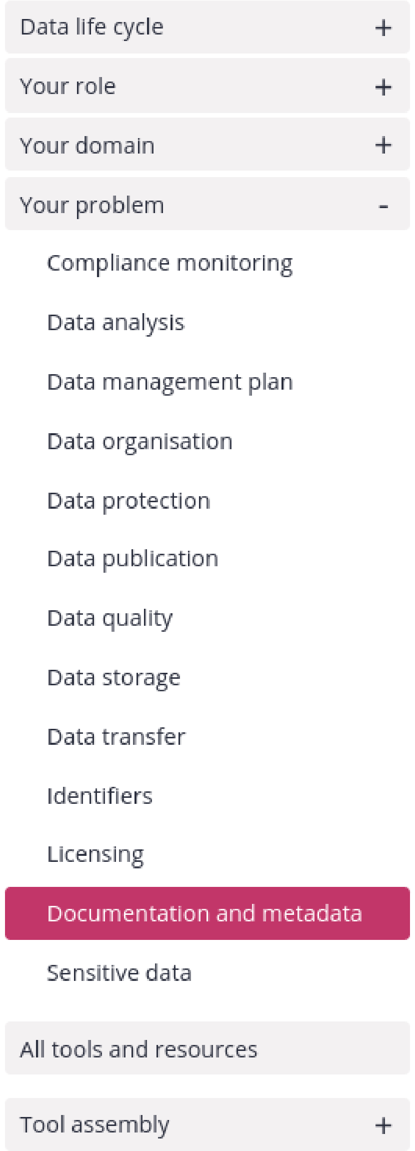
Metadata & Persistent identifiers

Espen Åberg
Data Steward
ELIXIR Norway/BioMedData



Data life cycle +

Your role +

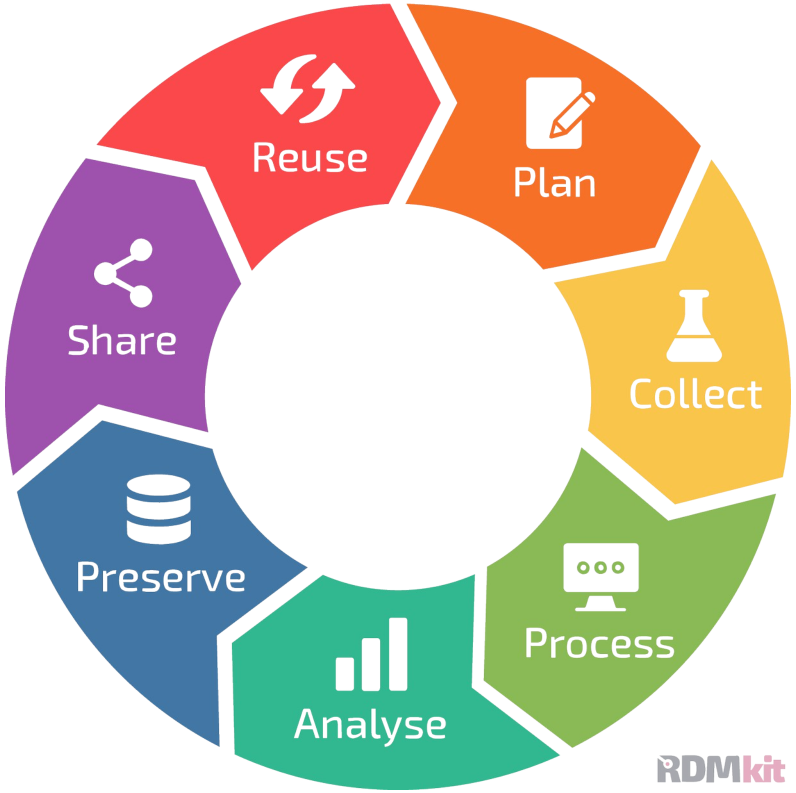
Your domain +

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 +



RDMkit

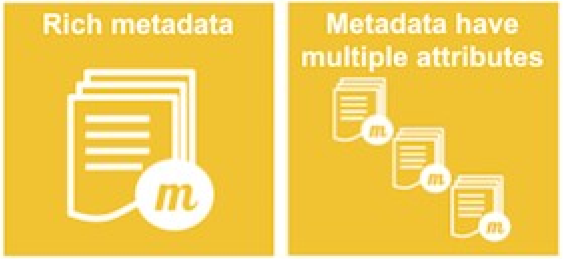
Link to RDMkit: https://rdmkit.elixir-europe.org/

“Metadata is constructed, constructive, and actionable.”

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

“data about data”

What is metadata?



“information about something”

“Data is content, and metadata is context”

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

— Adam Rauh

"Metadata is a Love Note to the Future"

FAIR illustrations icons from ARDC 2018 licensed under Creative Commons Attribution 4.0 International License

“Metadata is constructed, constructive, and actionable.”

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

“data about data”

What i a?

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.

“information about something”

“Data is content, a

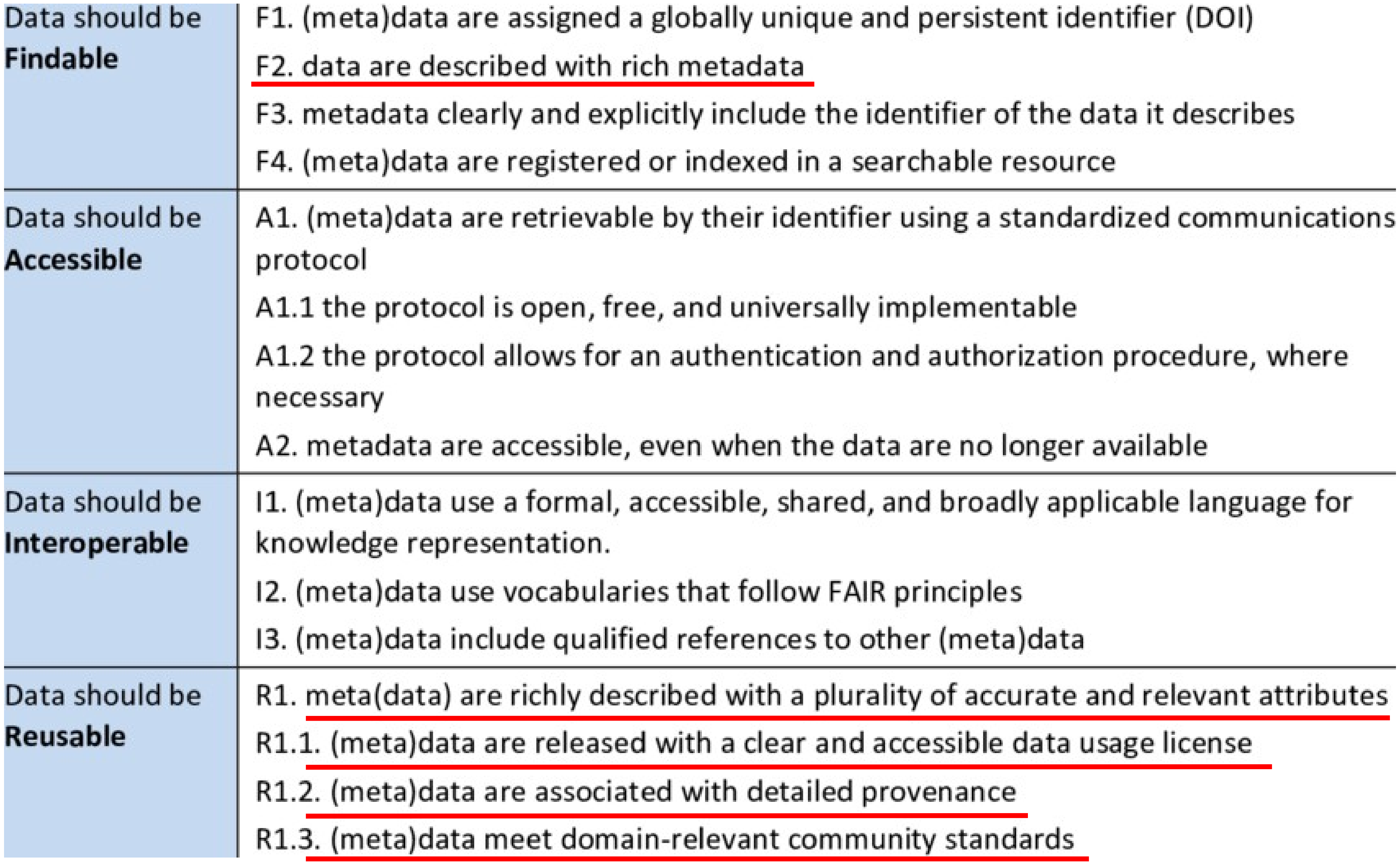
“If data is the new oil, metadata is the refinery"

— Adam Rauh

etadata is a Love Note to the Future"

FAIR illustrations icons from ARDC 2018 licensed under Creative Commons Attribution 4.0 International License

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).

Experimental design

“Data”

“Metadata”

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

Environment
Compound
Infection
Inhibitor
siRNA
sgRNA
Dose
Time

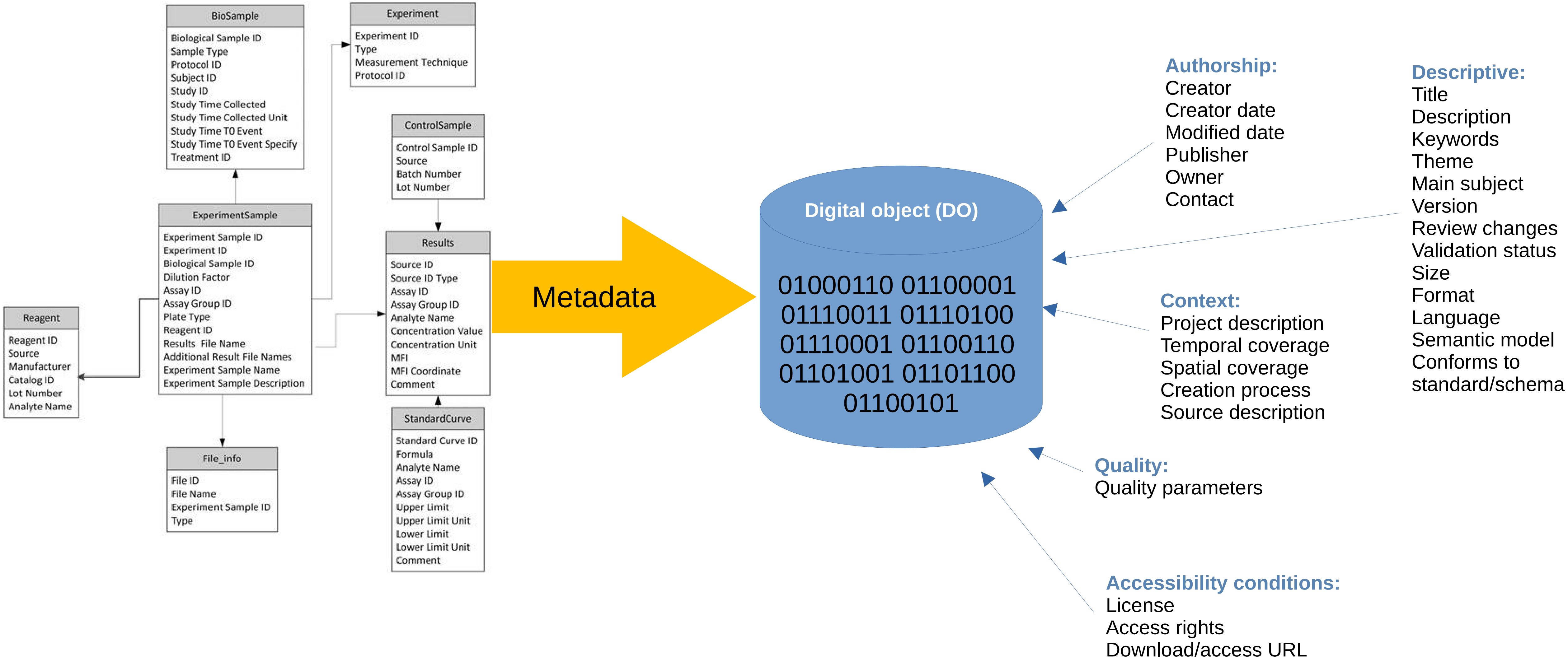
Sex
Age
Weight
Litter
Genotype
Species
Cell line

Operator
Batch
Plate
Cage
Array
Flowcell
Instrument
Day
Order
Source

Experimental
Treatment
Sampling
Measurement

Lazic, S.E. (2016). Cambridge University Press.

“Rich” Metadata



BioSample
- Biological Sample ID
- Sample Type
- Protocol ID
- Subject ID
- Study ID
- Study Time Collected
- Study Time Collected Unit
- Study Time T0 Event
- Study Time T0 Event Specify
- Treatment ID

Experiment
- Experiment ID
- Type
- Measurement Technique
- Protocol ID

ControlSample
- Control Sample ID
- Source
- Batch Number
- Lot Number

ExperimentSample
- Experiment Sample ID
- Experiment ID
- Biological Sample ID
- Dilution Factor
- Assay ID
- Assay Group ID
- Plate Type
- Reagent ID
- Results File Name
- Additional Result File Names
- Experiment Sample Name
- Experiment Sample Description

Results
- Source ID
- Source ID Type
- Assay ID
- Assay Group ID
- Analyte Name
- Concentration Value
- Concentration Unit
- MFI
- MFI Coordinate
- Comment

Reagent
- Reagent ID
- Source
- Manufacturer
- Catalog ID
- Lot Number
- Analyte Name

File_info
- File ID
- File Name
- Experiment Sample ID
- Type

StandardCurve
- Standard Curve ID
- Formula
- Analyte Name
- Assay ID
- Assay Group ID
- Upper Limit
- Upper Limit Unit
- Lower Limit
- Lower Limit Unit
- Comment

Authorship:
Creator
Creator date
Modified date
Publisher
Owner
Contact

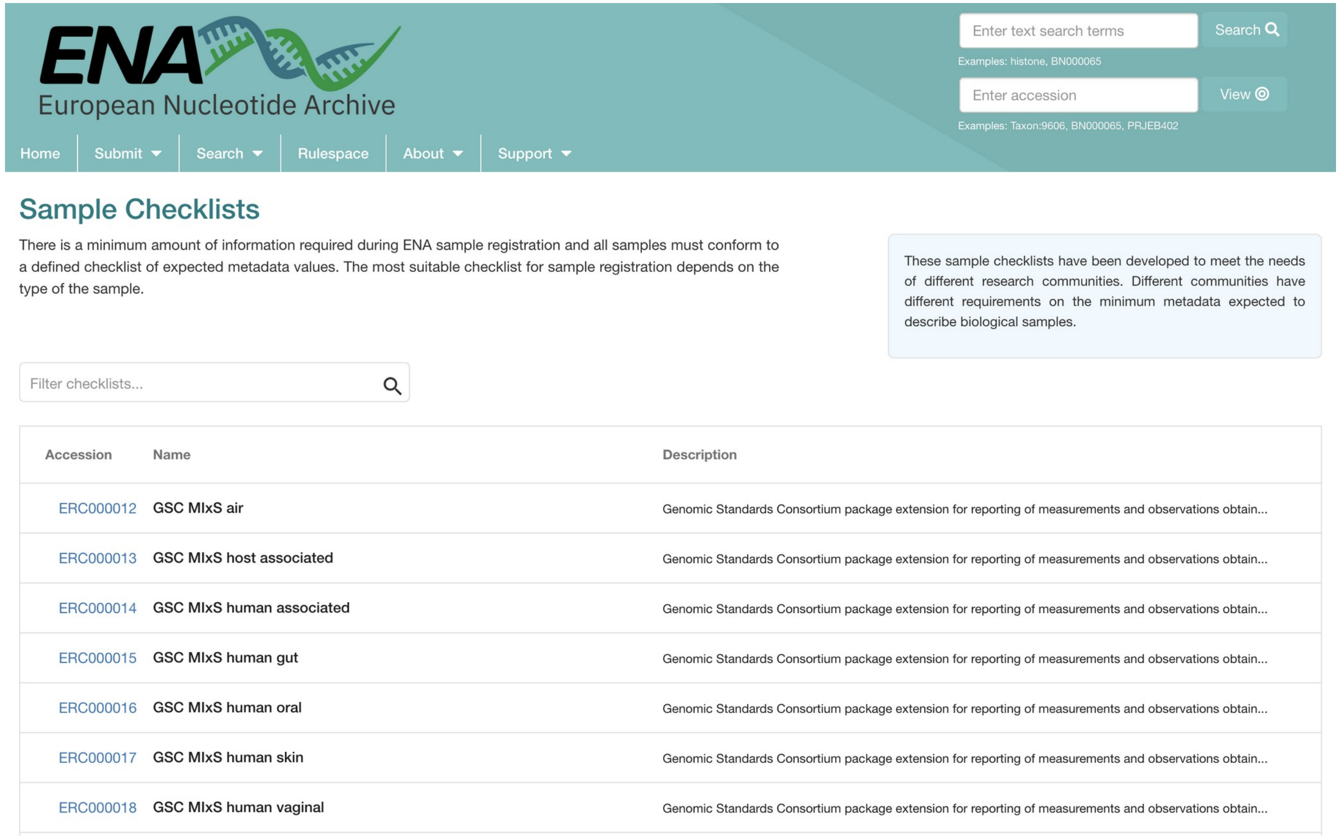
Descriptive:
Title
Description
Keywords
Theme
Main subject
Version
Review changes
Validation status
Size
Format
Language
Semantic model
Conforms to standard/schema

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

Quality:
Quality parameters

Accessibility conditions:
License
Access rights
Download/access URL

Metadata templates/checklists



ENA
European Nucleotide Archive
Enter text search terms
Search
Examples: histone, BN000065
Enter accession
View
Examples: Taxon:9606, BN000065, PRJEB402
Home
Submit
Search
Rulespace
About
Support

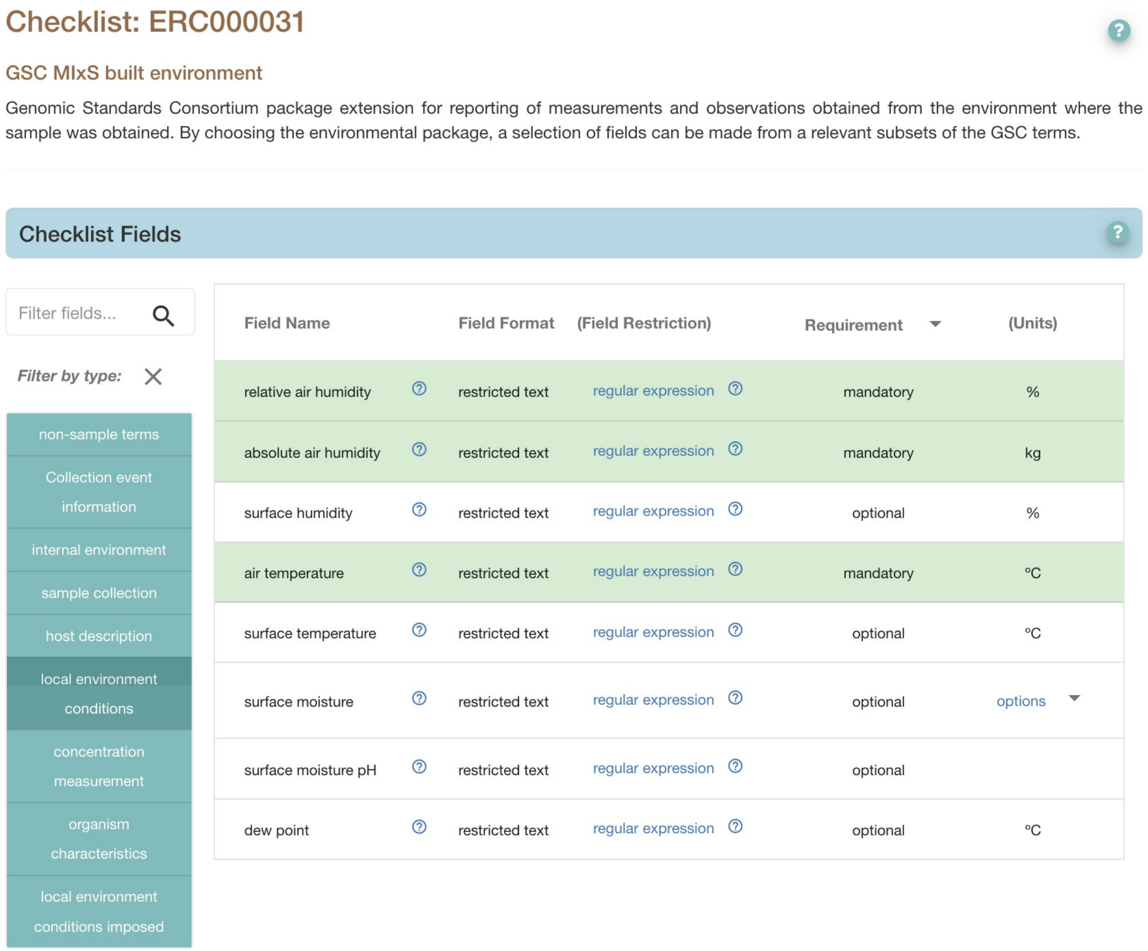
Sample Checklists

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

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	ºC
surface temperature	restricted text	regular expression	optional	ºC
surface moisture	restricted text	regular expression	optional	options
surface moisture pH	restricted text	regular expression	optional	
dew point	restricted text	regular expression	optional	ºC

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

Metadata Submission Workflow

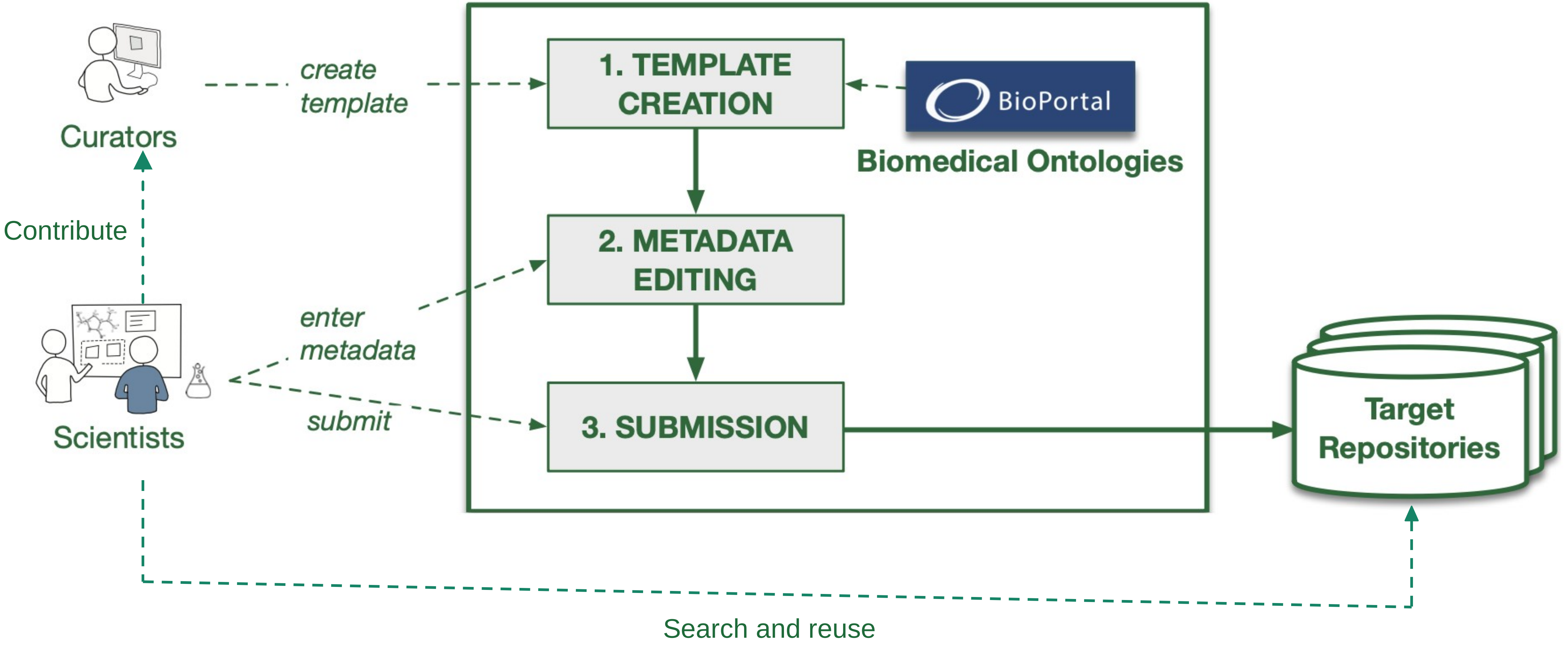
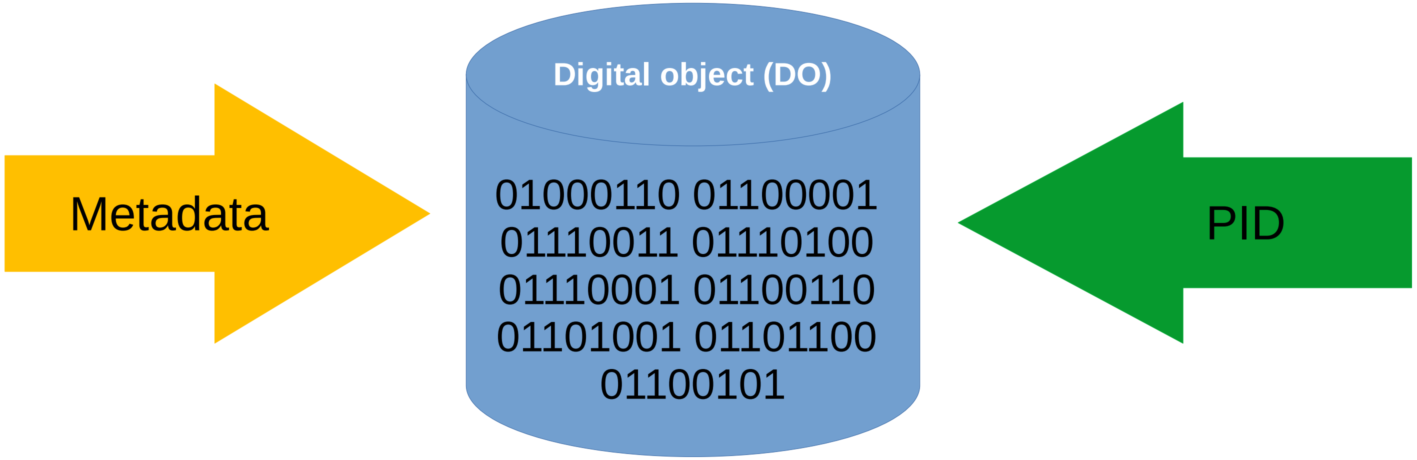


Figure 2 from: Using Semantic Technologies to Enhance Metadata Submissions to Public Repositories in Biomedicine



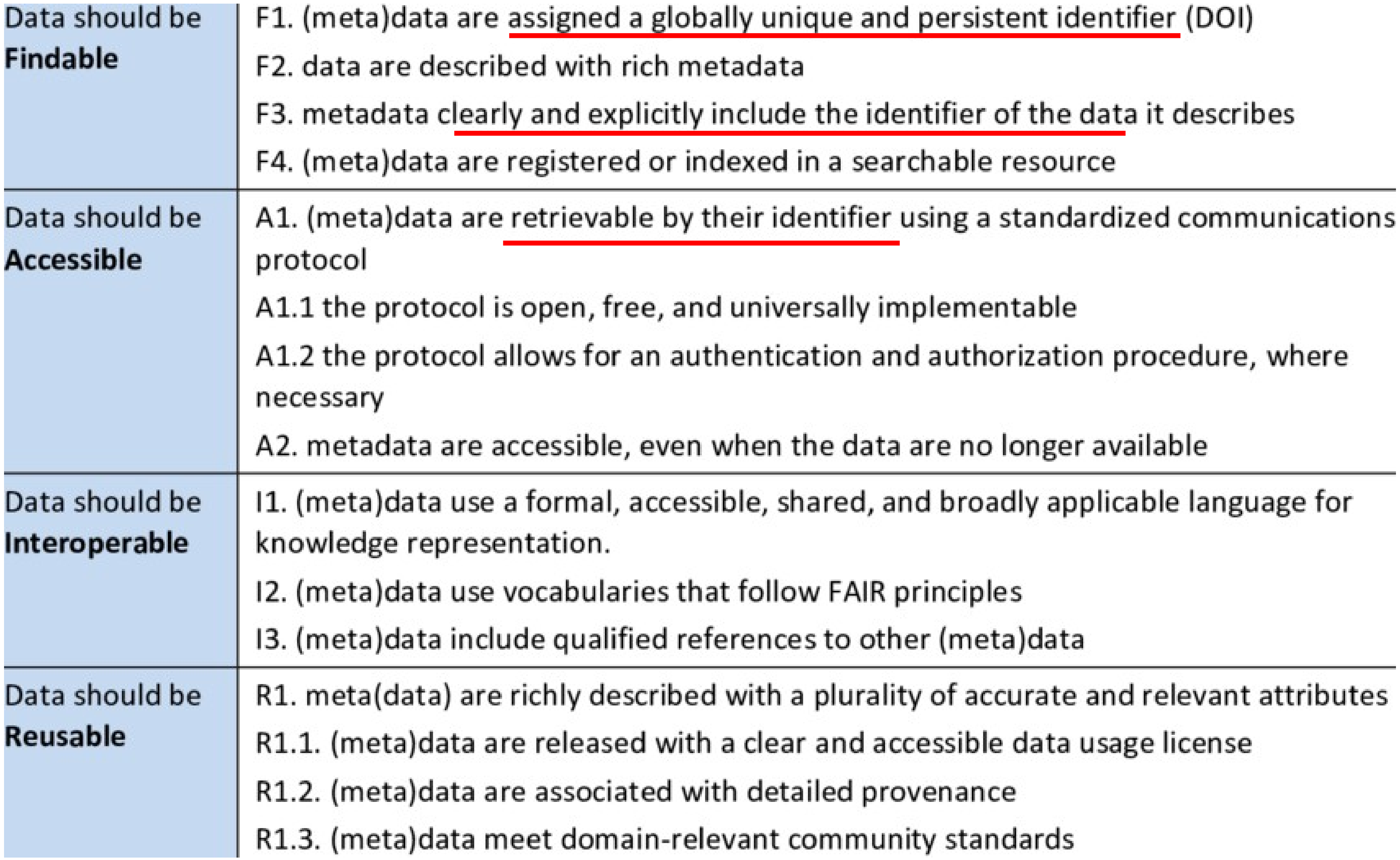
Metadata

Digital object (DO)

01000110 01100001
01110011 01110100
01110001 01100110
01101001 01101100
01100101

PID

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

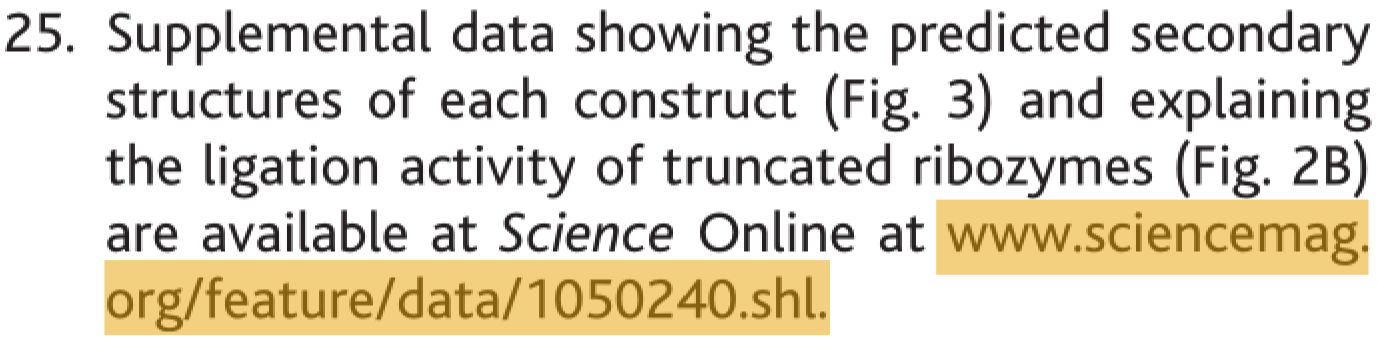
Medyckyj-Scott, David et al. (2016).

Why not just use a URL?

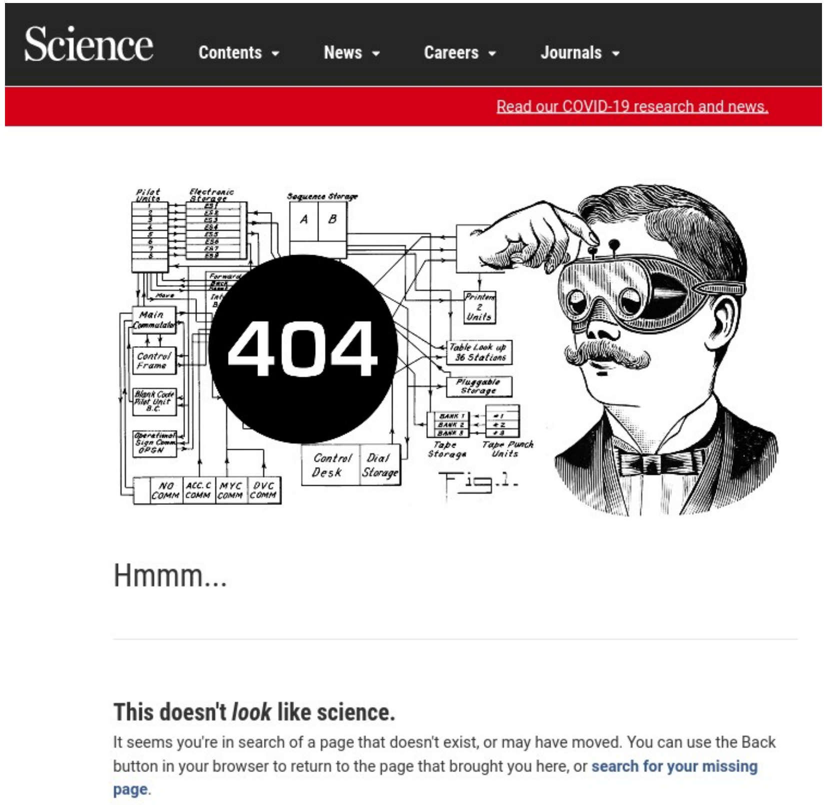
domain may change

resource may be relocated

URL may change



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...

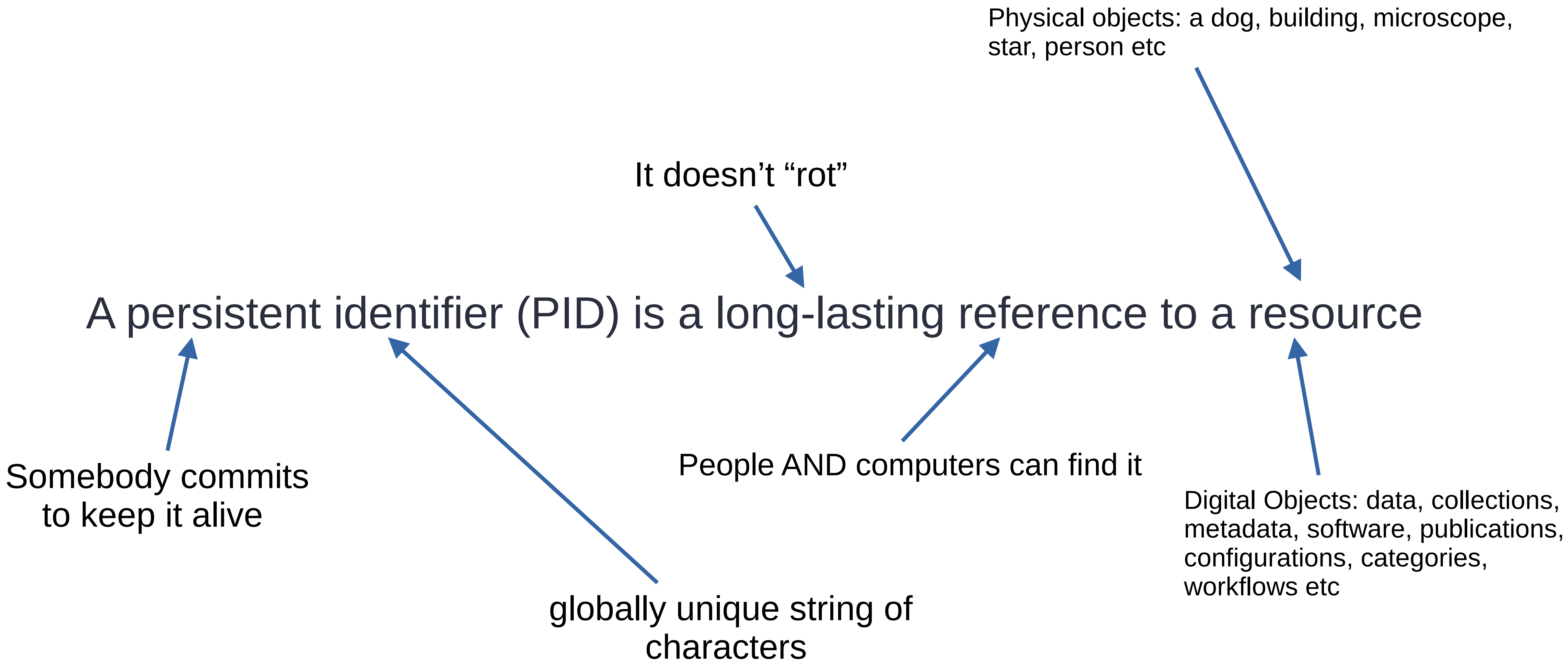
This doesn't *look* like science.

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**.

farm3.staticflickr.com

“Link rot”

PID?



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

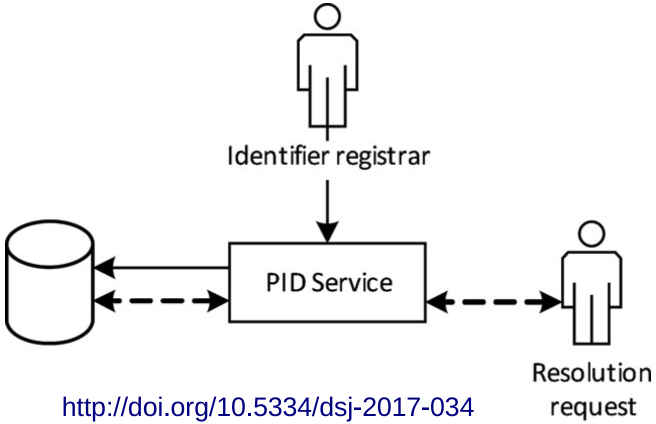
People AND computers can find it

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

A PID consists of two components:

1. a unique identifier

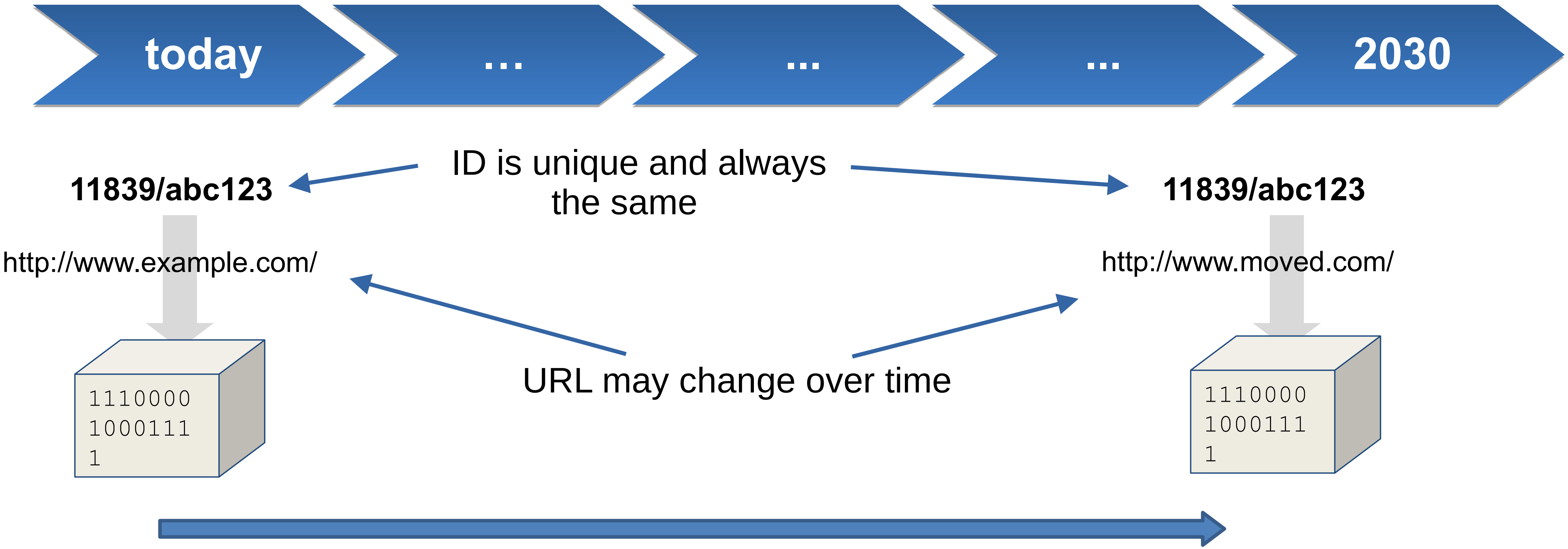
2. a service that locates the resource over time even when it's location changes



http://doi.org/10.5334/dsj-2017-034

Persistent over time

.. by design



today

...

...

...

2030

11839/abc123

ID is unique and always the same

11839/abc123

http://www.example.com/

http://www.moved.com/

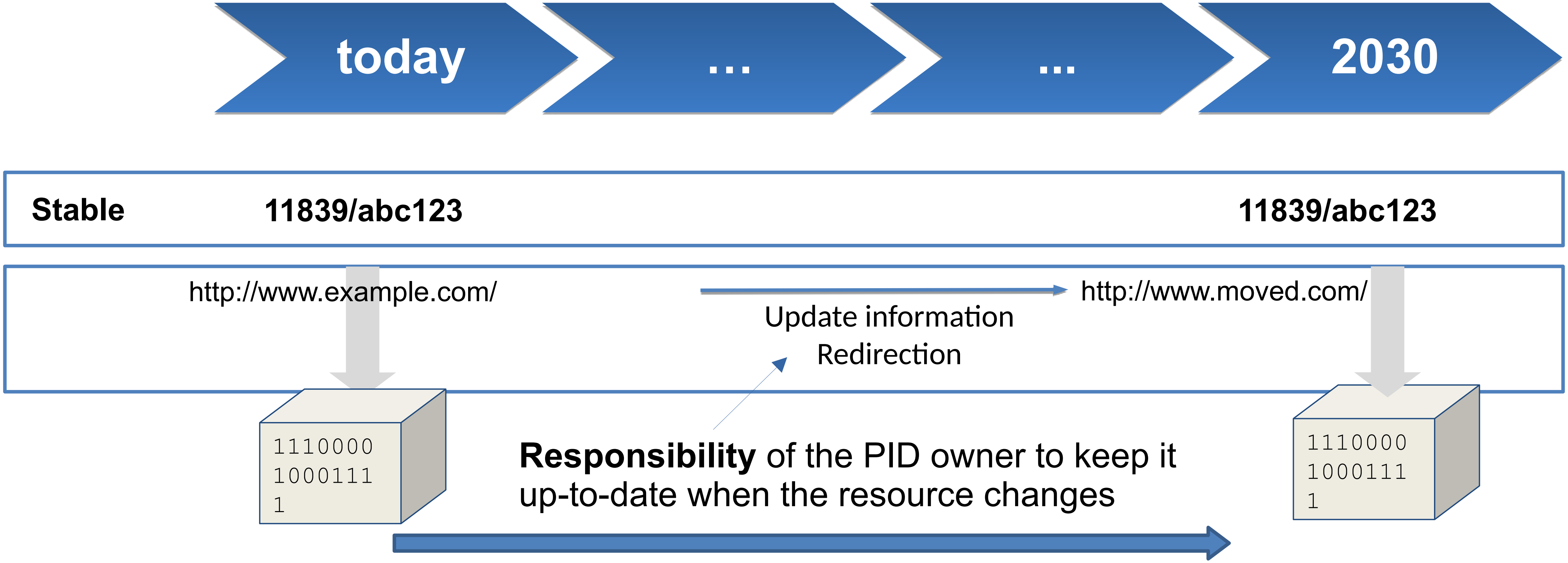
URL may change over time

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

Modified from https://b2drop.eudat.eu/s/TCYdyf3eqwE0Ust

Persistent over time

.. by design



today

...

...

2030

Stable

11839/abc123

11839/abc123

http://www.example.com/

http://www.moved.com/

Update information

Redirection

1110000
1000111
1

1110000
1000111
1

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

Modified from https://b2drop.eudat.eu/s/TCYdyf3eqwE0Ust

Examples for digital objects

Digital Object Identifiers doi

Handles Handle.Net®

Archival Resource Keys (ARK)

Persistent Uniform Resource Locator (URL)

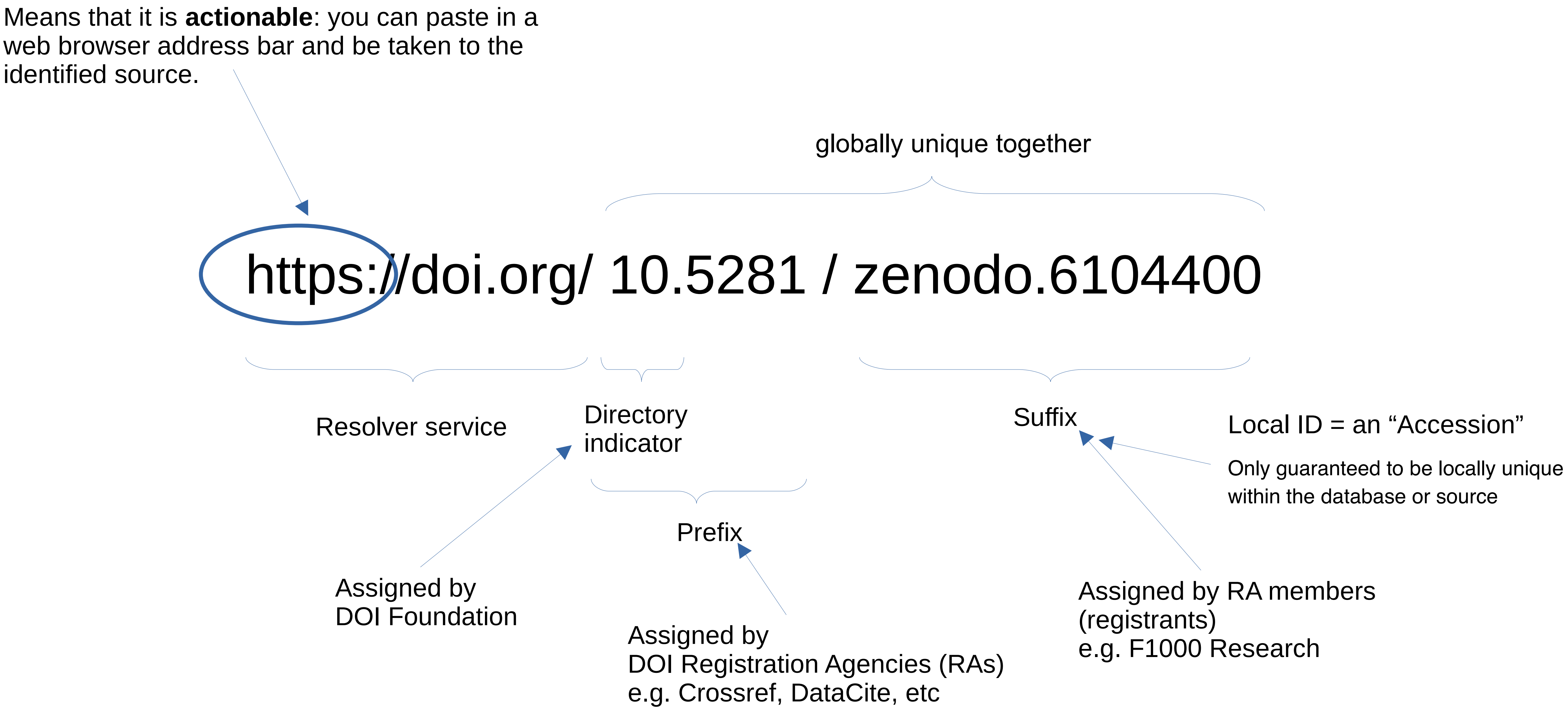
Identifiers.org

N2T.net

How to 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.

globally unique together

https://doi.org/ 10.5281 / zenodo.6104400

Resolver service

Directory indicator

Prefix

Suffix

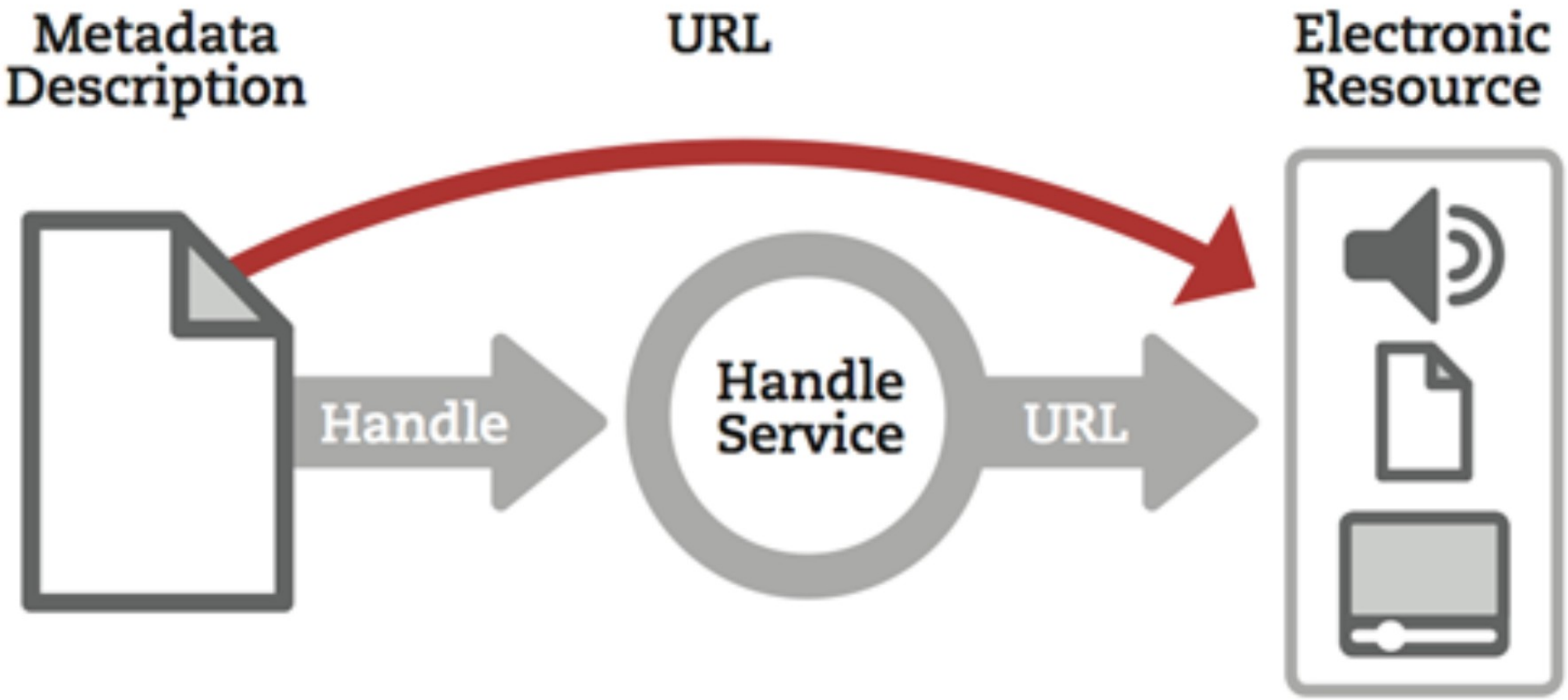
Local ID = an “Accession”

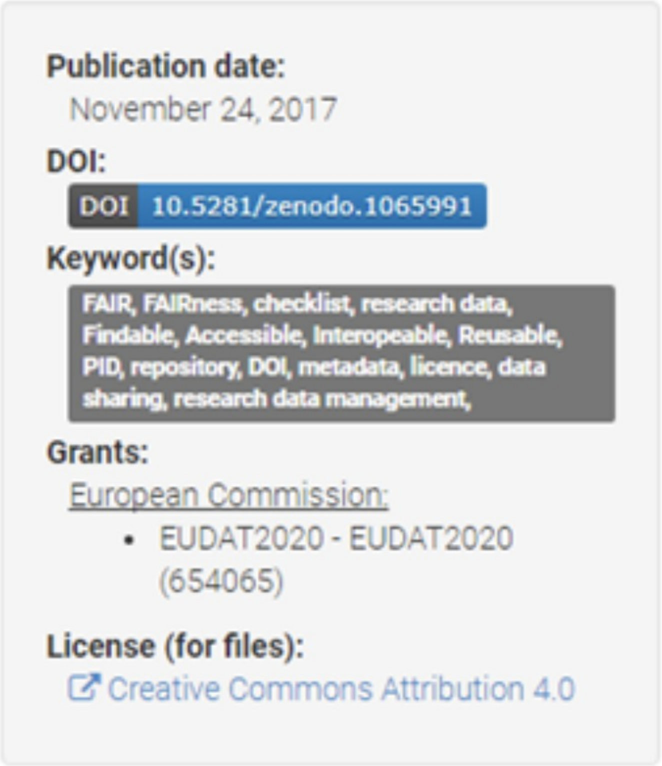
Only guaranteed to be locally unique within the database or source

Assigned by
DOI Foundation

Assigned by
DOI Registration Agencies (RAs)
e.g. Crossref, DataCite, etc

Assigned by RA members (registrants)
e.g. F1000 Research





Publication date:

November 24, 2017

DOI:

DOI 10.5281/zenodo.1065991

Keyword(s):

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

Grants:

European Commission:

- EUDAT2020 - EUDAT2020 (654065)

License (for files):

Creative Commons Attribution 4.0

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

PIDs for...

People ORCID isni

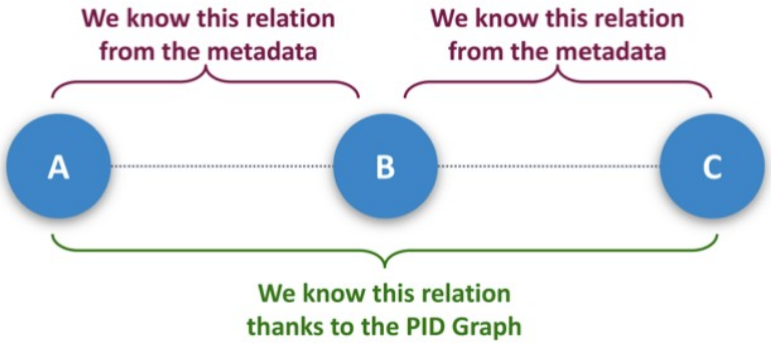
$ Funding bodies Crossref

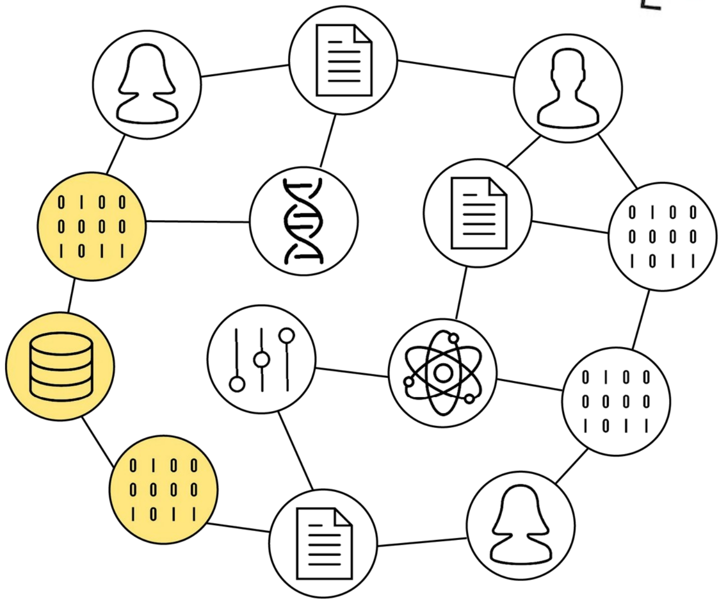
Institutions GRID ROR

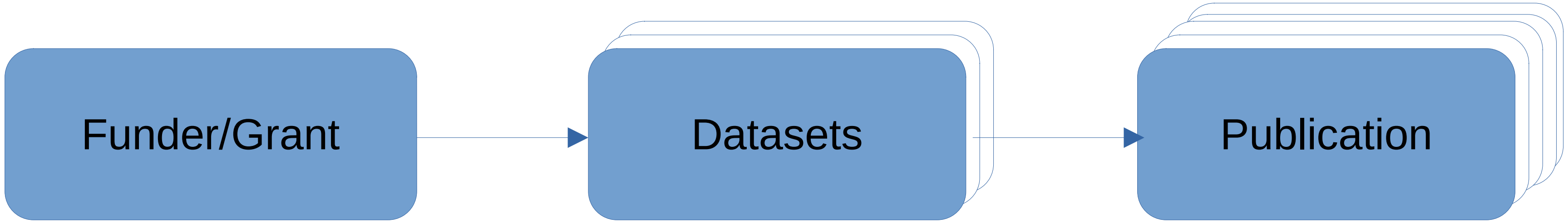
Instruments (soon) PiD

PIDs assembled into graphs

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


We know this relation from the metadata
We know this relation from the metadata
A
B
C
We know this relation thanks to the PID Graph





Funder/Grant

Datasets

Publication

Crossref

DataCite
FIND, ACCESS, AND REUSE DATA

Crossref