



Metadata & Persistent identifiers



Espen Åberg
Data Steward
ELIXIR Norway/BioMedData

NeLS

Norwegian e-Infrastructure for Life Sciences



 **nor seq**



 **Galaxy**

 **BioStudies.**

 **ENA**
European Nucleotide Archive

 **IDR**

 **ArrayExpress**

 **MetaboLights**

 **JWS ONLINE**

 **PRIDE Archive**
Proteomics IDentifications Database



sensitive
data



 **EUROPEAN
GENOME-PHENOME
ARCHIVE**

- Data life cycle +
- Your role +
- Your domain +
- Your problem -
- Compliance monitoring
- Data analysis
- Data management plan
- Data organisation
- Data protection
- Data publication
- Data quality
- Data storage
- Data transfer
- Identifiers
- Licensing
- Documentation and metadata
- Sensitive data
- All tools and resources
- Tool assembly +



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

“Data is content, and metadata is context. Metadata can be much more revealing than data, especially when collected in the aggregate.”

— Bruce Schneier, Data and Goliath

“data on data”

What is metadata?



“information about something”

“If data is the new oil, metadata is the refinery”

— Adam Rauh

Experimental design

Outcome = Treatment effect + Biological effect + Technical effects + Error

Experimental design

Outcome = Treatment effect + Biological effect + Technical effects + Error

Environment

Compound

Infection

Inhibitor

siRNA

sgRNA

Dose

Time

Experimental design

Outcome = Treatment effect + Biological effect + Technical effects + Error

Environment

Compound

Infection

Inhibitor

siRNA

sgRNA

Dose

Time

Sex

Age

Weight

Litter

Genotype

Species

Cell line

Experimental design

Outcome = Treatment effect + Biological effect + Technical effects + Error

Environment

Compound

Infection

Inhibitor

siRNA

sgRNA

Dose

Time

Sex

Age

Weight

Litter

Genotype

Species

Cell line

Operator

Batch

Plate

Cage

Array

Flowcell

Instrument

Day

Order

Source

Experimental design

Outcome = Treatment effect + Biological effect + Technical effects + Error

Environment
Compound
Infection
Inhibitor
siRNA
sgRNA
Dose
Time

Sex
Age
Weight
Litter
Genotype
Species
Cell line

Operator
Batch
Plate
Cage
Array
Flowcell
Instrument
Day
Order
Source

Experimental
Treatment
Sampling
Measurement

Experimental design

“Data”

“Metadata”

Outcome = Treatment effect + Biological effect + Technical effects + Error

Environment
Compound
Infection
Inhibitor
siRNA
sgRNA
Dose
Time

Sex
Age
Weight
Litter
Genotype
Species
Cell line

Operator
Batch
Plate
Cage
Array
Flowcell
Instrument
Day
Order
Source

Experimental
Treatment
Sampling
Measurement

Experimental design

“Data”

“Metadata”

Outcome = Treatment effect + **Pi**

Environment

Comp

Experimental

Treatment

Sampling

Measurement

And when you put all of this in files, the names of the files, their storage descriptions, types, sizes, versions, how they relate to each other, access rights, licensing etc, all become metadata in your project.

Array

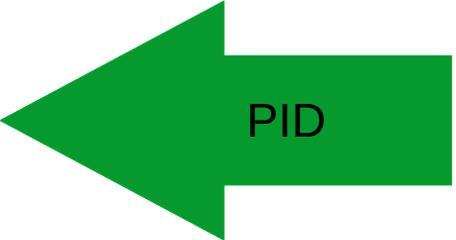
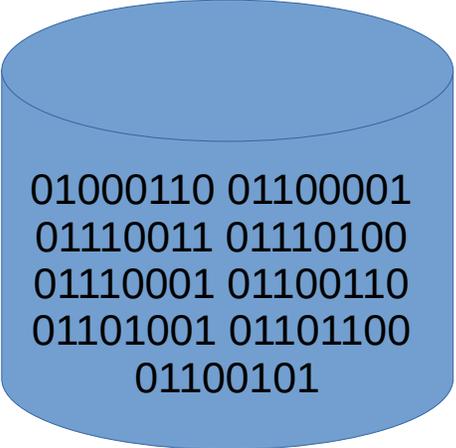
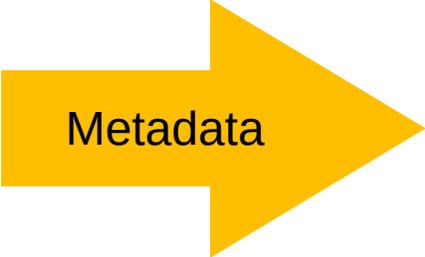
Flowcell

Instrument

Day

Order

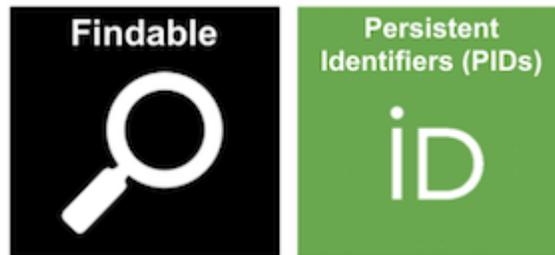
Source



PIDs helps make data FAIR

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

What is a persistent identifier (PID)?



PID?

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

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

Somebody commits to keeping it alive

globally unique string of characters

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

Why not just use a URL?

domain may change

resource may be
relocated

URL may change



“Link rot”



REPORT

One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds

Erik A. Schultes, David P. Bartel*

Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142, USA.

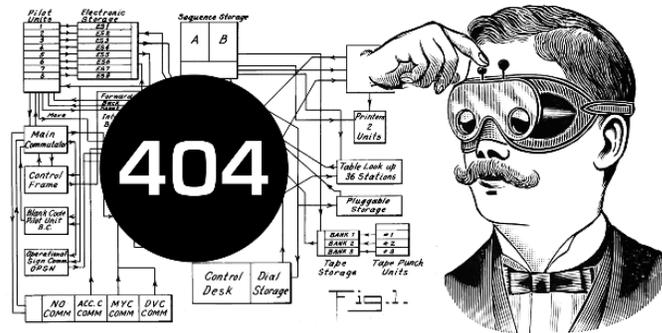
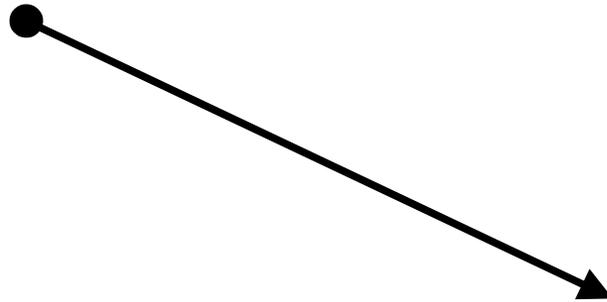
- Hide authors and affiliations

Science 21 Jul 2000:
Vol. 289, Issue 5478, pp. 448-452

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.

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.

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.



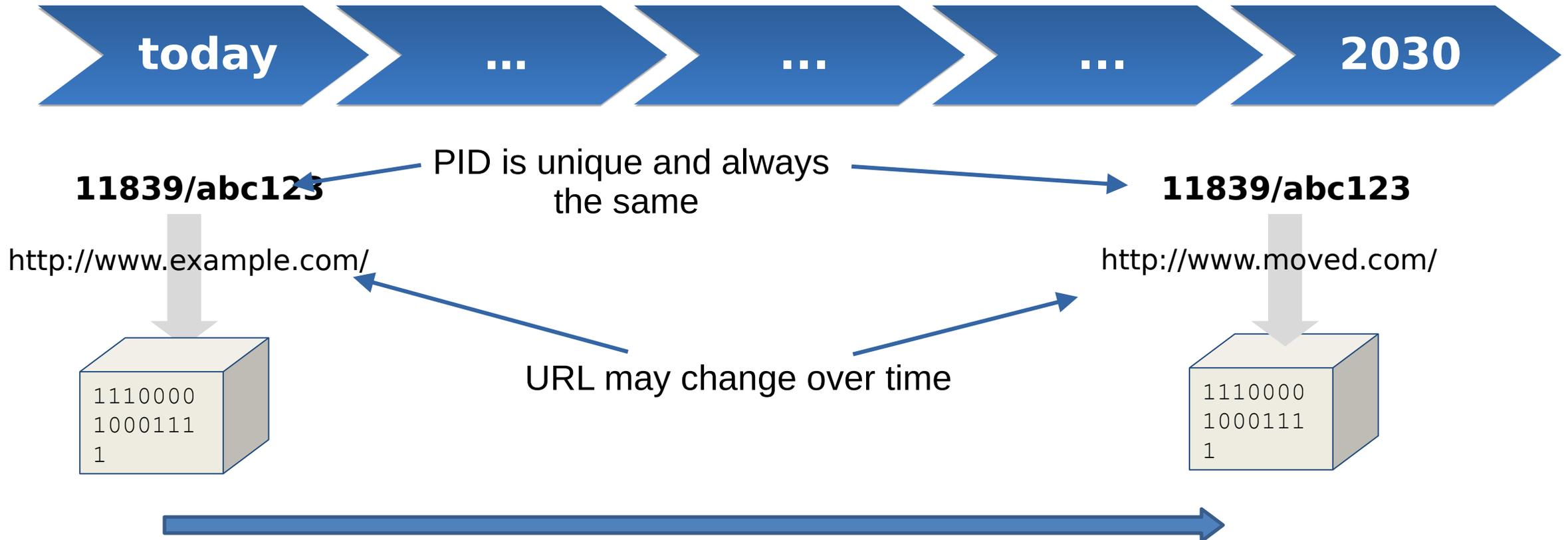
Hmmm...

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](#).

Persistent over time

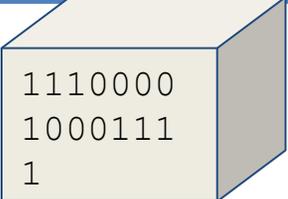
.. by design



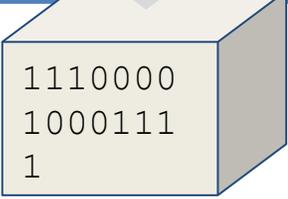
Supports access to resource as it moves from one location to another.

Persistent over time

.. by design



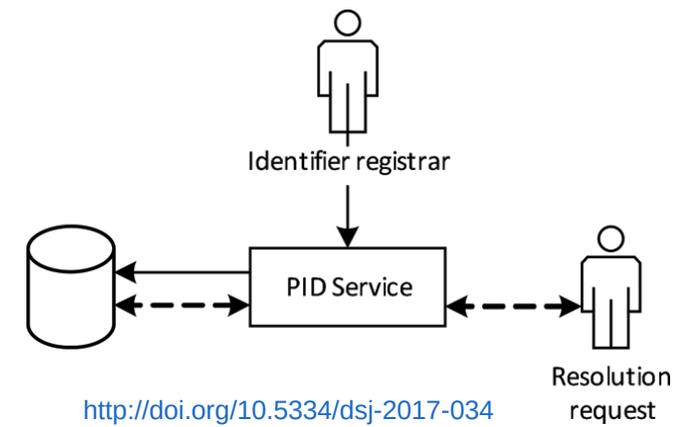
Responsibility of the PID owner to keep it up-to-date when the resource changes



A PID consists of 2 components:

1. a unique identifier

2. a service that locates the resource over time even when it's location changes



Examples for digital objects

Digital Object Identifiers



Handles

Handle.Net®

Archival Resource Keys (ARK)

Persistent Uniform Resource Locator
(URL)

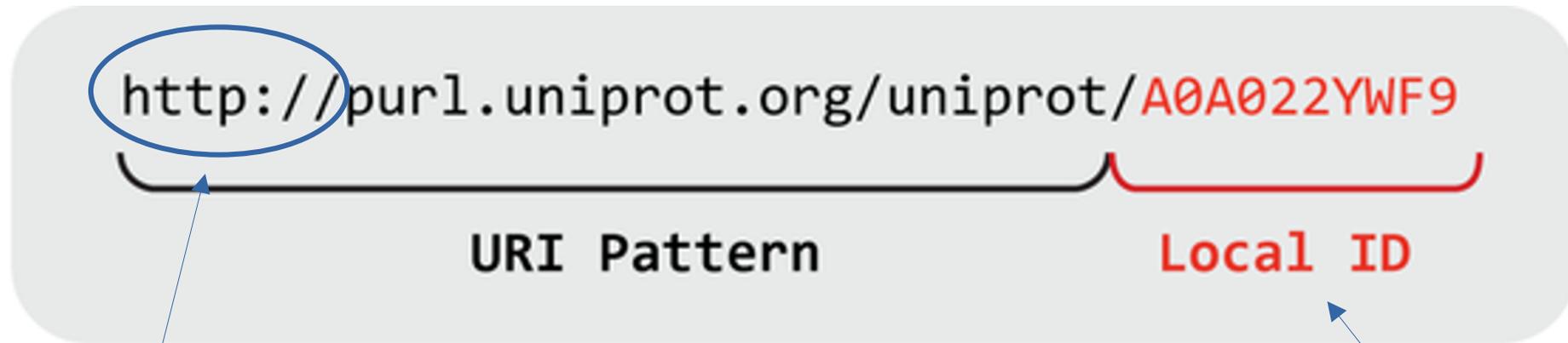


Identifiers.org

Anatomy of a web-based identifier

URL or URN

URI



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

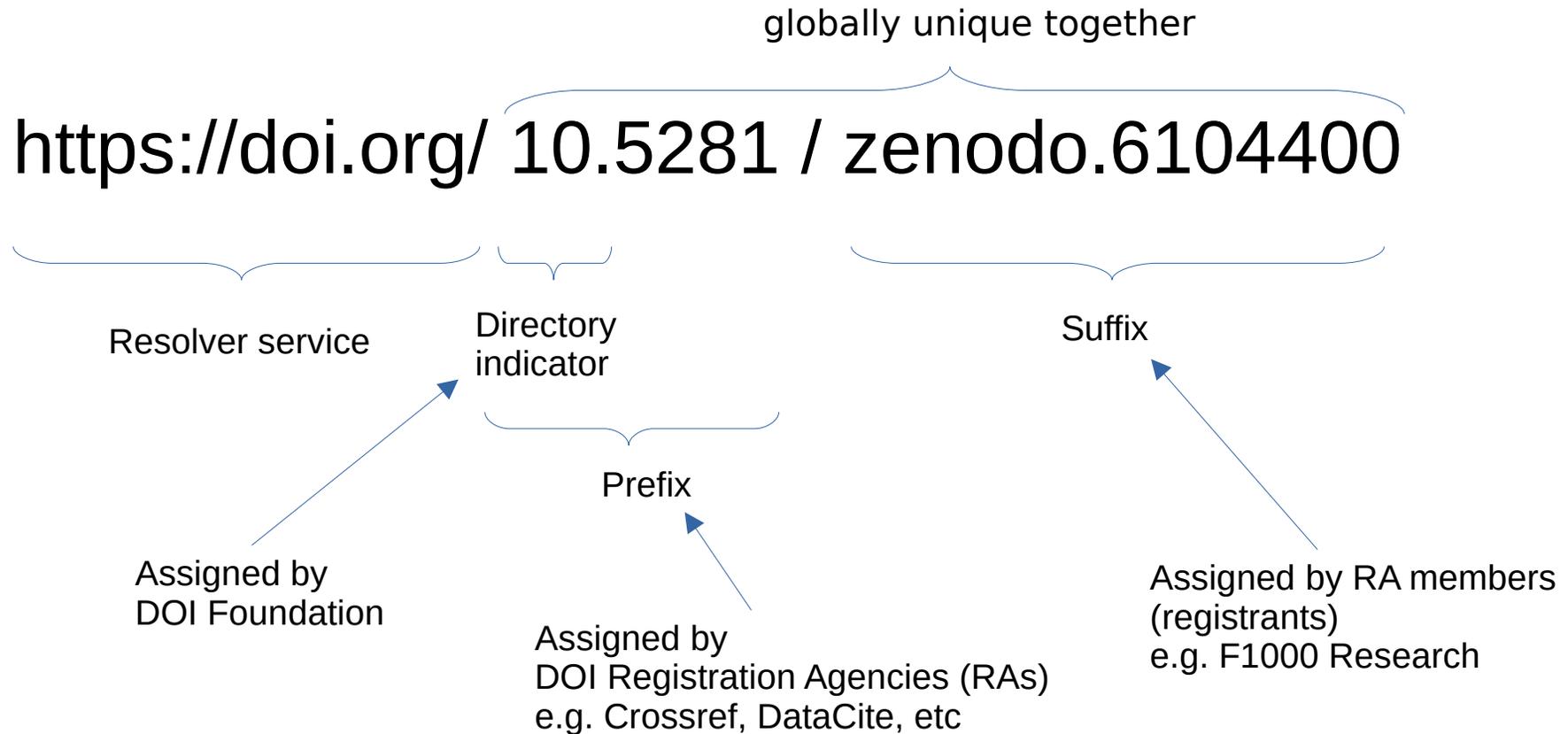
= an "Accession"

Only guaranteed to be locally unique within the database or source

How to recognize a PID

DOI: [10.5281/zenodo.6104400](https://doi.org/10.5281/zenodo.6104400)

Anatomy of a DOI



**Metadata
Description**

URL

**Electronic
Resource**



Handle

**Handle
Service**

URL



Publication date:

November 24, 2017

DOI:

DOI [10.5281/zenodo.1065991](https://doi.org/10.5281/zenodo.1065991)

Keyword(s):

FAIR, FAIRness, checklist, research data, Findable, Accessible, Interoperable, Reusable, PID, repository, DOI, metadata, licence, data sharing, research data management,

Grants:

European Commission:

- EUDAT2020 - EUDAT2020 (654065)

License (for files):

[CC BY](https://creativecommons.org/licenses/by/4.0/) Creative Commons Attribution 4.0

PIDs for

5

People

ORCID

isni

PIDs for

5

People

ORCID

isni

\$

Funding bodies


Crossref

PIDs for

5

People

ORCID

isni

\$

Funding bodies


Crossref

~

Institutions


GRID
Global Research Identifier Database

ROR 

PIDs for

5

People

ORCID

isni

\$

Funding bodies


Crossref

~

Institutions

GRID 

ROR 

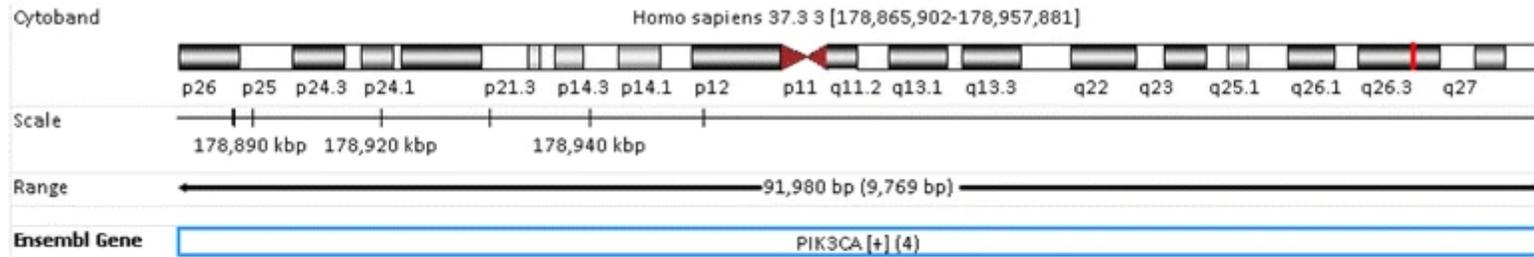


Instruments (soon)



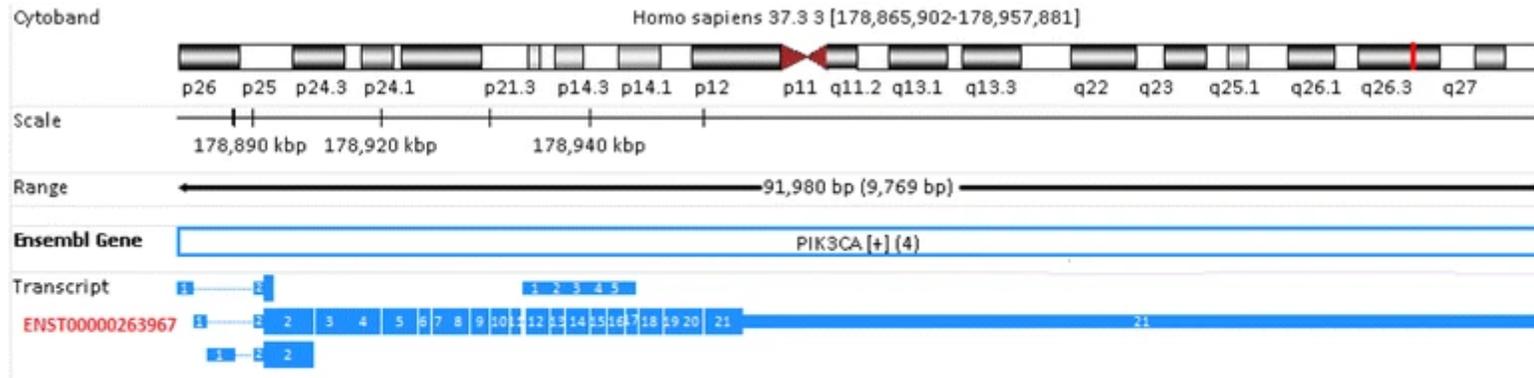
Sequence identifiers:

XXX Gene: PIK3CA



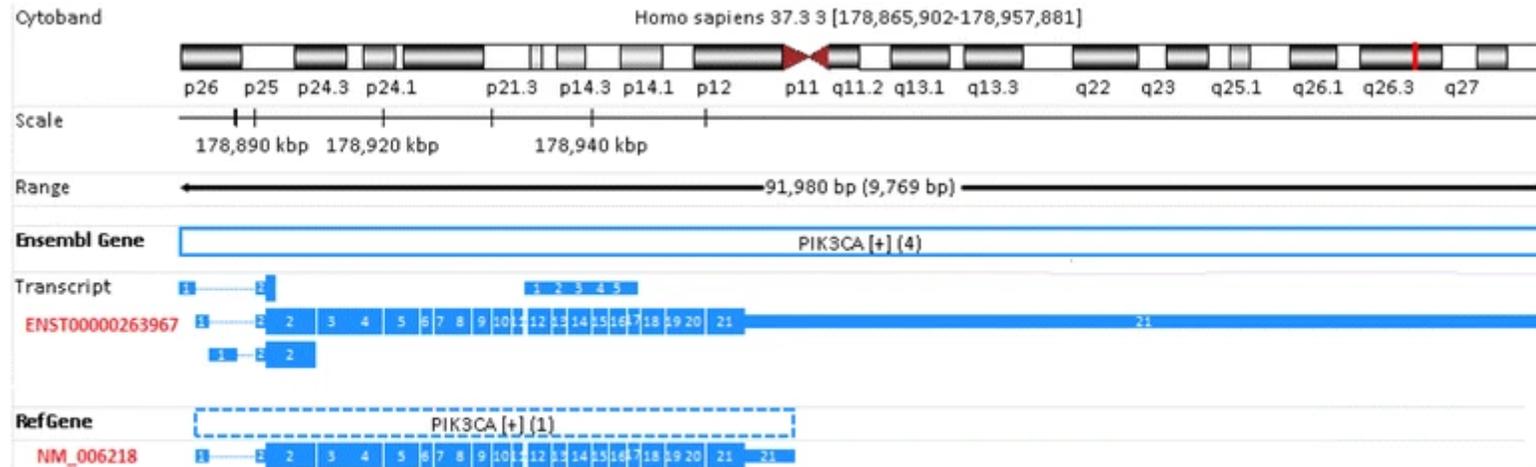
Sequence identifiers:

XXX Gene: PIK3CA



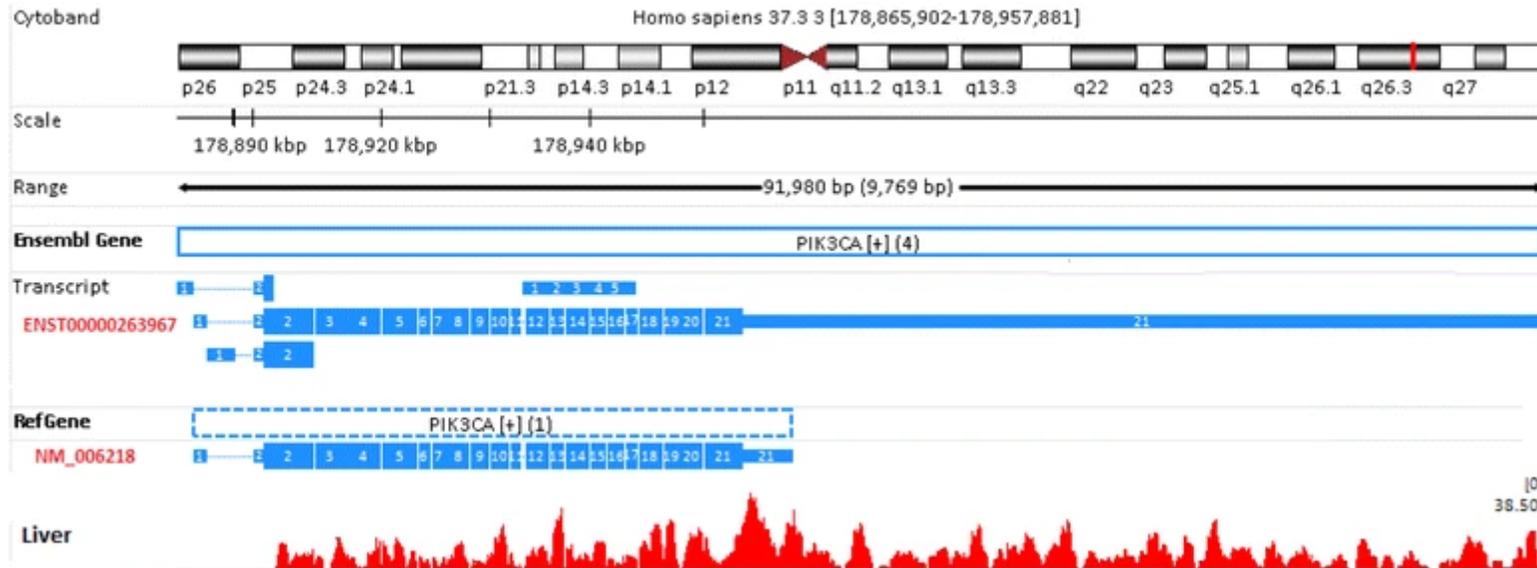
Sequence identifiers:

XXX Gene: PIK3CA



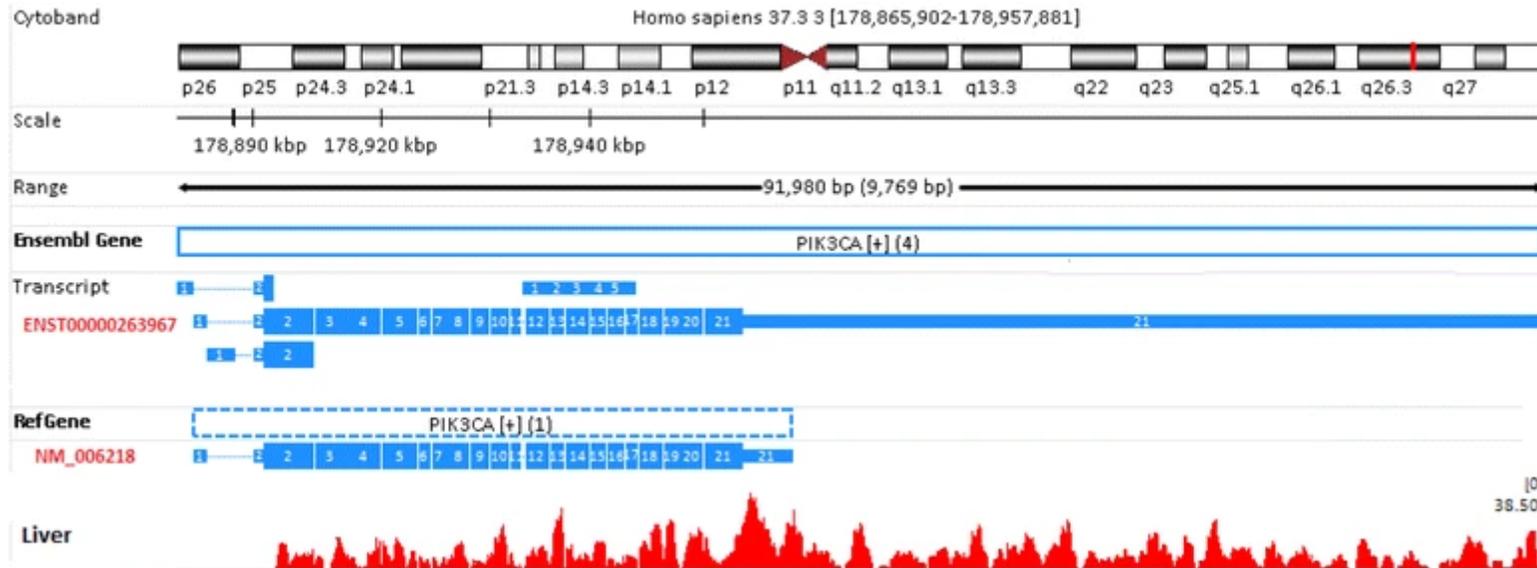
Sequence identifiers:

XXX Gene: PIK3CA



Sequence identifiers:

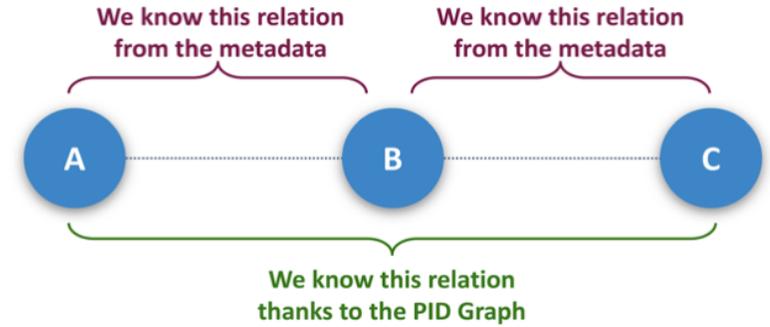
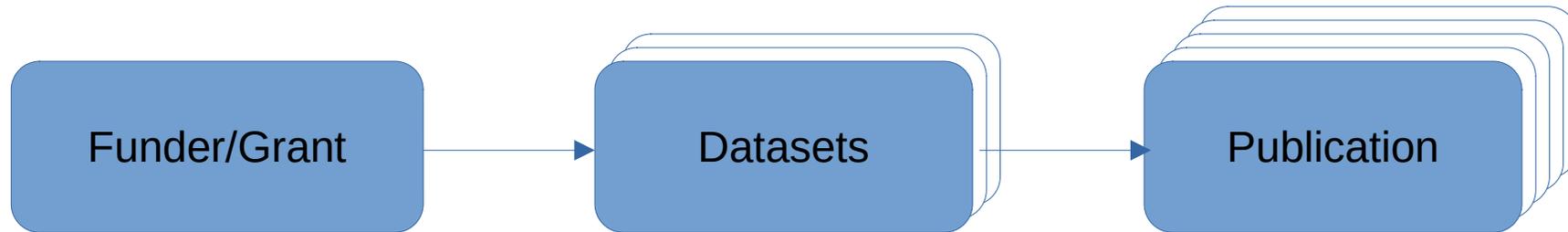
XXX Gene: PIK3CA



ENST00000263967.2
3
4

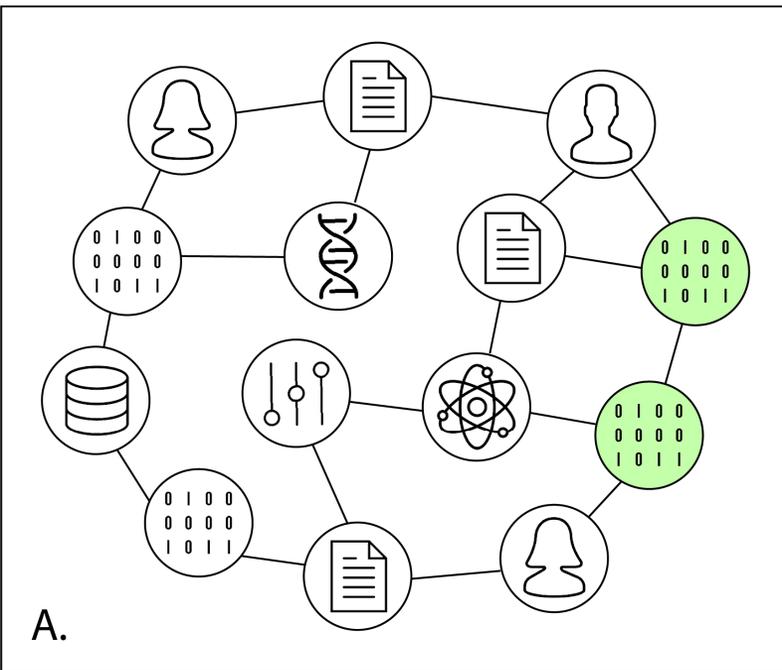
PIDs can do more

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

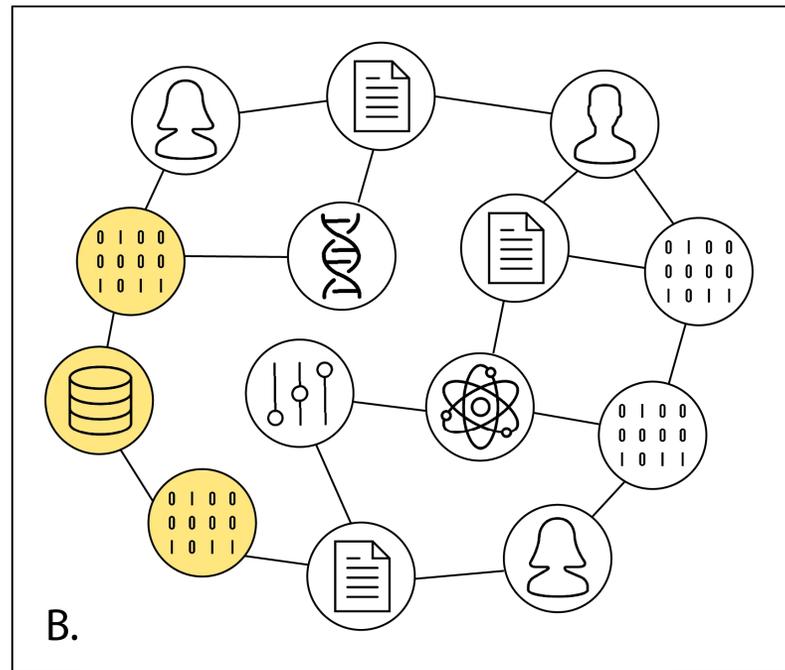




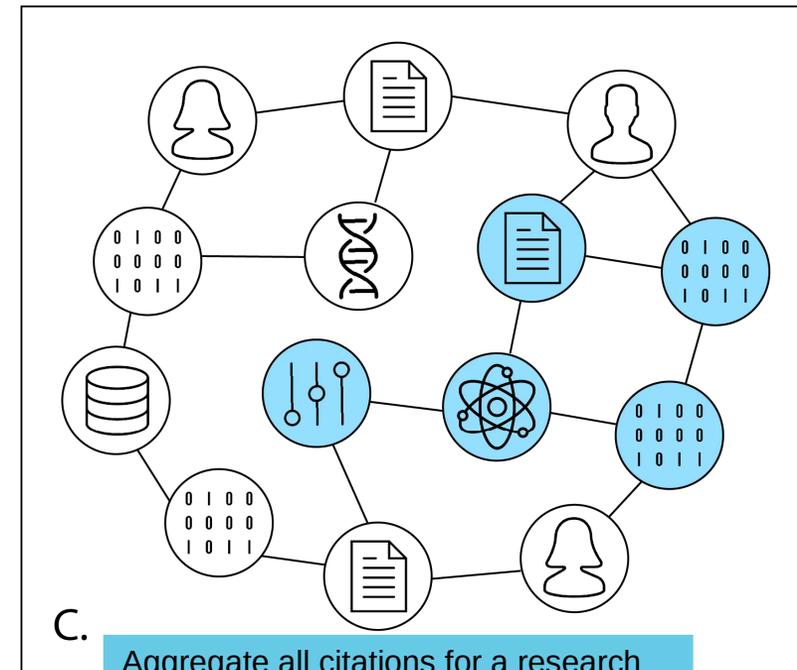
PID Graphs



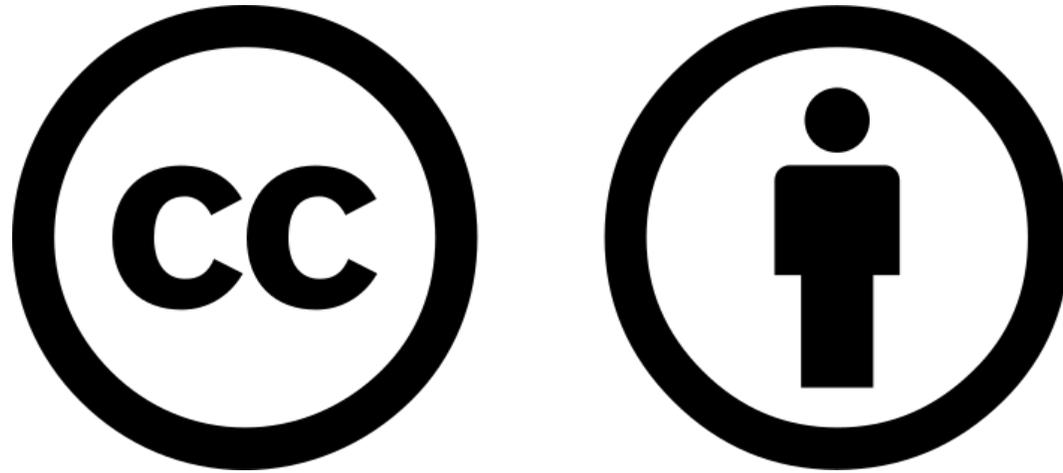
Aggregate the citations for all versions of a dataset or software source code



Aggregate the citations for all datasets hosted in a particular repository, funded by a particular funder, or created by a particular researcher



Aggregate all citations for a research object: a publication, the data underlying the findings in the paper, and the software, samples, and reagents used to create those datasets.



Except where otherwise noted, this work is licensed under:
<https://creativecommons.org/licenses/by/4.0/>