



Metadata & Persistent identifiers



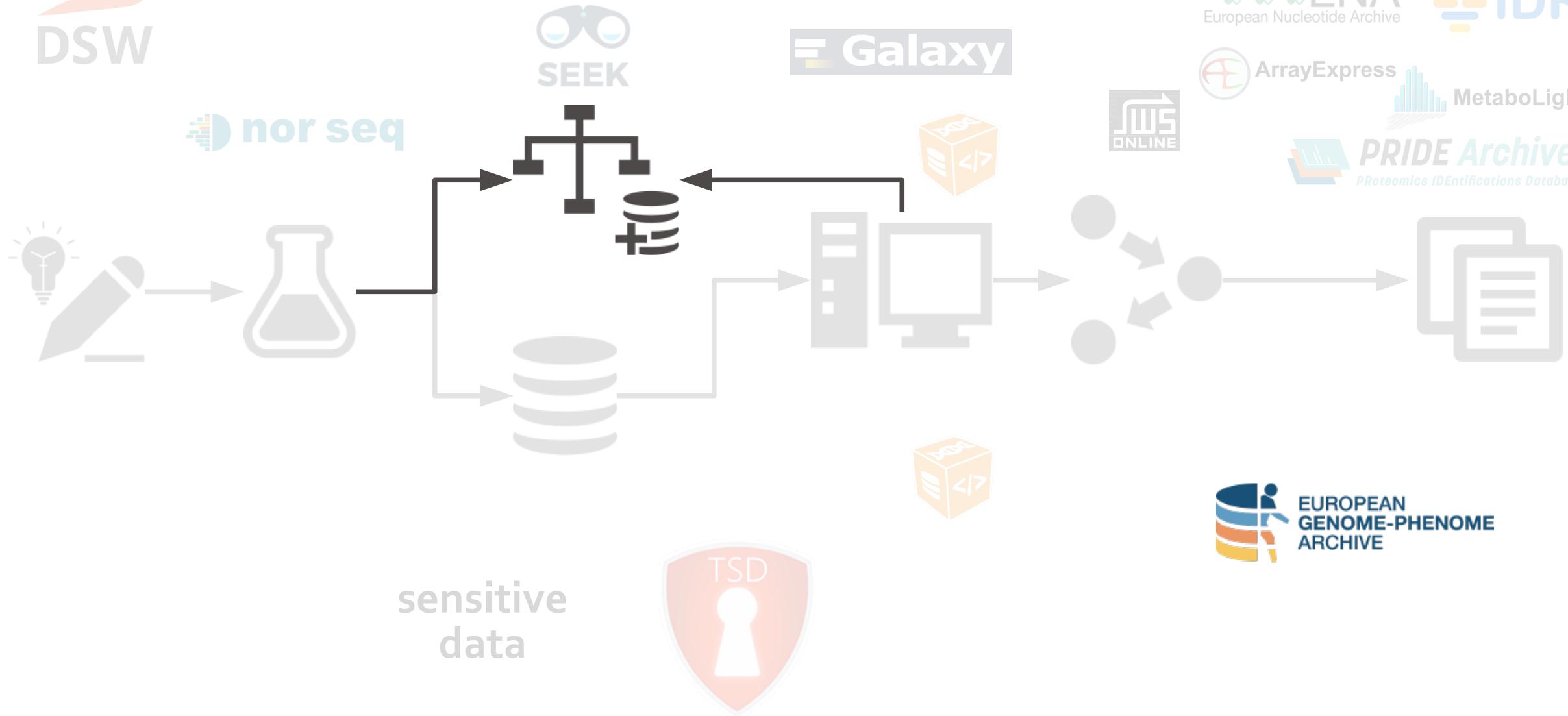
CENTRE FOR
DIGITAL LIFE
NORWAY

Espen Åberg
Data Steward
ELIXIR Norway/BioMedData



NeLS

Norwegian e-Infrastructure for Life Sciences



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	Sex
Compound	Age
Infection	Weight
Inhibitor	Litter
siRNA	Genotype
sgRNA	Species
Dose	Cell line
Time	

Experimental design

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

Environment	Sex	Operator
Compound	Age	Batch
Infection	Weight	Plate
Inhibitor	Litter	Cage
siRNA	Genotype	Array
sgRNA	Species	Flowcell
Dose	Cell line	Instrument
Time		Day
		Order
		Source

Experimental design

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

Environment	Sex	Operator	Experimental
Compound	Age	Batch	Treatment
Infection	Weight	Plate	Sampling
Inhibitor	Litter	Cage	Measurement
siRNA	Genotype	Array	
sgRNA	Species	Flowcell	
Dose	Cell line	Instrument	
Time		Day	
		Order	
		Source	

Experimental design

“Data”

“Metadata”

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

Environment	Sex	Operator	Experimental
Compound	Age	Batch	Treatment
Infection	Weight	Plate	Sampling
Inhibitor	Litter	Cage	Measurement
siRNA	Genotype	Array	
sgRNA	Species	Flowcell	
Dose	Cell line	Instrument	
Time		Day	
		Order	
		Source	

Experimental design

“Data”

Outcome = Treatment effect + Bias

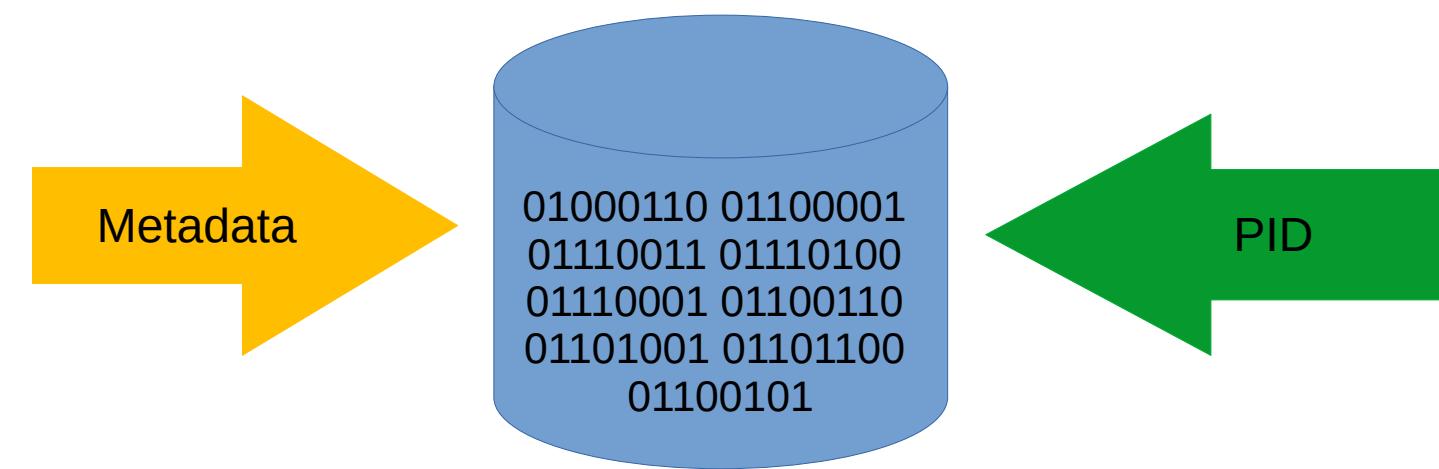
Environment
Components

“Metadata”

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

Environmental
Treatment
Sampling
Measurement



PIDs helps make data FAIR

Data should be Findable	F1. (meta)data are <u>assigned a globally unique and persistent identifier (DOI)</u> F2. data are described with rich metadata F3. metadata <u>clearly and explicitly include the identifier of the data it describes</u> F4. (meta)data are registered or indexed in a searchable resource
Data should be Accessible	A1. (meta)data are <u>retrievable by their identifier using a standardized communications protocol</u> 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

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”



[farm3.staticflickr.com](#)

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.

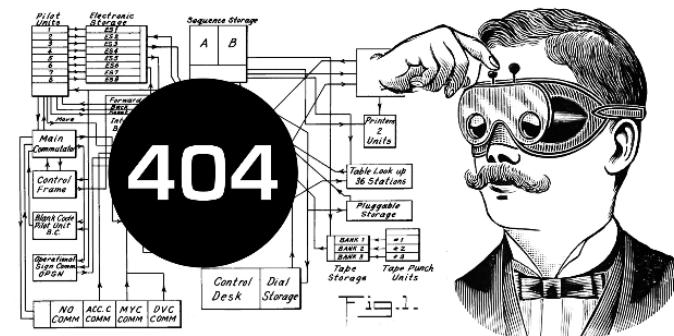
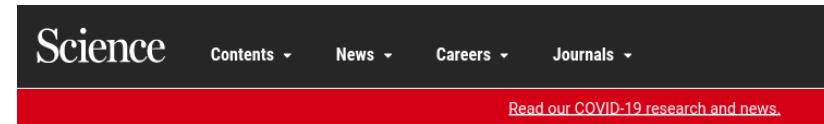
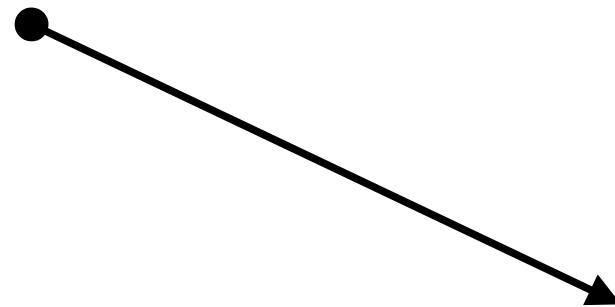
[- Hide authors and affiliations](#)

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.



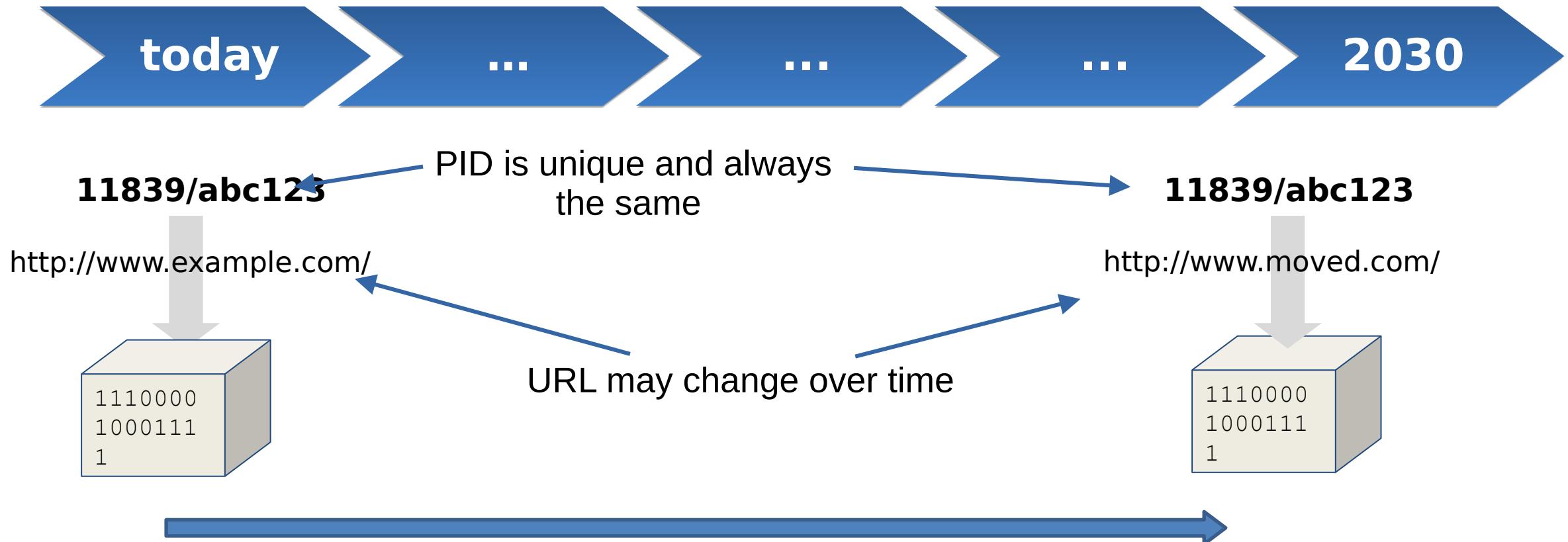
Hmmm...

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

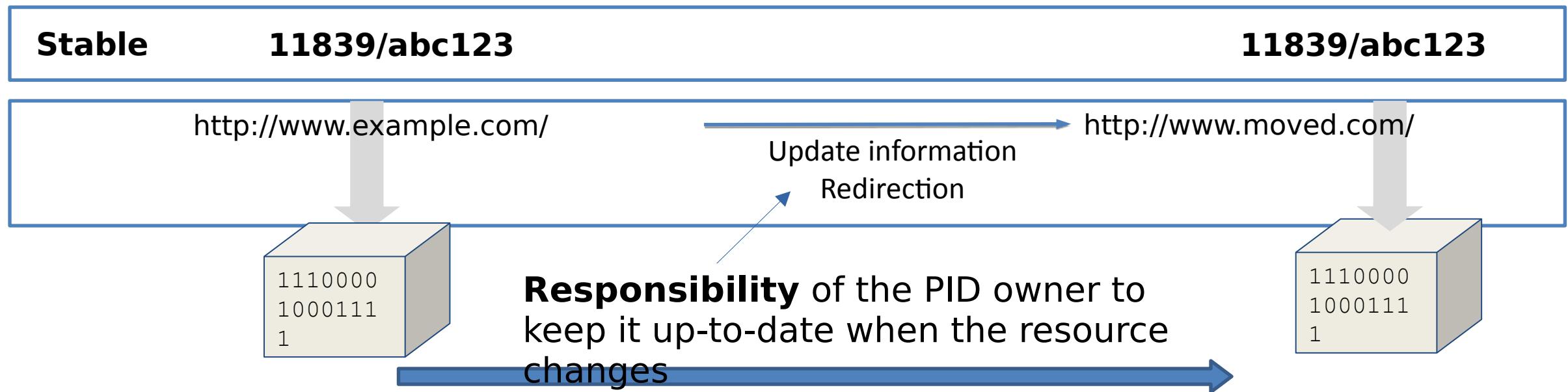
.. by design



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

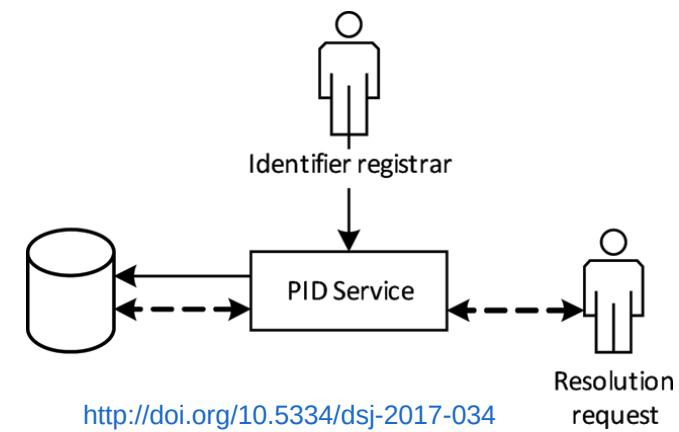
Persistent over time

.. by design



A PID consists of 2 components:

1. a unique identifier
2. a service that locates the resource over time even when its location changes



Examples for digital objects

Digital Object Identifiers



Handles

Handle.Net®

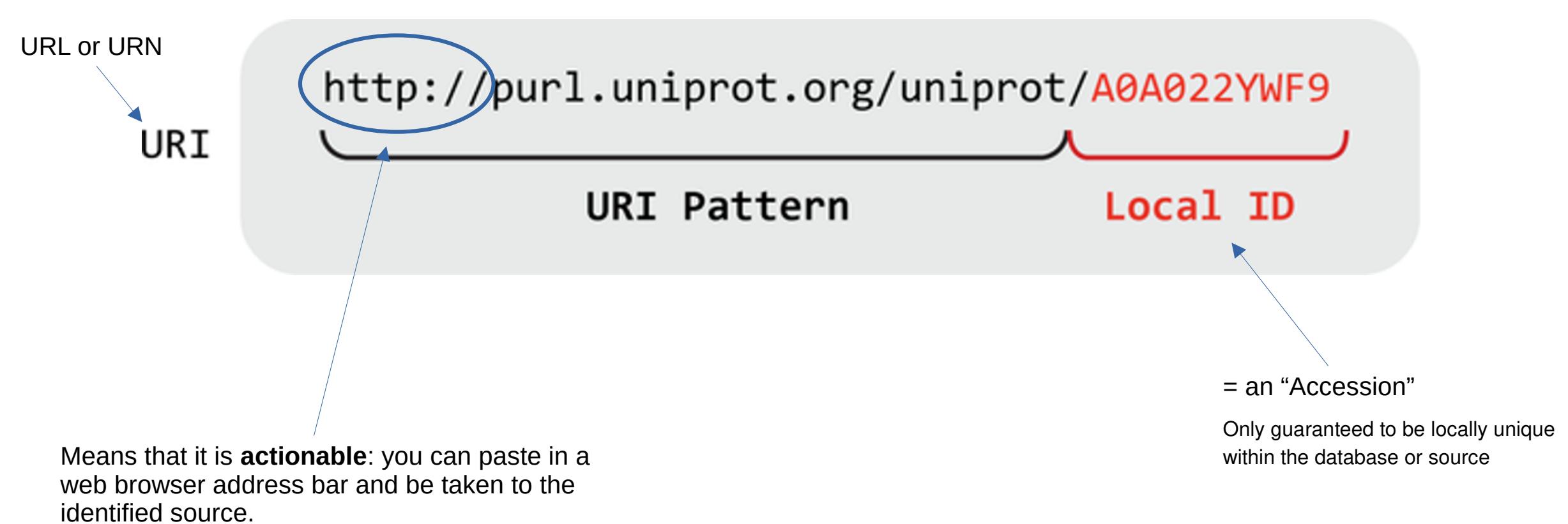
Archival Resource Keys (ARK)

Persistent Uniform Resource Locator
(URL)



Identifiers.org

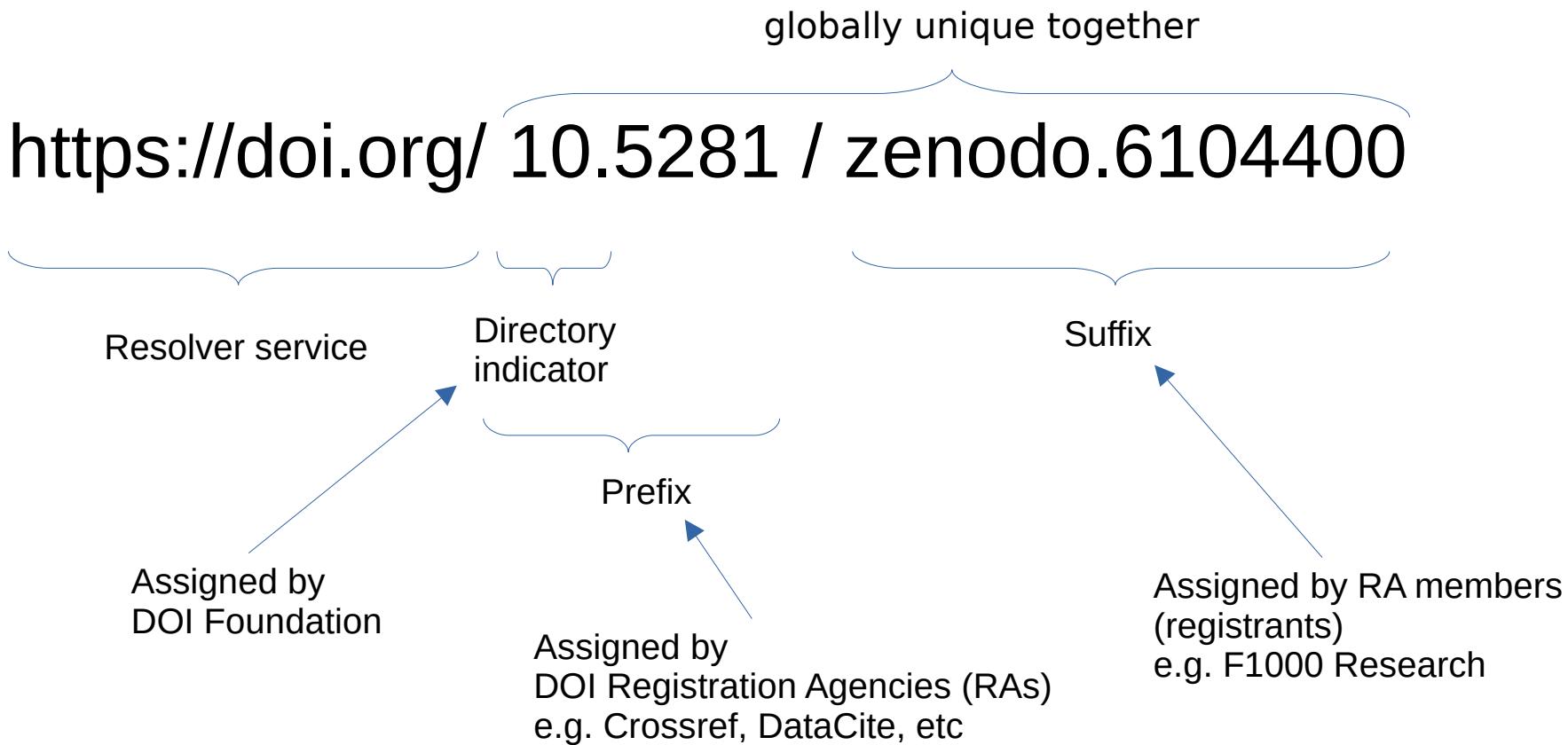
Anatomy of a web-based identifier

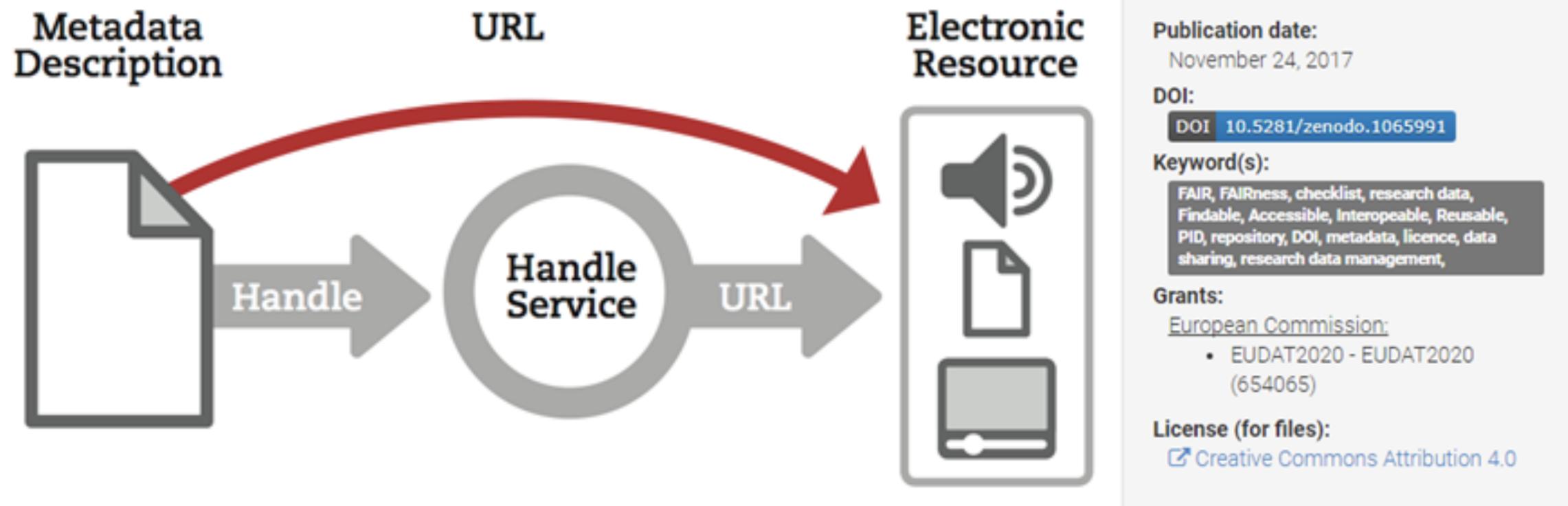


How to recognize a PID

DOI: 10.5281/zenodo.6104400

Anatomy of a DOI





PIDs for

5

People

ORCID

isni

PIDs for

5

People

ORCID

isni

\$

Funding bodies



PIDs for

5

People



\$

Funding bodies



~

Institutions



PIDs for

5

People



\$

Funding bodies



~

Institutions

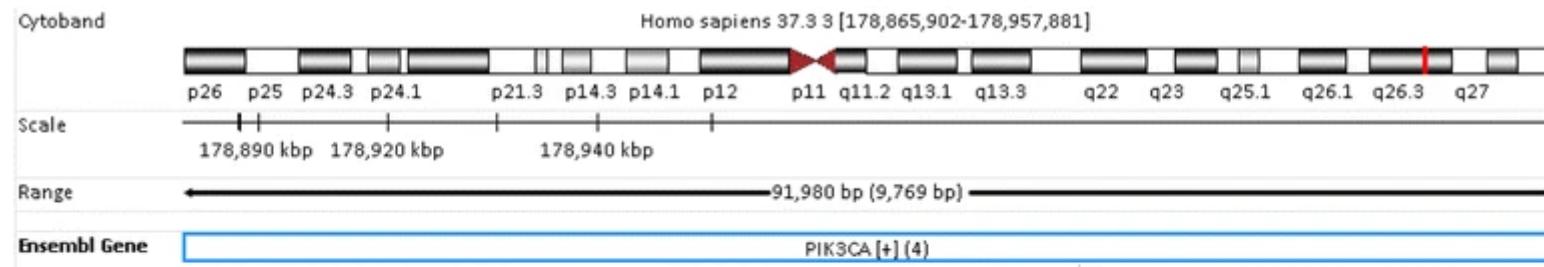


Instruments (soon)



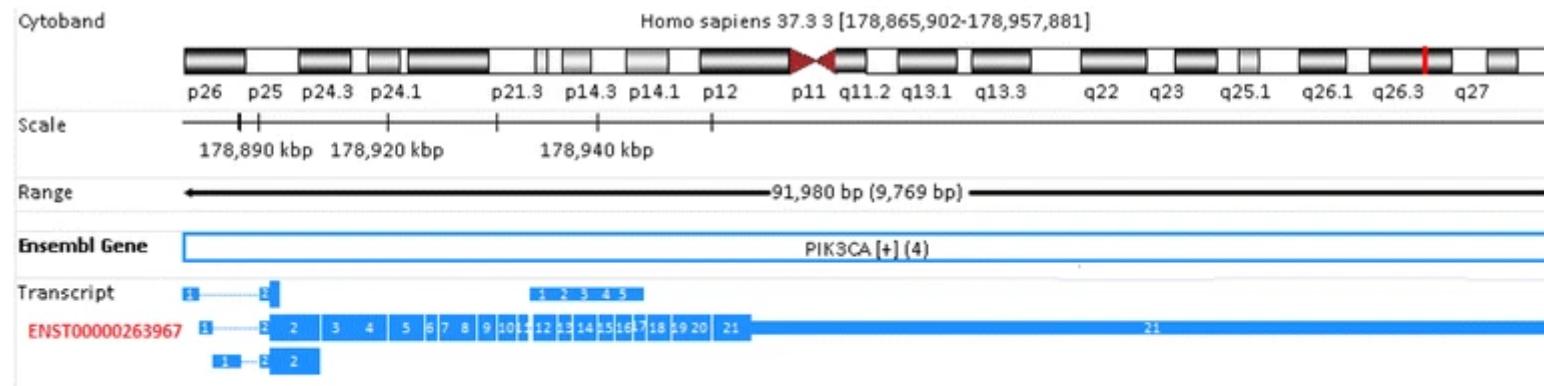
Sequence identifiers:

XXX Gene: PIK3CA



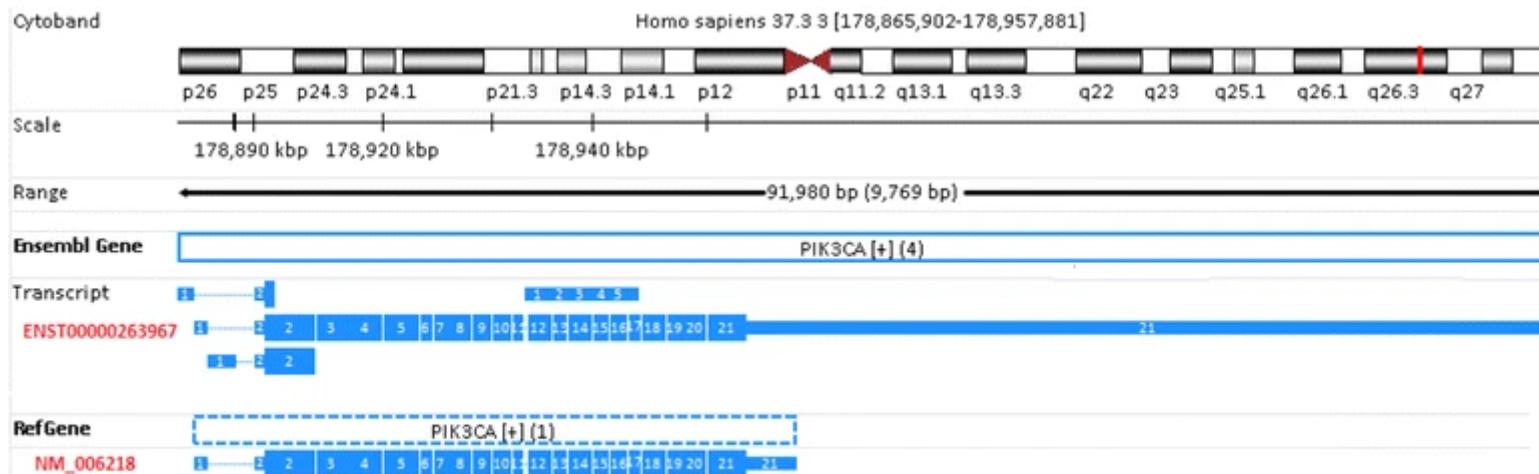
Sequence identifiers:

XXX Gene: PIK3CA



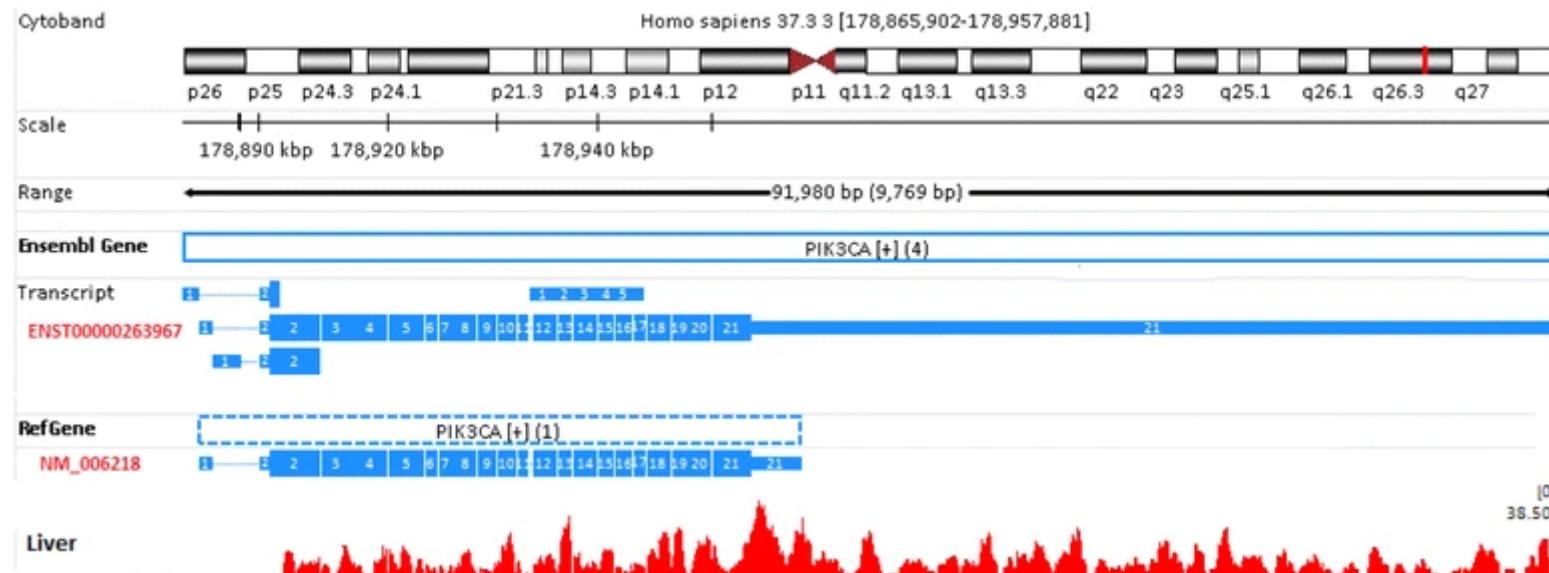
Sequence identifiers:

XXX Gene: PIK3CA



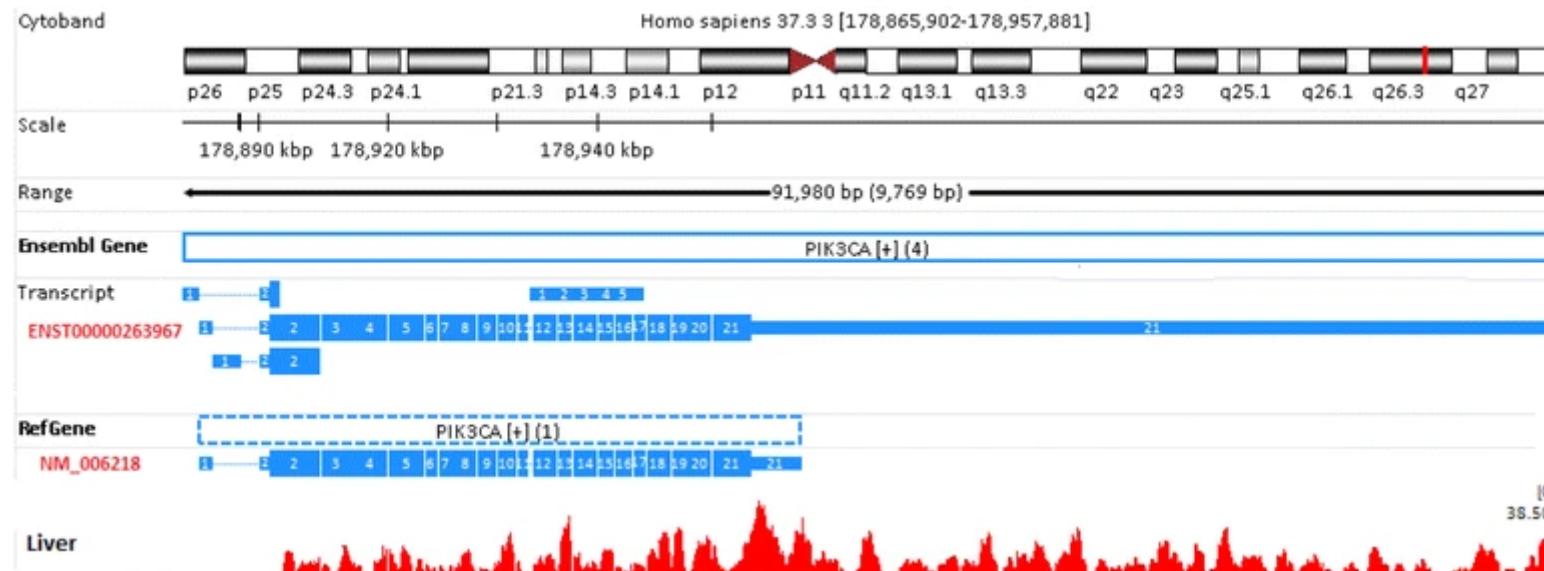
Sequence identifiers:

XXX Gene: PIK3CA



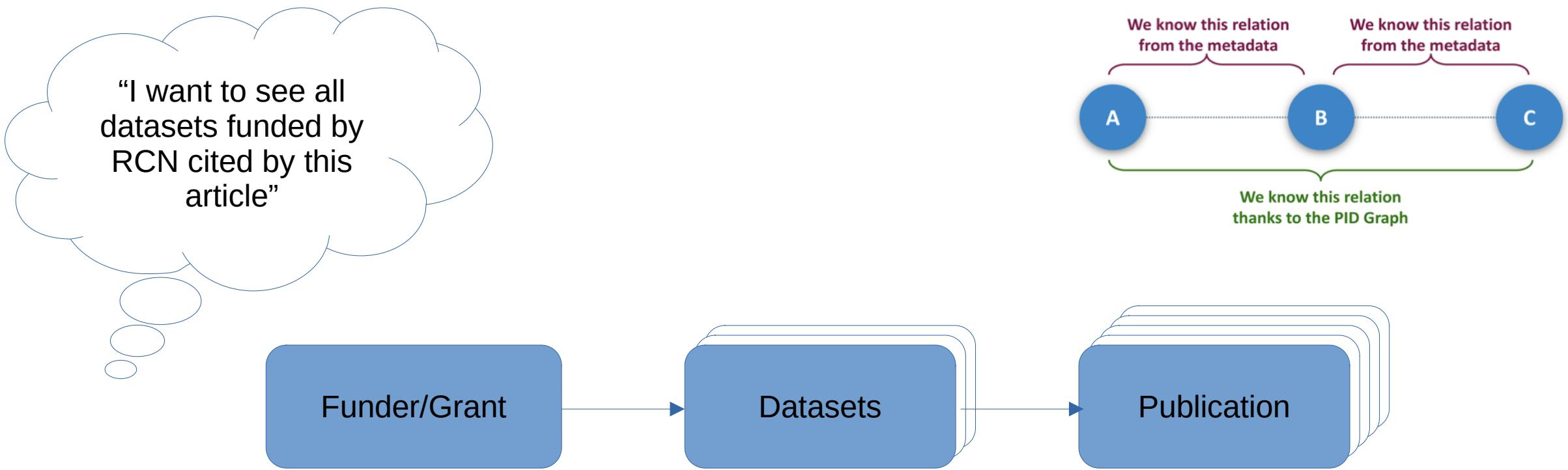
Sequence identifiers:

XXX Gene: PIK3CA



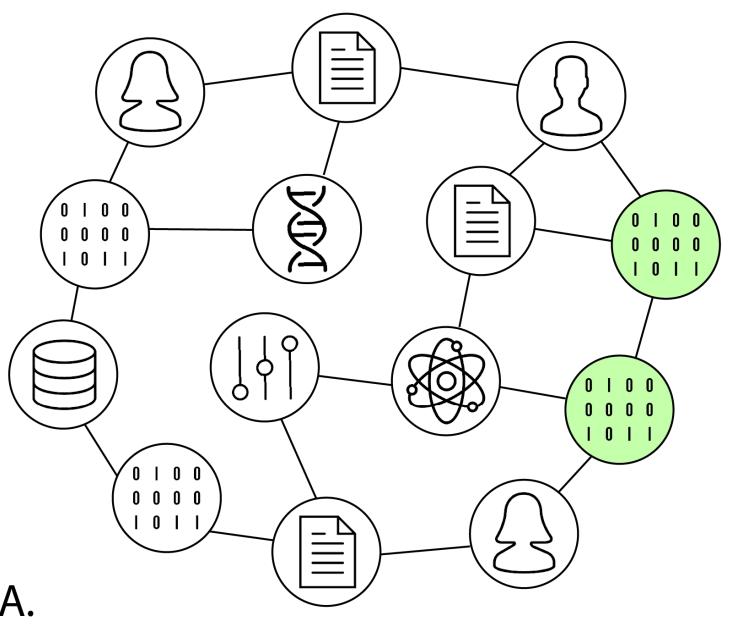
ENST00000263967.2
3
4

PIDs can do more

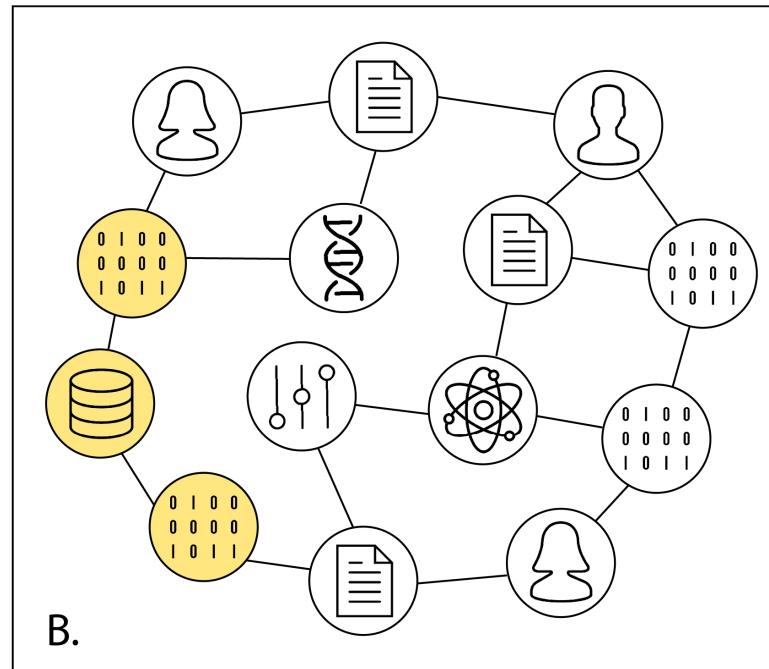




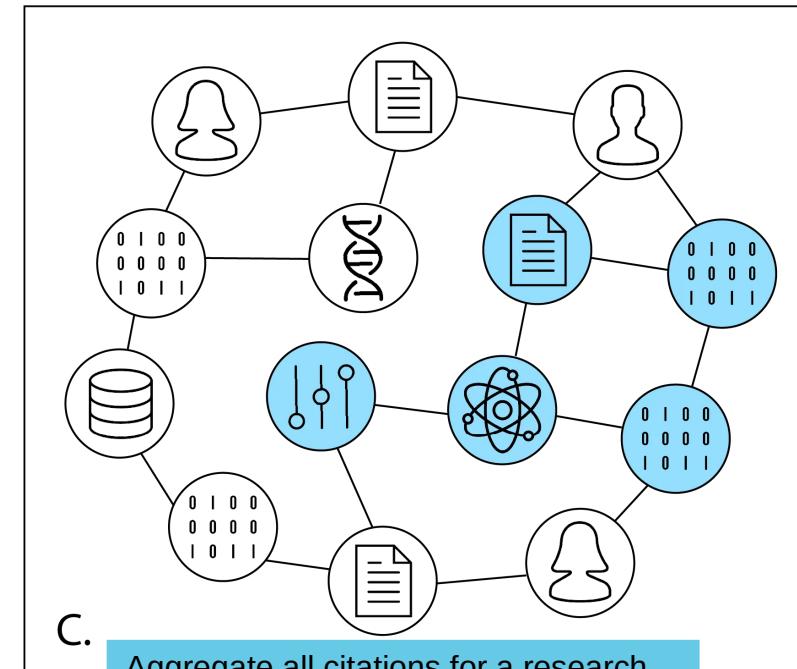
PID Graphs



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/>