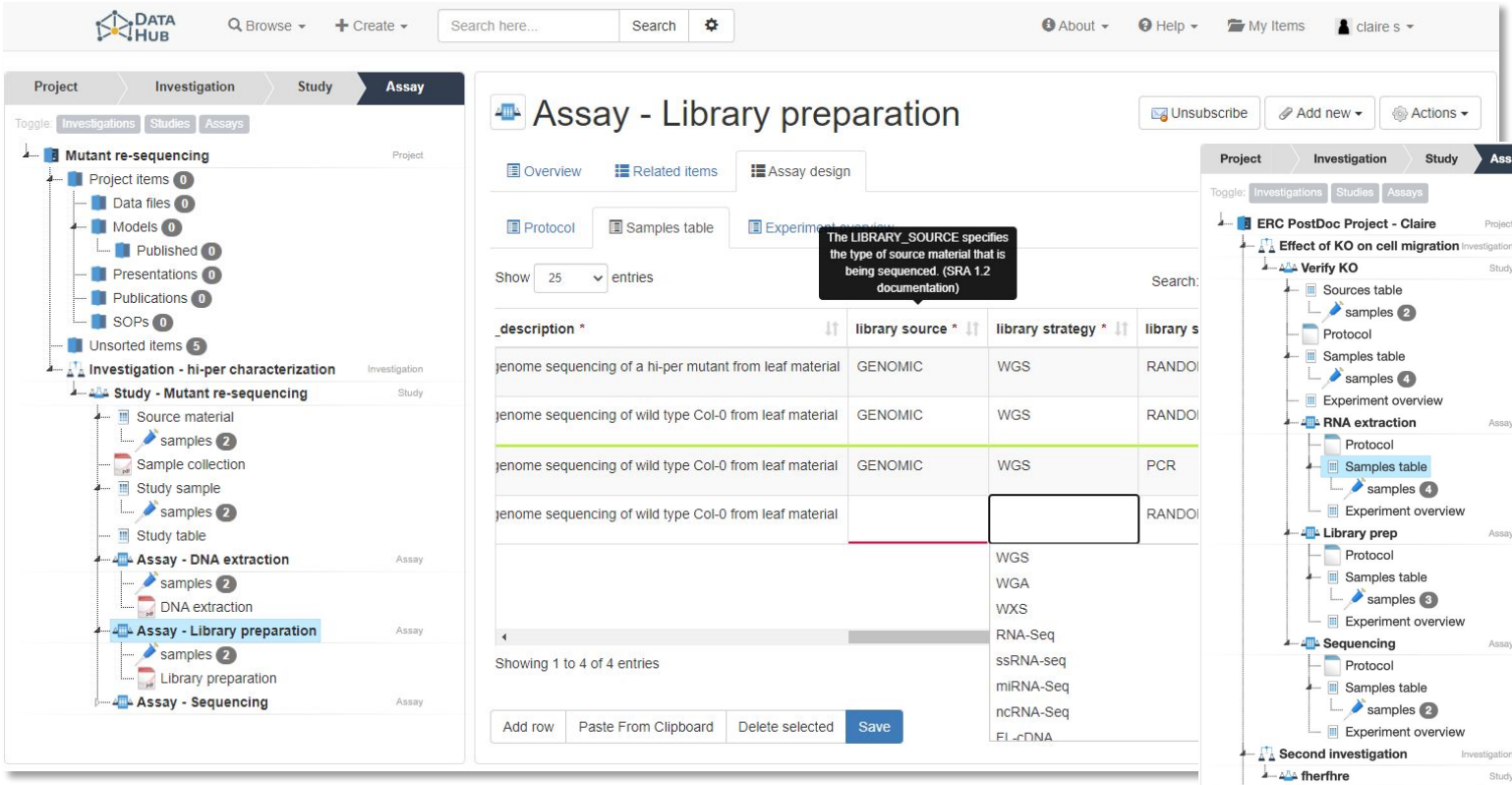


https://doi.org/10.1093/gigascience/giab060

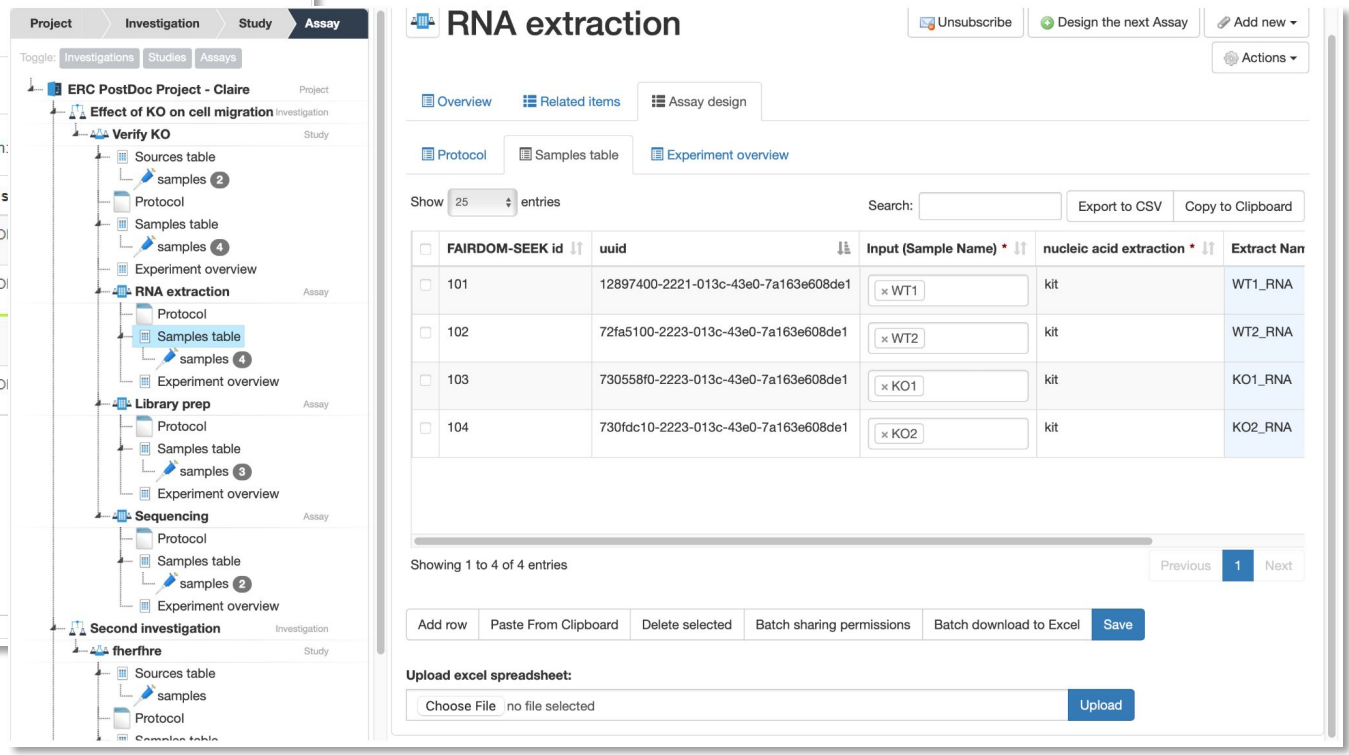
ISA illustration: Rafael Andrade Buono

FAIRDOM SEEK in DataHub

Dynamic table for samples metadata

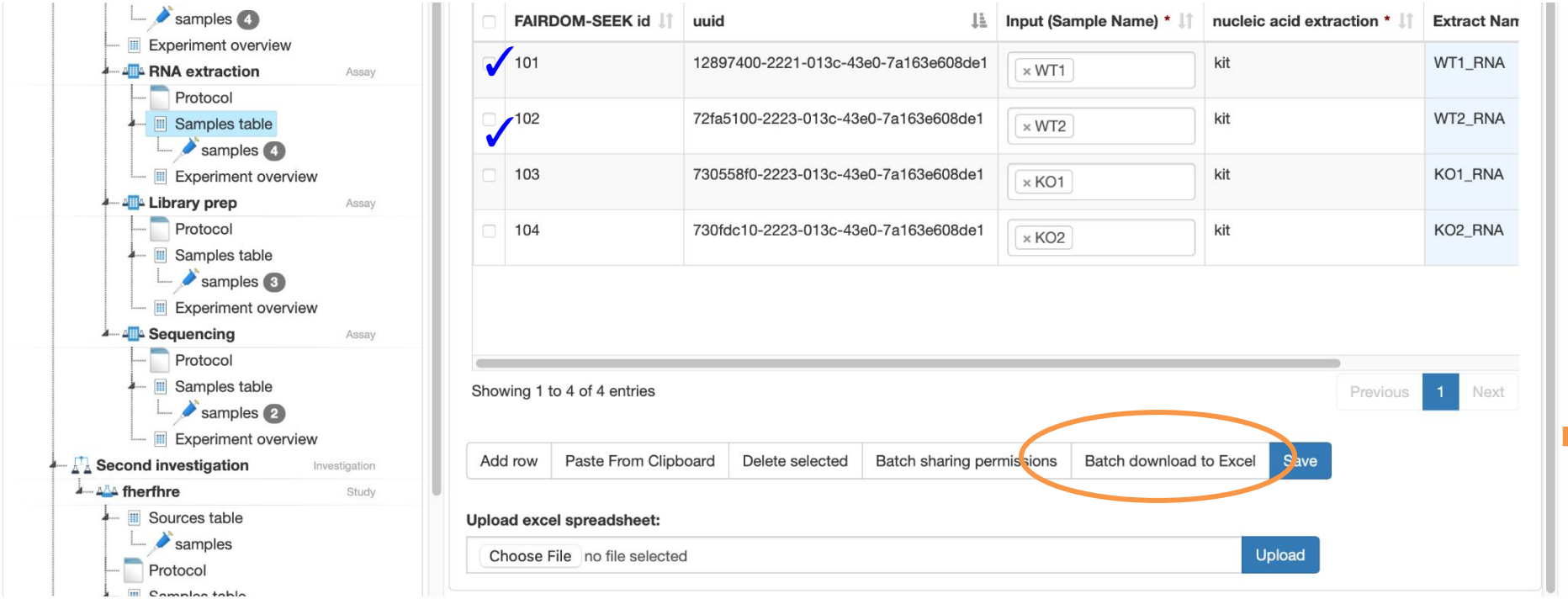


Batch sharing permissions for samples from the dynamic table



FAIRDOM SEEK in DataHub

Download to Excel

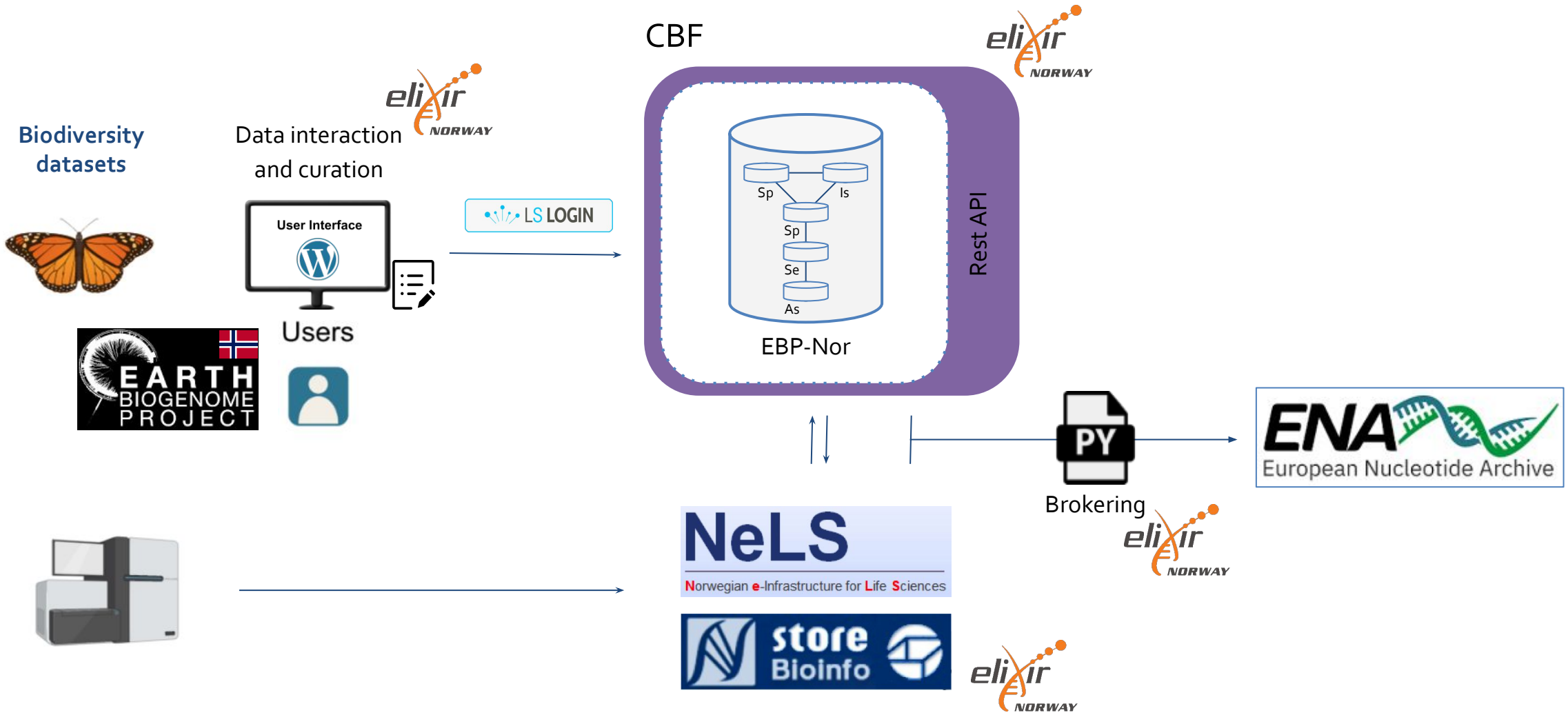


A	B	C	D	E	F	G
id	uuid	Input (Sample Name) *	nucleic	Extract Na	Material typ *	
101	12897400-2221-013c-43e0-7a163e608de1	[{"id"=>97, "type"=>"Sample", "title"=>"WT1"}]	kit	WT1_RNA	RNA	
102	72fa5100-2223-013c-43e0-7a163e608de1	[{"id"=>98, "type"=>"Sample", "title"=>"WT2"}]	kit	WT2_RNA	RNA	

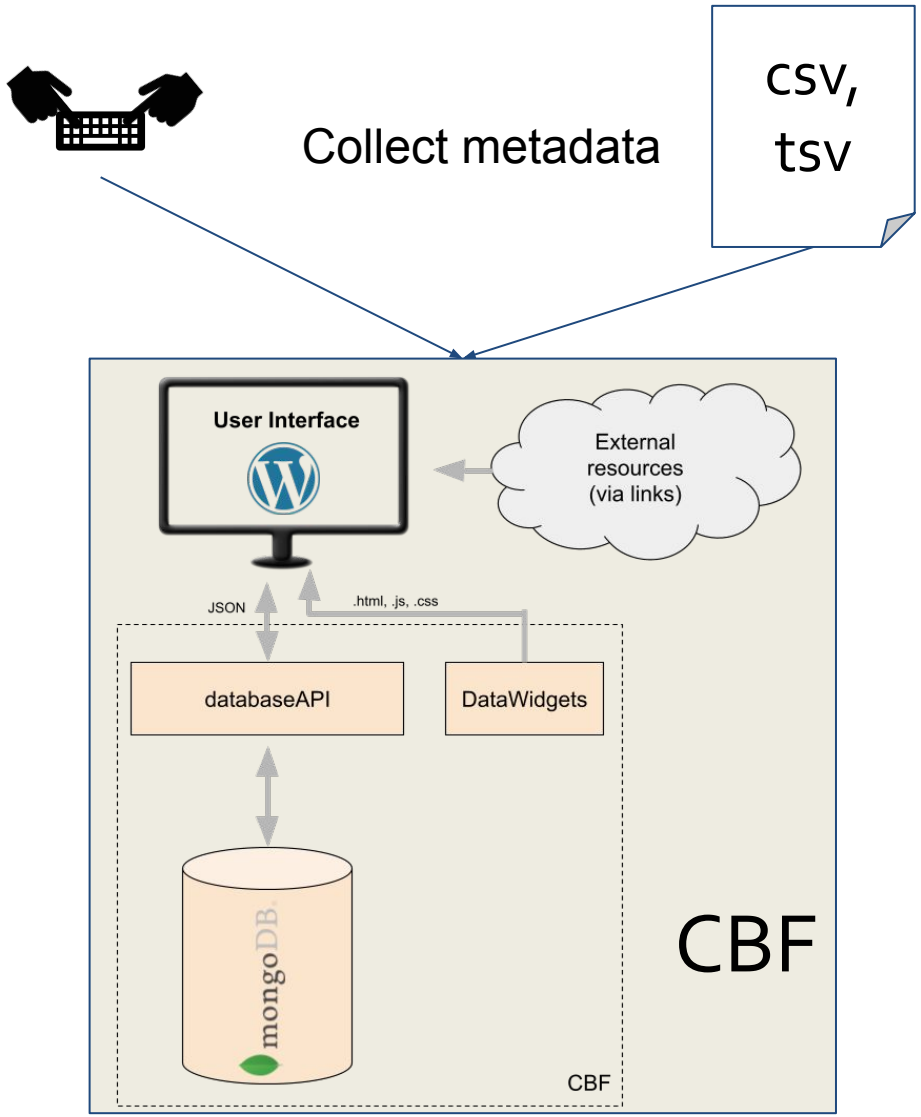
Locked cells

Material type
Choose a valid option. This value is REQUIRED!

EBP-Nor use case - Brokering



Contextual Biodata Framework (CBF)



Visualize & explore

- Plots & timelines
- Cart & geo location
- Rich filtering

Share

- API
- Download

Communicate

- Wordpress plugin (easy web integration)

FAIR

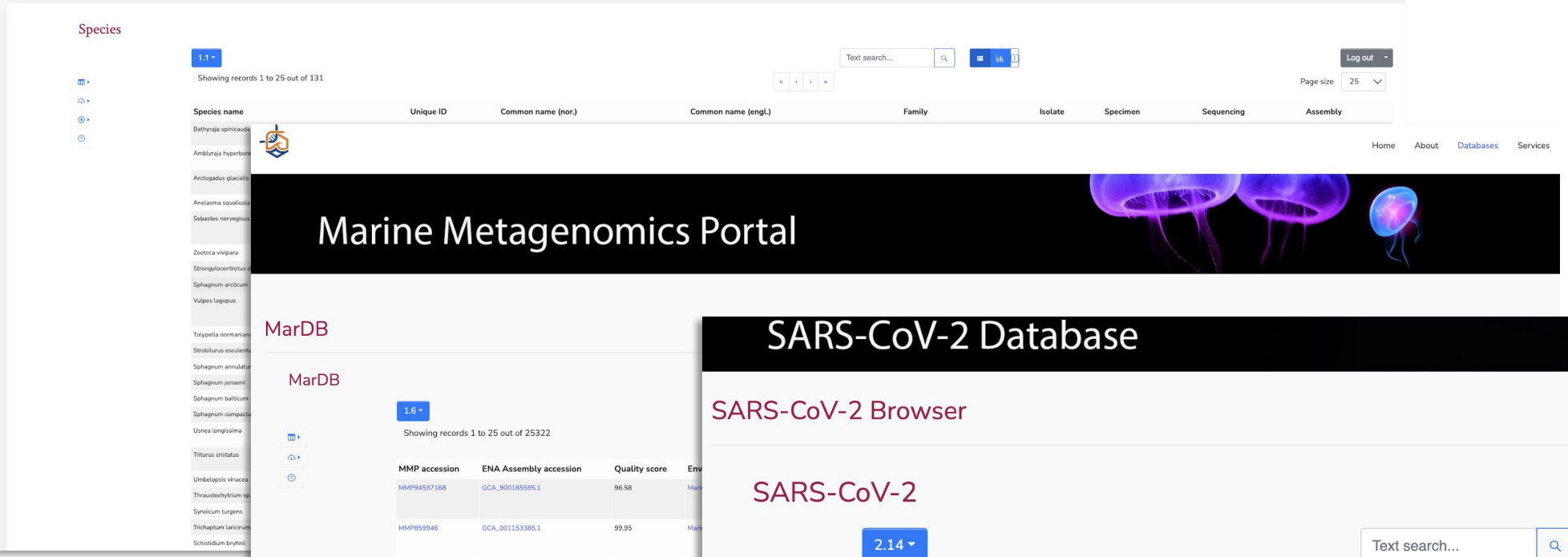
- Versioning
- Linked data
- Annotations

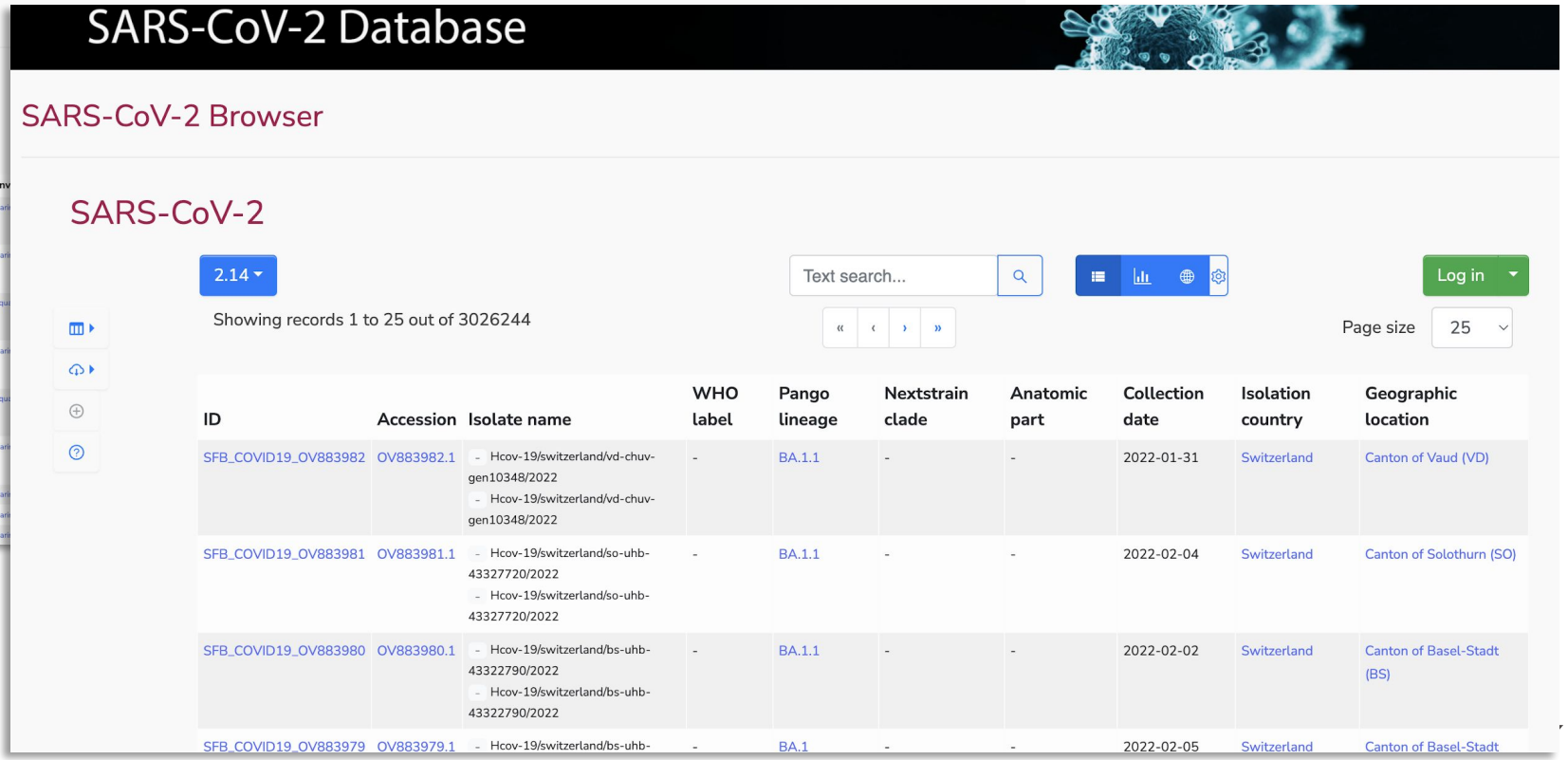
Slide: Terje Klemetsen

EBP-Nor

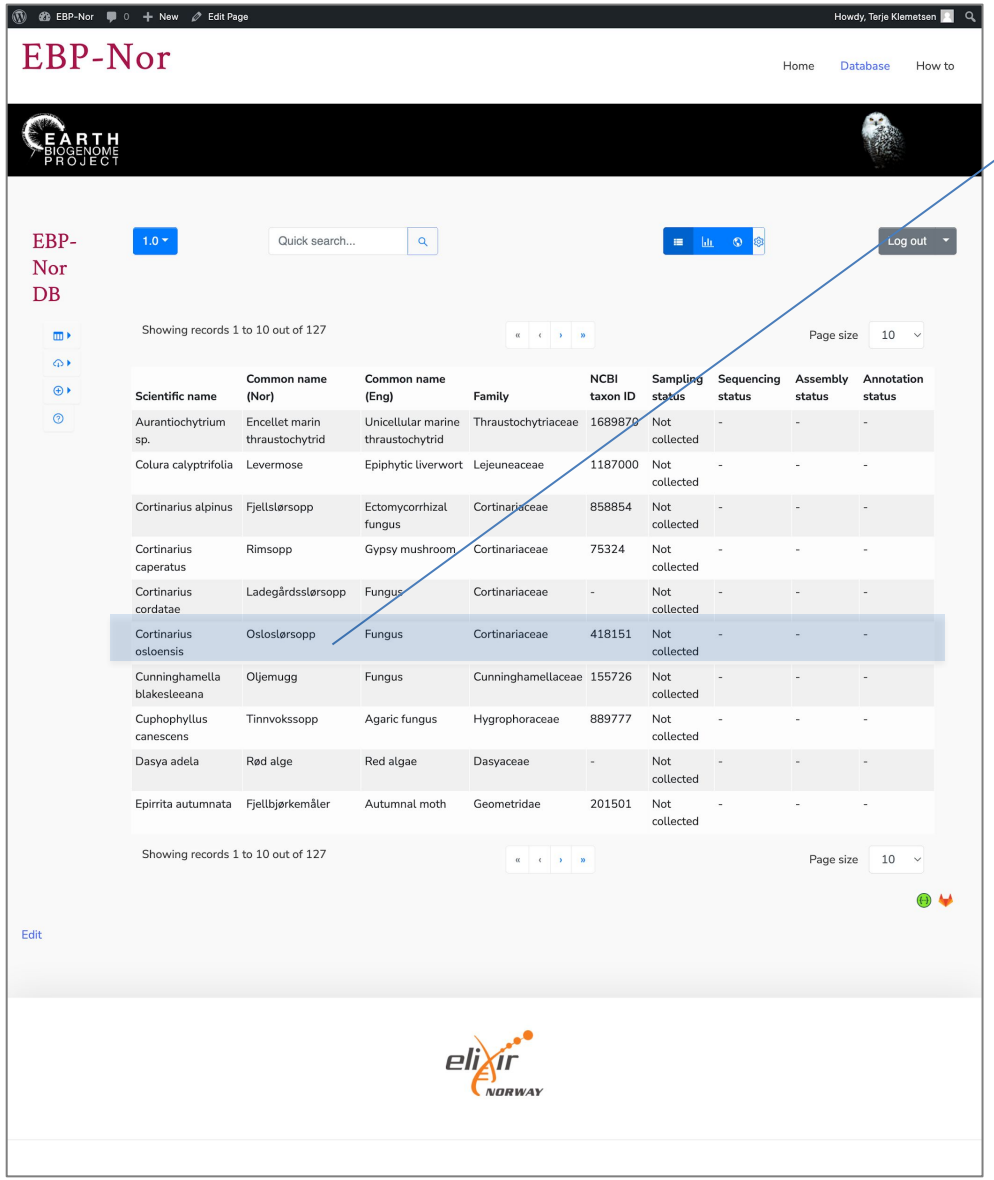
CBF

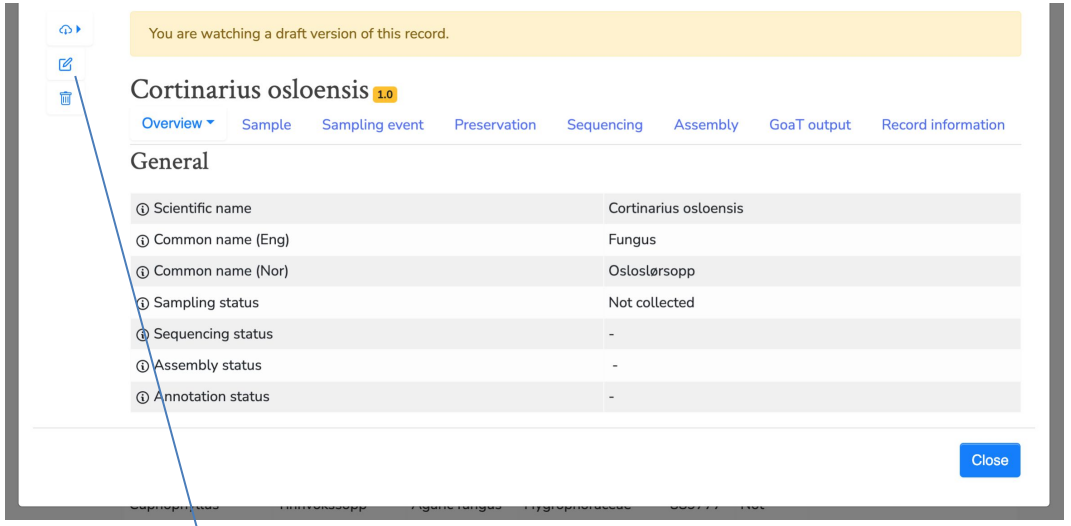

EARTH
BIOGENOME
PROJECT

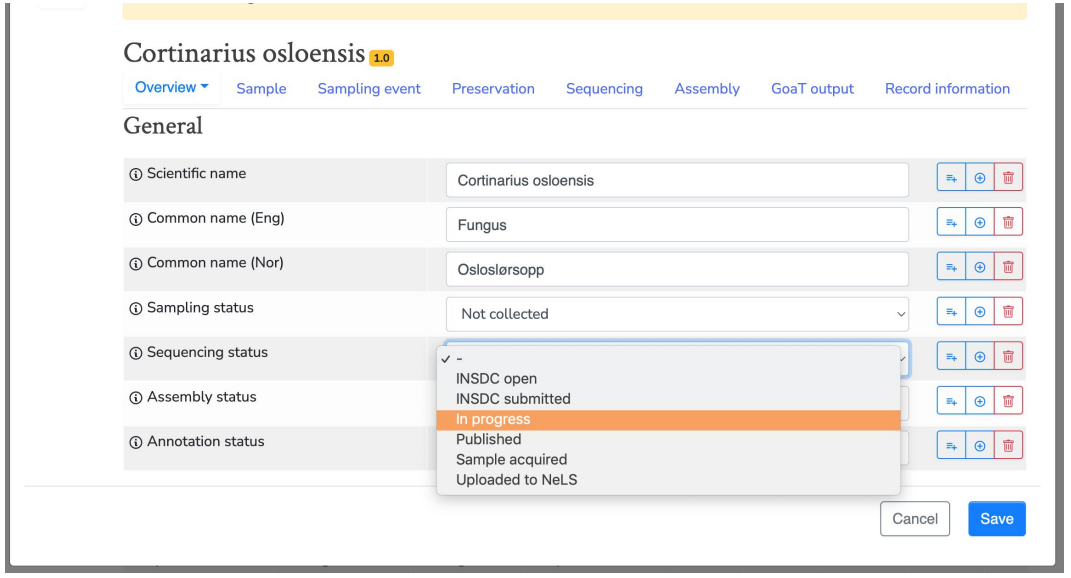

Species
Showing records 1 to 25 out of 131
Species name
Unique ID
Common name (nor.)
Common name (engl.)
Family
Isolate
Specimen
Sequencing
Assembly
Home
About
Databases
Services
Marine Metagenomics Portal
MarDB
MarDB
Showing records 1 to 25 out of 25322
MMP accession
ENA Assembly accession
Quality score


SARS-CoV-2 Database
SARS-CoV-2 Browser
SARS-CoV-2
Text search...
Log in
Showing records 1 to 25 out of 3026244
Page size
25
ID
Accession
Isolate name
WHO label
Pango lineage
Nextstrain clade
Anatomic part
Collection date
Isolation country
Geographic location
SFB_COVID19_OV883982
OV883982.1
Hcov-19/switzerland/vd-chuv-gen10348/2022
Hcov-19/switzerland/vd-chuv-gen10348/2022
BA.1.1
2022-01-31
Switzerland
Canton of Vaud (VD)
SFB_COVID19_OV883981
OV883981.1
Hcov-19/switzerland/so-uhb-43327720/2022
Hcov-19/switzerland/so-uhb-43327720/2022
BA.1.1
2022-02-04
Switzerland
Canton of Solothurn (SO)
SFB_COVID19_OV883980
OV883980.1
Hcov-19/switzerland/bs-uhb-43322790/2022
Hcov-19/switzerland/bs-uhb-43322790/2022
BA.1.1
2022-02-02
Switzerland
Canton of Basel-Stadt (BS)
SFB_COVID19_OV883979
OV883979.1
BA.1
2022-02-05
Switzerland
Canton of Basel-Stadt

Functionalities - Editing an entry

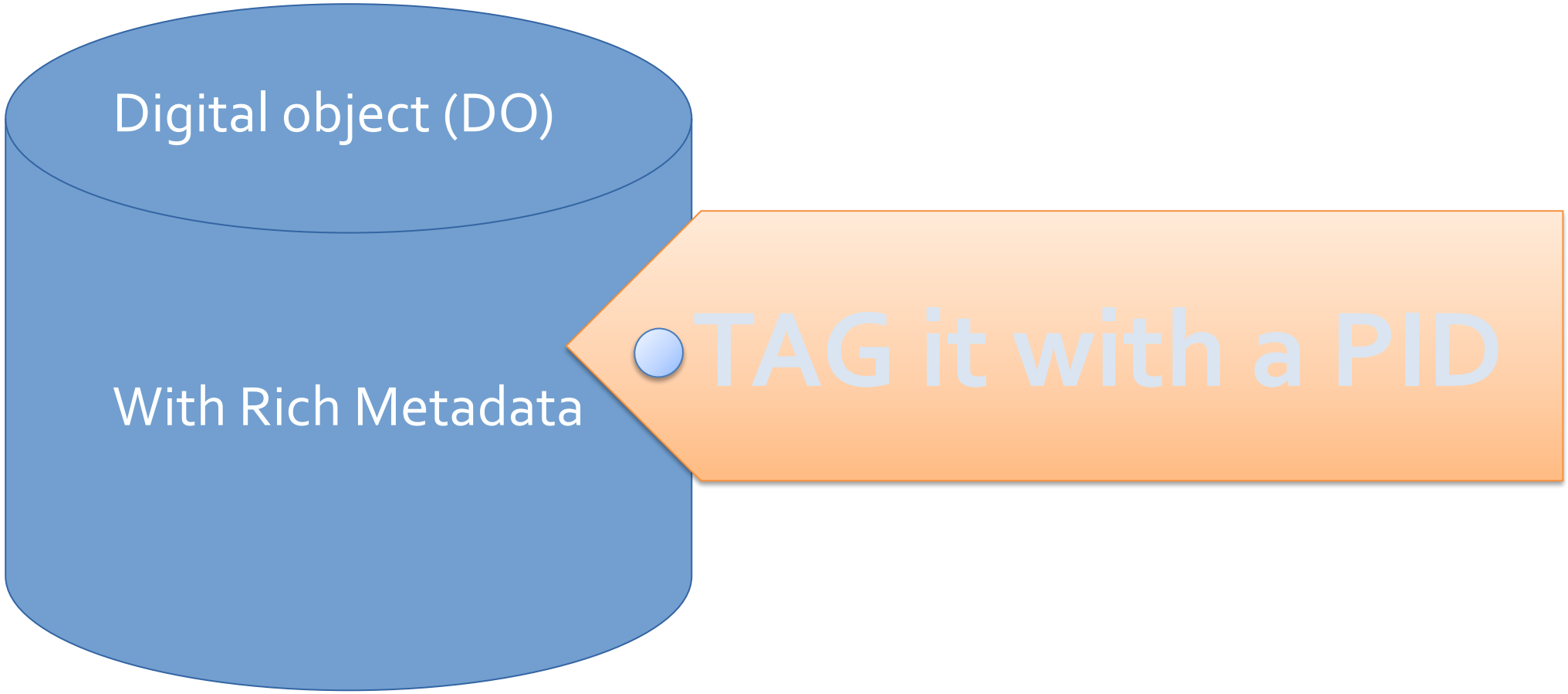






Slide: Terje Klemetsen

Make it visible



PIDs helps make data FAIR

Data should be **Findable**	F1. (meta)data are assigned a globally unique and persistent identifier (DOI) F2. data are described with rich metadata F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource
Data should be **Accessible**	A1. (meta)data are retrievable by their identifier using a standardized communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary A2. metadata are accessible, even when the data are no longer available
Data should be **Interoperable**	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles I3. (meta)data include qualified references to other (meta)data
Data should be **Reusable**	R1. meta(data) are richly described with a plurality of accurate and relevant attributes R1.1. (meta)data are released with a clear and accessible data usage license R1.2. (meta)data are associated with detailed provenance R1.3. (meta)data meet domain-relevant community standards

Medyckyj-Scott, David et al. (2016).

Why don’t I just use a link (URL)?

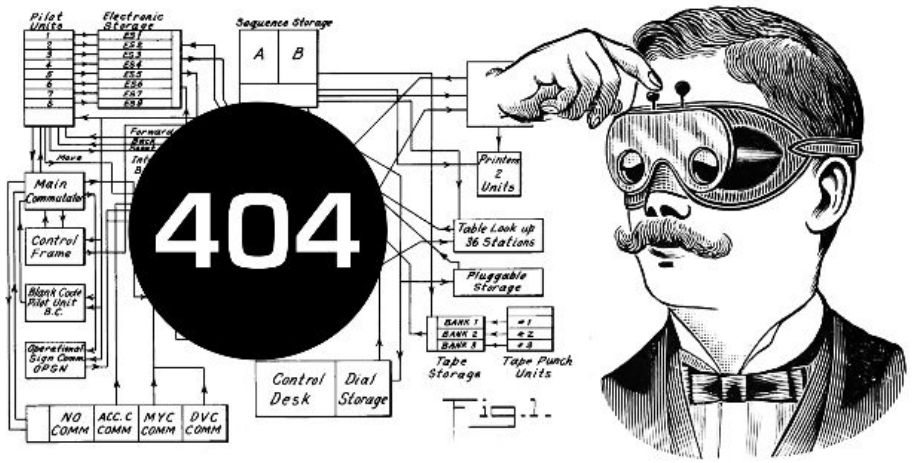
25. Supplemental data showing the predicted secondary structures of each construct (Fig. 3) and explaining the ligation activity of truncated ribozymes (Fig. 2B) are available at *Science* Online at www.sciencemag.org/feature/data/1050240.shl.

farm3.staticflickr.com

“LINK ROT”

Science Contents News Careers Journals

Read our COVID-19 research and news.



Hmmm...

This doesn't *look* like science.

It seems you're in search of a page that doesn't exist, or may have moved. You can use the Back button in your browser to return to the page that brought you here, or **search for your missing page**.

PID = PDI = GUID

PID = Persistent Identifier
PDI = Persistent Digital Identifier
GUID = Globally Unique Identifier

Physical objects: a dog, building, microscope, star, person etc

It doesn't "rot"

A persistent identifier (PID) is a long-lasting reference to a resource

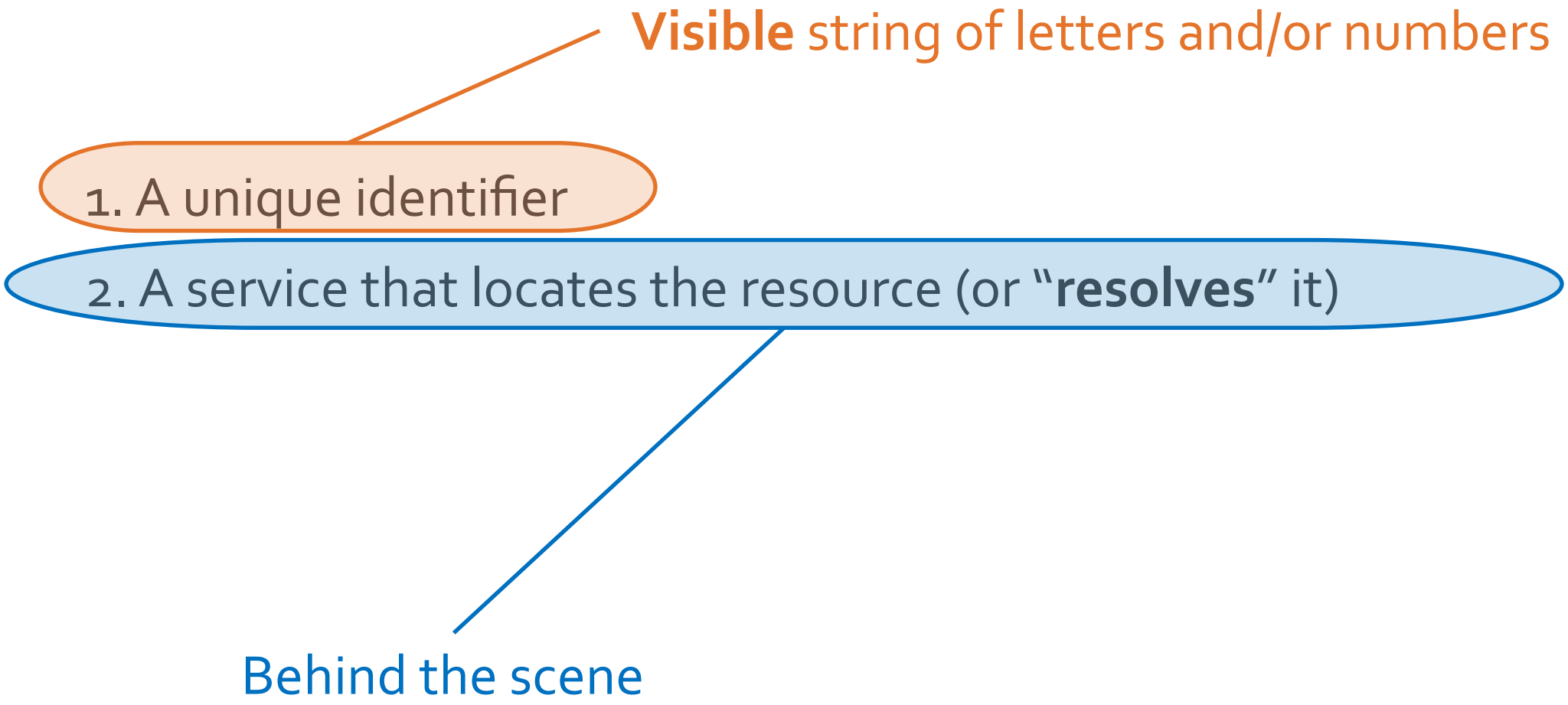
Somebody commits to keep it alive

globally unique string of characters

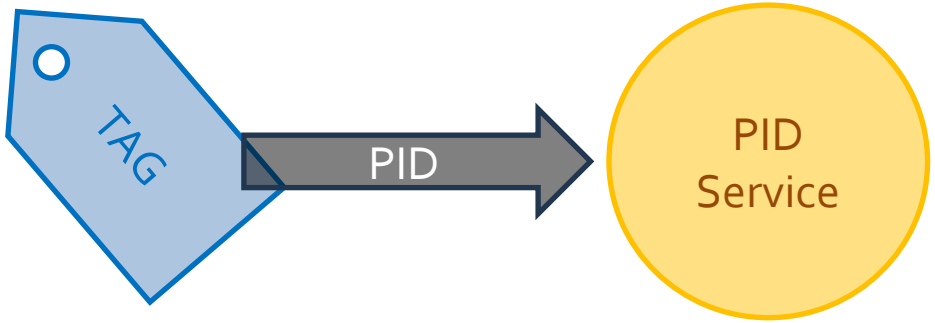
People AND computers can find it

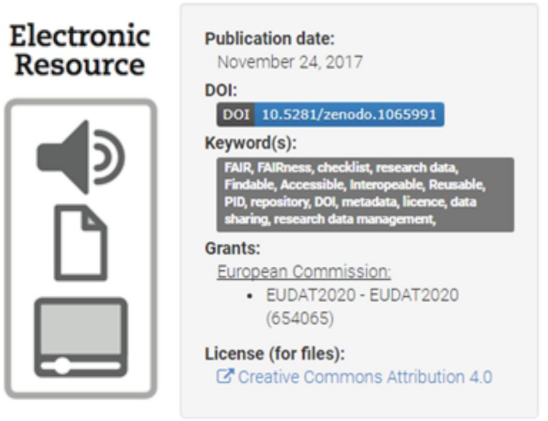
Digital Objects: data, collections, metadata, software, publications, configurations, categories, workflows etc

A PID consists of two components:

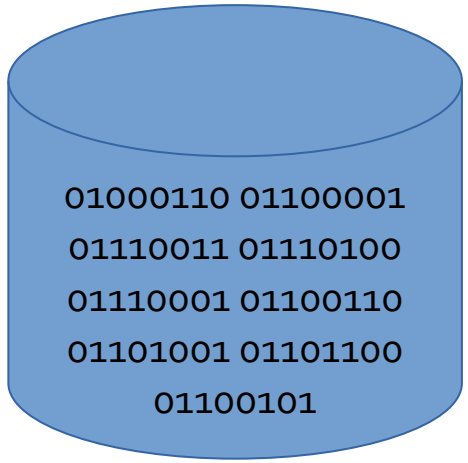


Priciple:





https://www.clarin.eu/sites/default/files/pid-CLARIN-ShortGuide.pdf



Example

DOI

10.25820/data.006172

Citation

Van Benschoten, W. Z., & Shepherd, J. J. (2022). *Dataset for "Piecewise Interaction Picture Density Matrix Quantum Monte Carlo"* [Data set]. University of Iowa. https://doi.org/10.25820/data.006172

URL

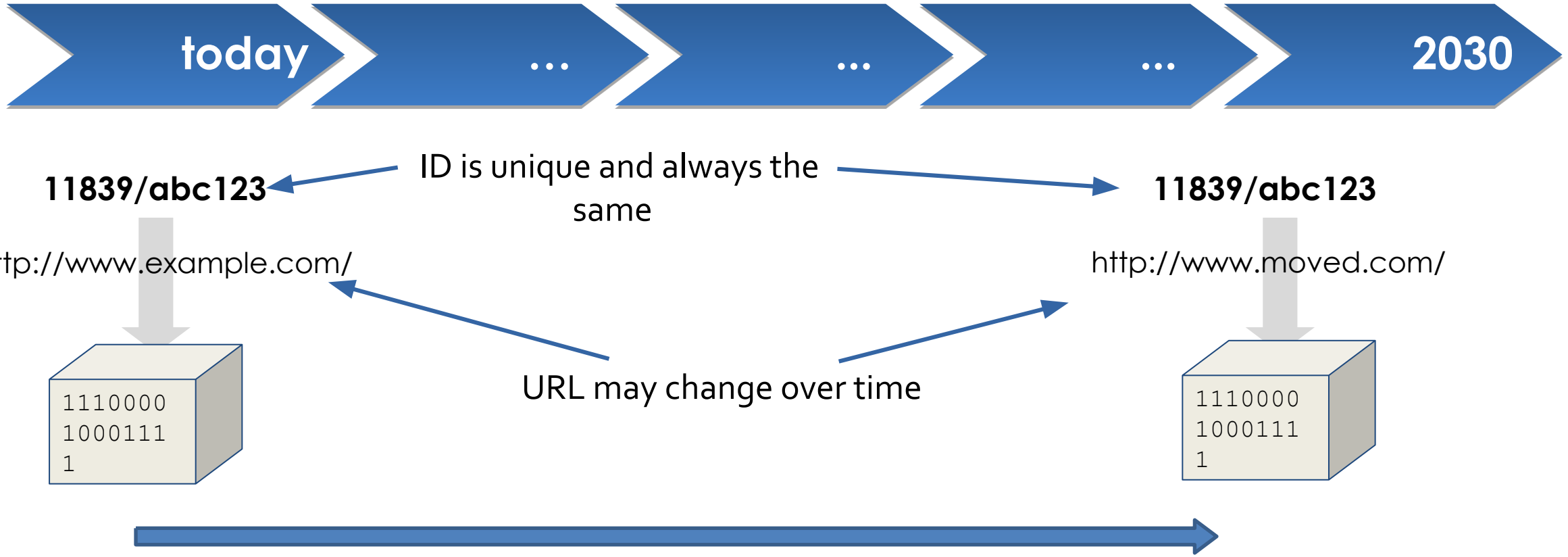
https://iro.uiowa.edu/esploro/outputs/dataset/9984240535802771

Landing Page

https://blog.lib.uiowa.edu/news/2023/03/07/the-power-of-persistent-identifiers-in-data-sharing/

Persistent over time

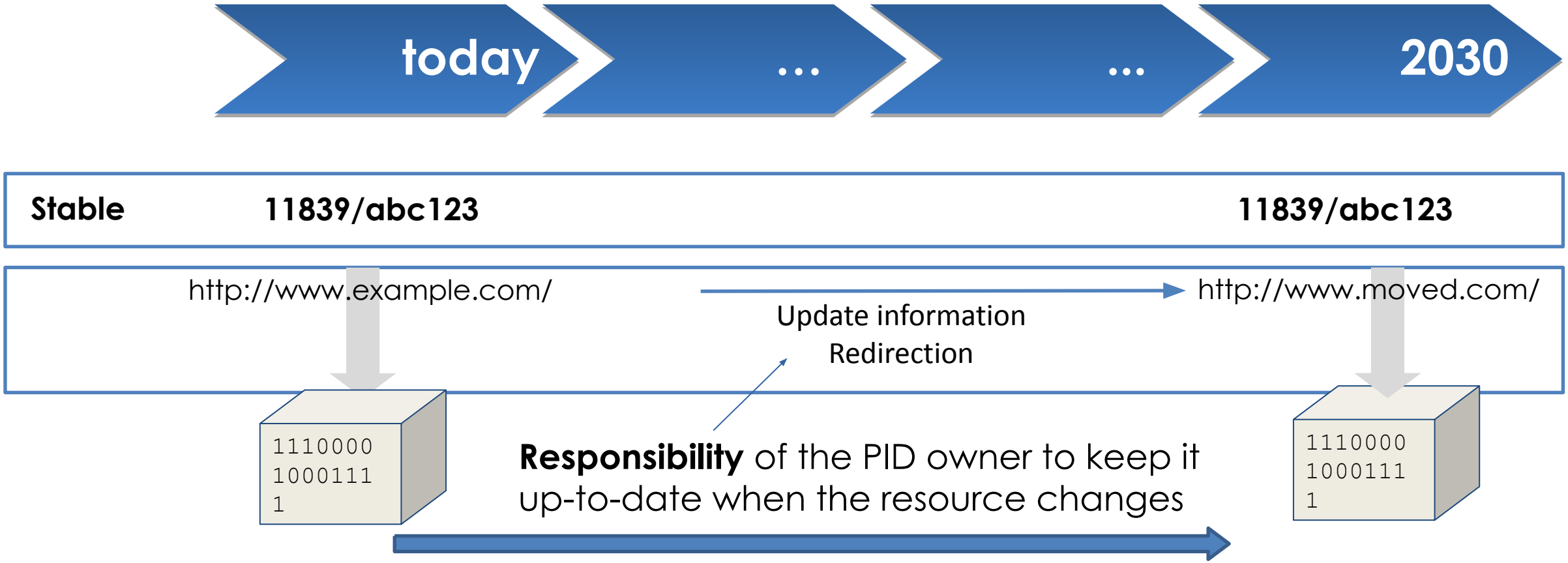
.. by design



Supports access to resource as it moves from one location to another.

Persistent over time

.. by design



Different systems

Some Common Identifiers:

Digital Object Identifiers (doi:10.1186/2041-1480-3-9)

Handles (hdl:2381/12775)

URN (urn:isbn:0451450523)

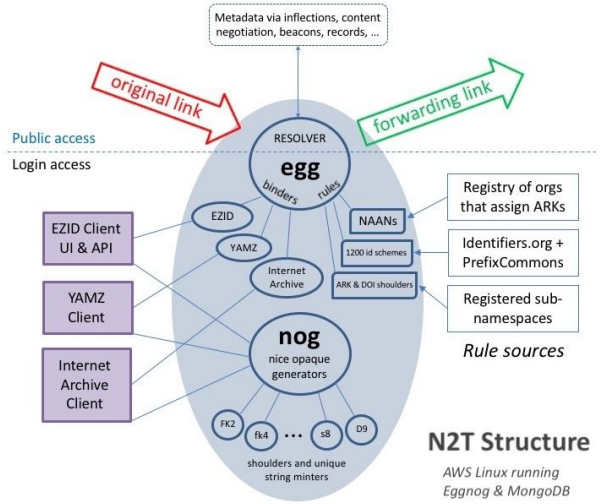
Archival Resource Keys (ARK) (ark:/13030/tf5p30086k)

Persistent Uniform Resource Locator (PURL)

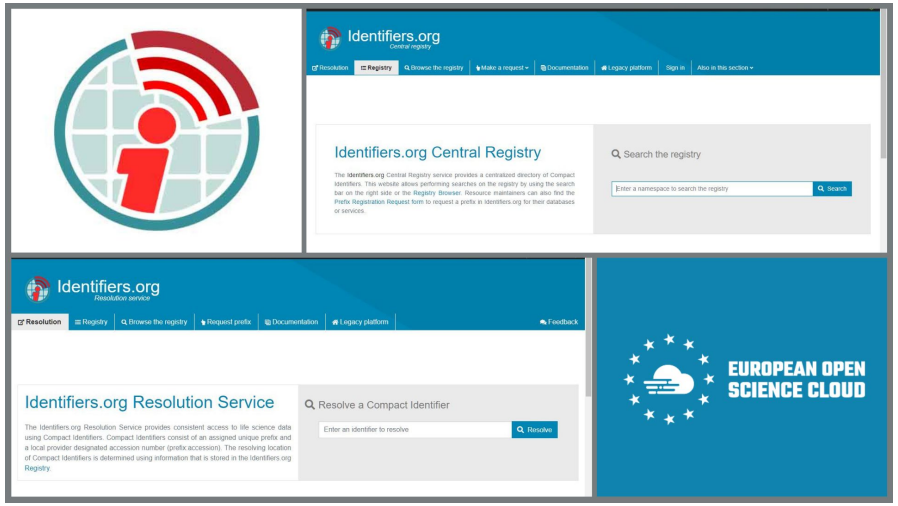
Resolver Services

N2T (Name-to-Thing)

Identifiers.org



https://arks.org/about/n2t-global-resolver/



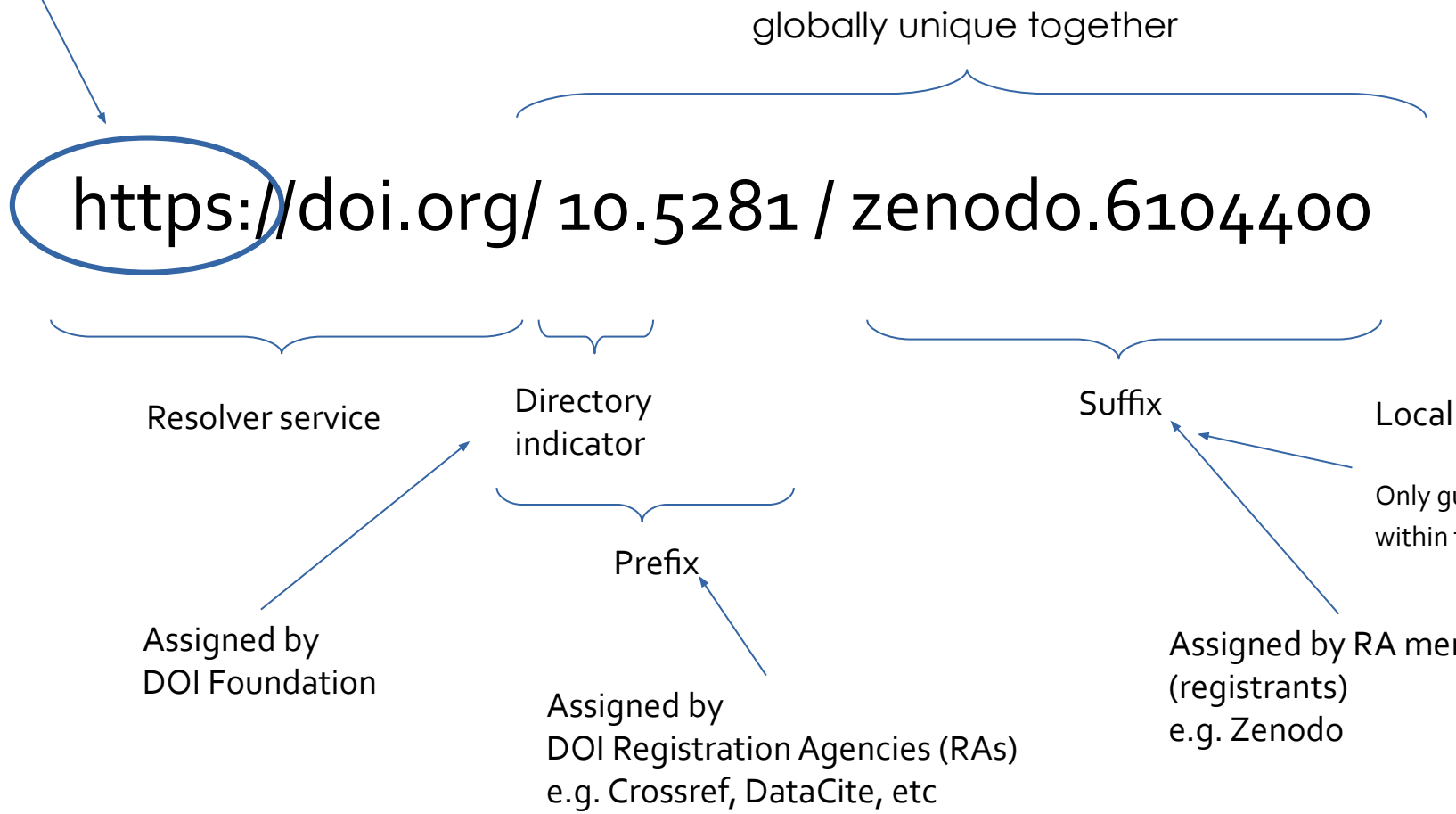
https://eosc-portal.eu/news-and-events/news/identifiers-ensuring-robust-and-reliable-access-life-sciences-data

How do I recognize a PID?

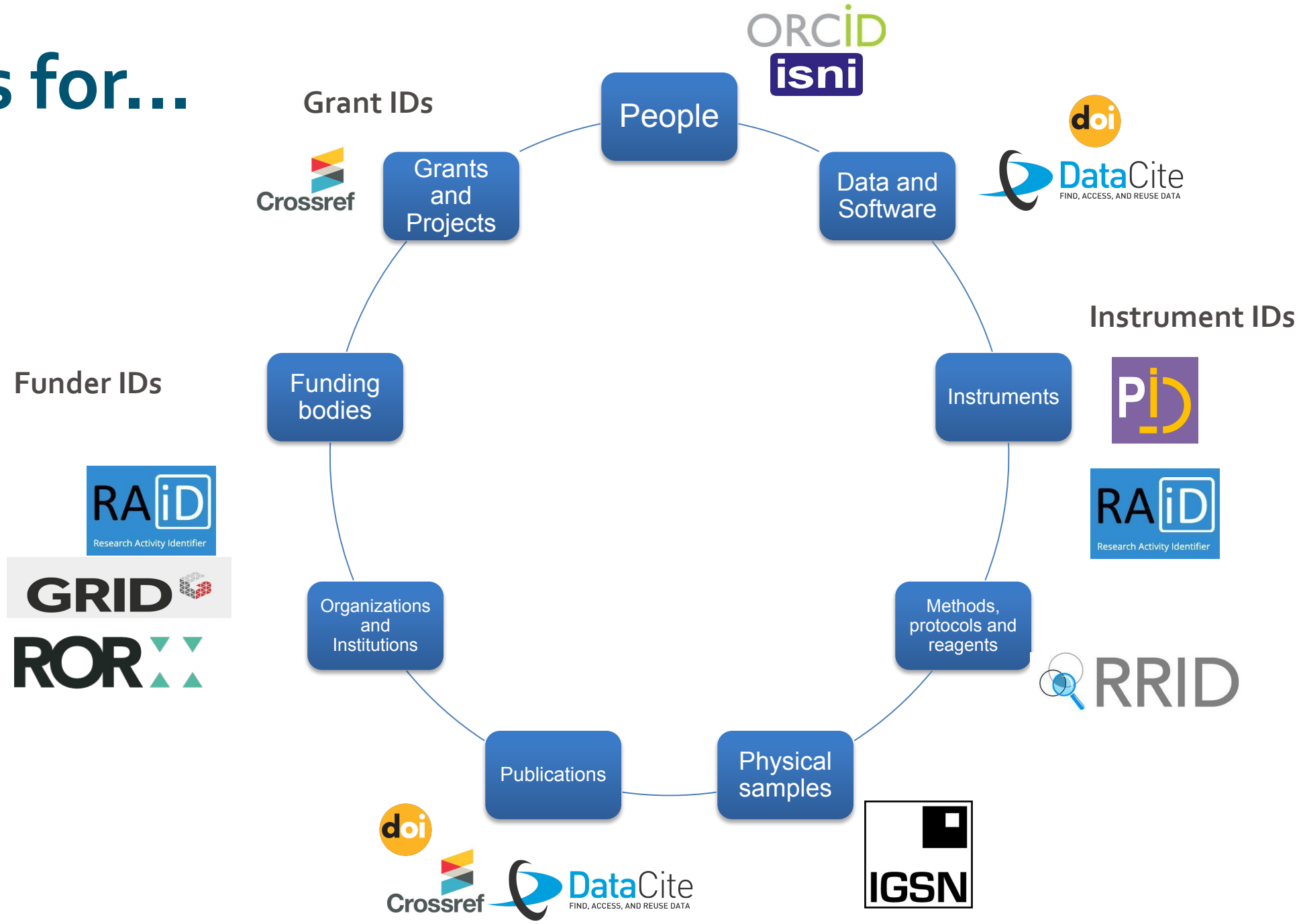
DOI: 10.5281/zenodo.6104400

Anatomy of a DOI

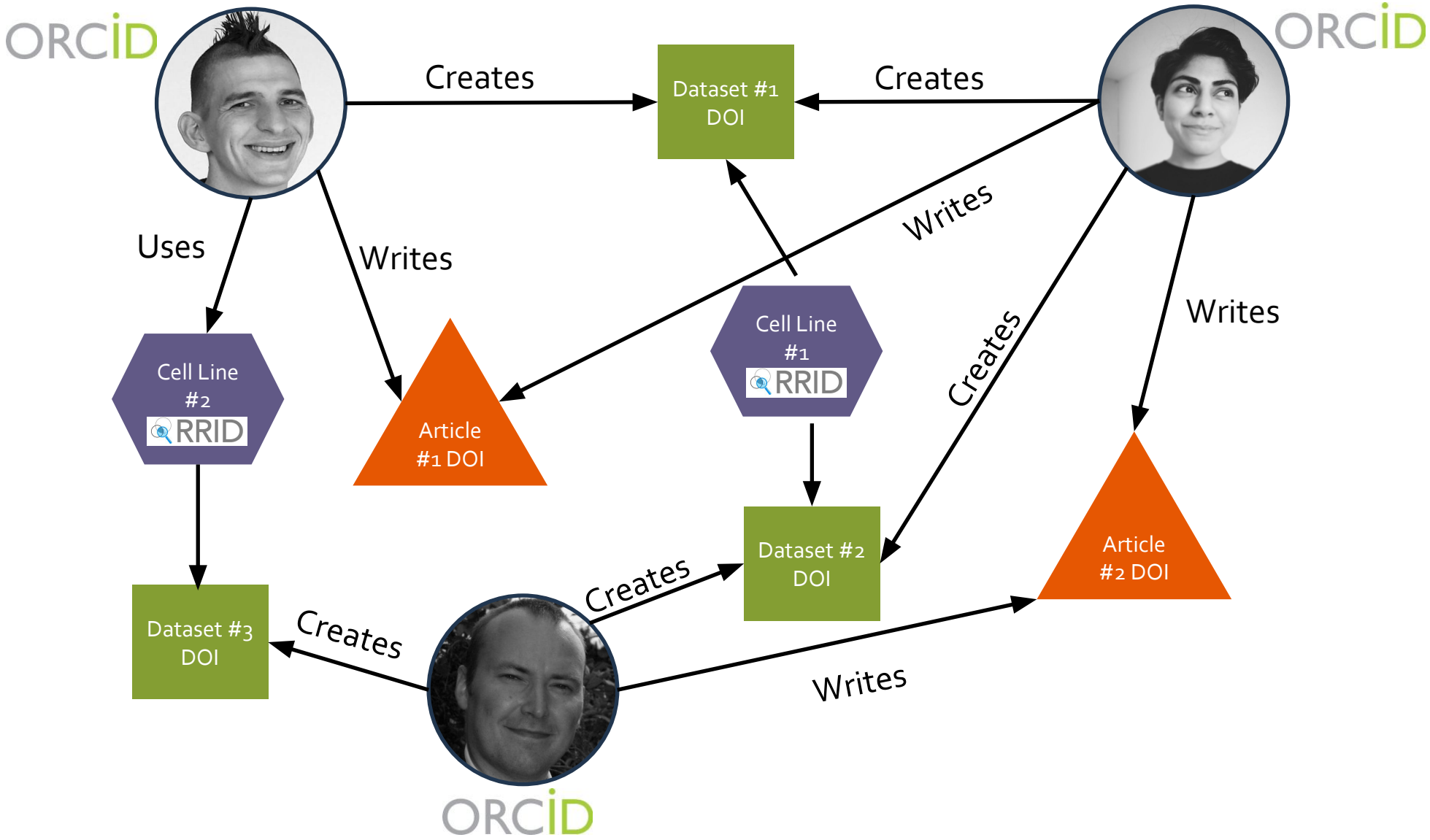
Means that it is actionable: you can paste in a web browser address bar and be taken to the identified source.



PIDs for...

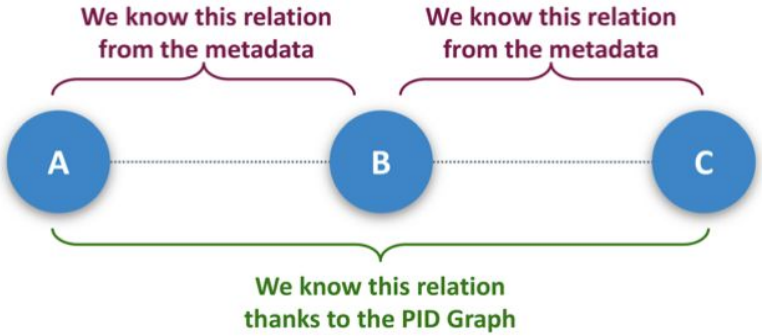

ORCID
isni
People
Grant IDs
Crossref
Grants and Projects
Data and Software
doi
DataCite
FIND, ACCESS, AND REUSE DATA
Instrument IDs
Funder IDs
Funding bodies
Instruments
PID
RAiD
Research Activity Identifier
GRID
ROR
Organizations and Institutions
Methods, protocols and reagents
RRID
Publications
Physical samples
doi
Crossref
DataCite
FIND, ACCESS, AND REUSE DATA
IGSN

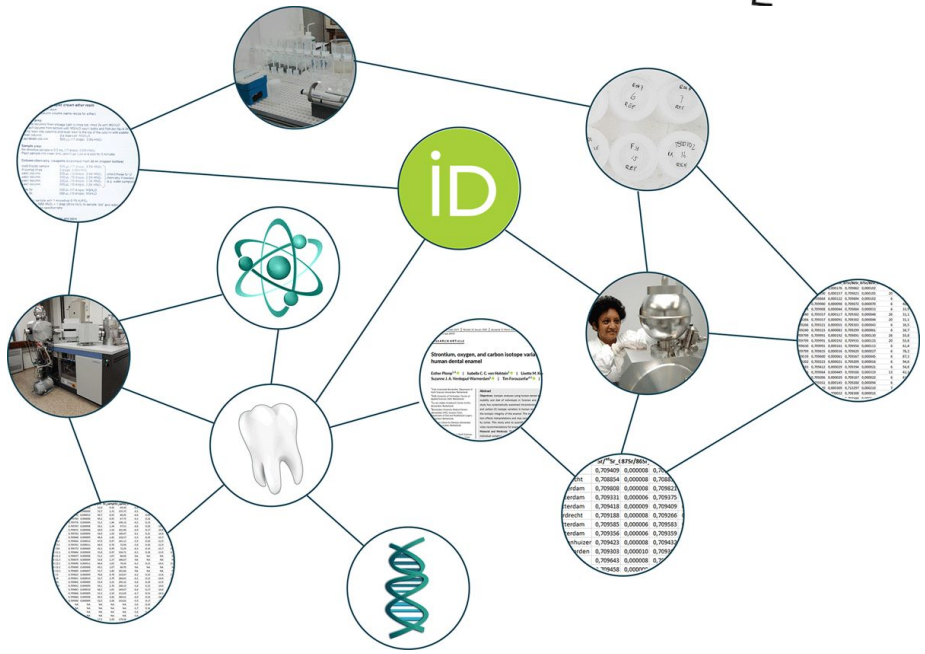
PIDs connect different entities in research



PID graphs

"I want to see all datasets funded by RCN cited by this article"





Crossref

DataCite
FIND, ACCESS, AND REUSE DATA

Crossref