DTL

DUTCH TECHCENTRE FOR LIFE SCIENCES

PERSISTENT IDENTIFIERS AND THEIR USE CASES

IMPROVING FINDABILITY OF DATA

CHRISTINE STAIGER
COORDINATOR DATA STEWARDSHIP

MAY, 2019
Helis Academy

Agenda

9:00 –9:15am	Arrival and Coffee
9:15 – 10:45am	Persistent Identifiers
10:45 – 11:15am	Coffee
11:15 - 11:45am	Data sharing, publishing and archiving
11:45 - 12:00pm	Intro exercise Dataverse
12:00 - 12:30pm	Exercise Dataverse
12:30 – 1:30pm	Lunch
1:30 – 3:00pm	Data sharing, publishing and archiving

PIDs – Why?

- Managing increasing numbers of **data objects**
- **Sharing data from different sources** amongst researchers

- Data needs to be **(globally) identifiable and addressable** to ensure reuse of data
- Data citation
- Linking data from different sources

- Challenges
  - Object locations change over time
  - Object migration between repositories

What do we want from data?

- Findable – Easy to find by both humans and computer systems
  → Expose Metadata

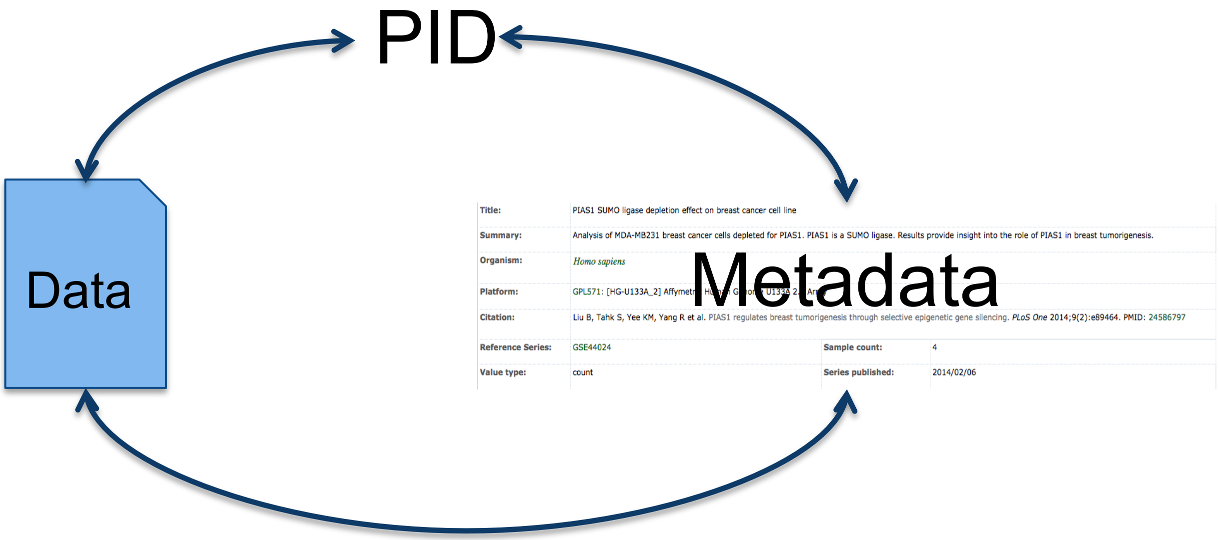
- Accessible – Stored for long term, accessed and/or downloaded with well-defined license and access

- Interoperable – Ready to be combined with other datasets by humans as well as computer systems;

- Reusable – Ready to be used for future research and to be processed further using computational methods.

→ Reference data and identify data
→ Infrastructure should take care of some aspects

Digital Object (DO)

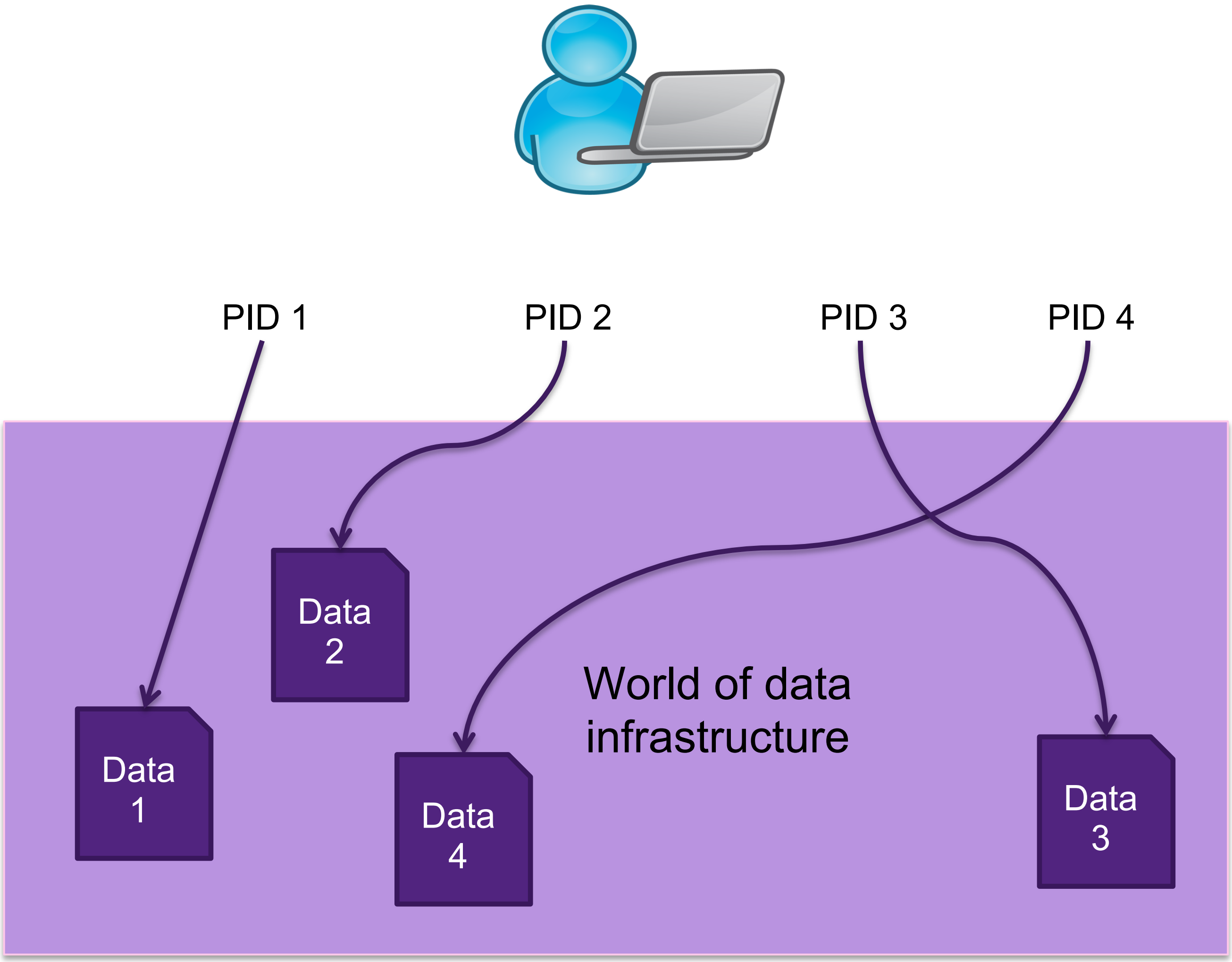


Title:	PIAS1 SUMO ligase depletion effect on breast cancer cell line		
Summary:	Analysis of MDA-MB231 breast cancer cells depleted for PIAS1. PIAS1 is a SUMO ligase. Results provide insight into the role of PIAS1 in breast tumorigenesis.		
Organism:	*Homo sapiens*		
Platform:	GPL571: [HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array		
Citation:	Liu B, Tahk S, Yee KM, Yang R et al. PIAS1 regulates breast tumorigenesis through selective epigenetic gene silencing. *PLoS One* 2014;9(2):e89464. PMID: 24586797		
Reference Series:	GSE44024	**Sample count:**	4
Value type:	count	**Series published:**	2014/02/06

- Persistent Identifier: reference and identify object, either metadata or data object

- Synchronise PID, Data and Metadata during creation, maintenance, update and deletion of a digital object!

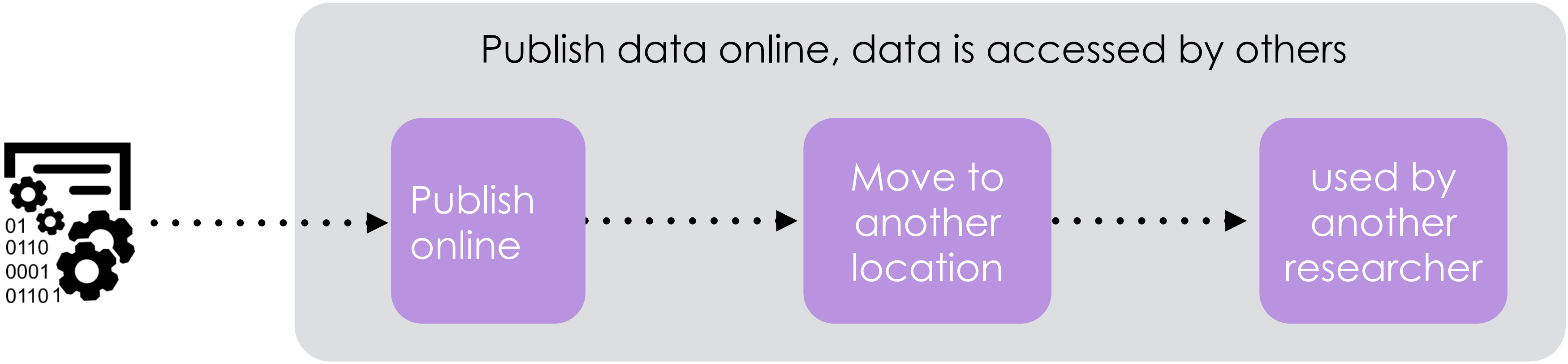
Pro: PIDs are static



Data:

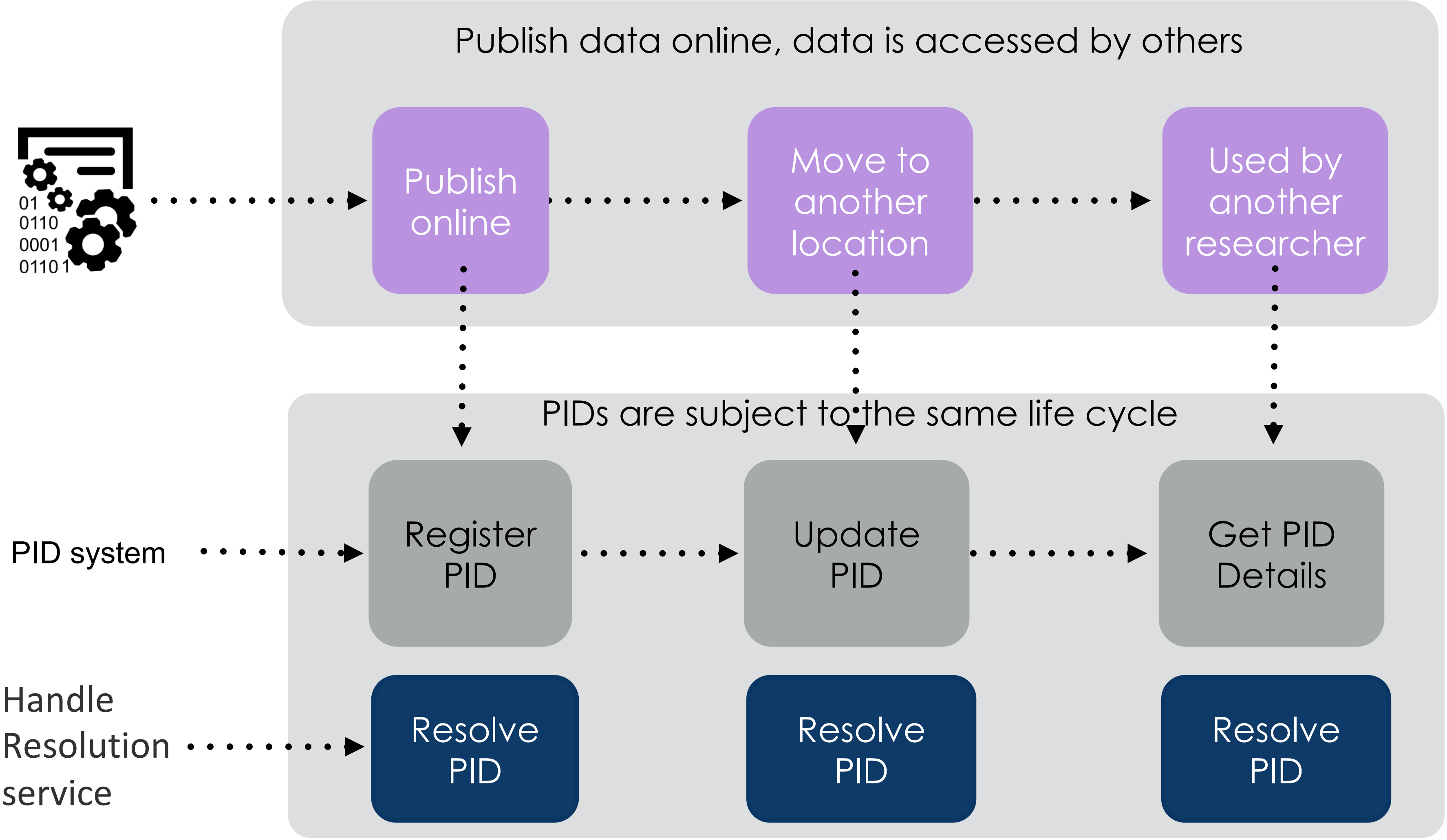
- Files
- Folders
- Webpages
- Sometimes even real world objects

Simple example of data sharing


01
0110
0001
0110 1
Publish data online, data is accessed by others
Publish online
Move to another location
used by another researcher

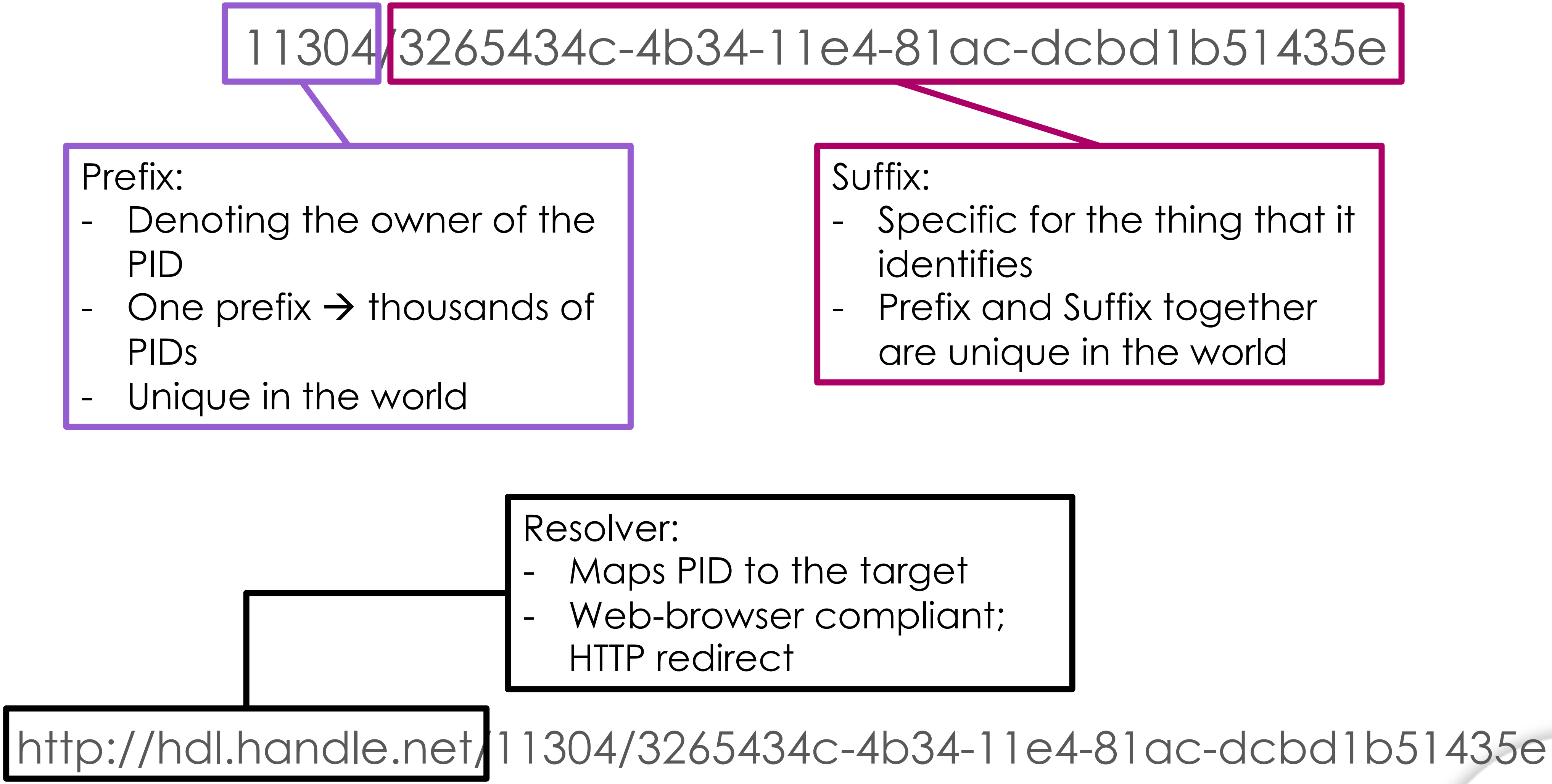
- Published online: http://www.test.com/test.html
- Other users may cite, access, re-use this url
- Relocate the resource at http://www.example.com/
- Other users are not informed -> 404

Simple example of data sharing


Publish data online, data is accessed by others
01
0110
0001
0110 1
Publish online
Move to another location
Used by another researcher
PIDs are subject to the same life cycle
PID system
Register PID
Update PID
Get PID Details
Handle Resolution service
Resolve PID
Resolve PID
Resolve PID

C Staiger, S Bendoukha, EUDAT Summer School 2017, Data discoverability and PIDs

Structure of a PID



11304/3265434c-4b34-11e4-81ac-dcbd1b51435e

Prefix:

- Denoting the owner of the PID
- One prefix → thousands of PIDs
- Unique in the world

Suffix:

- Specific for the thing that it identifies
- Prefix and Suffix together are unique in the world

Resolver:

- Maps PID to the target
- Web-browser compliant; HTTP redirect

http://hdl.handle.net/11304/3265434c-4b34-11e4-81ac-dcbd1b51435e

PID Use Cases

Use Case 1: Data publication

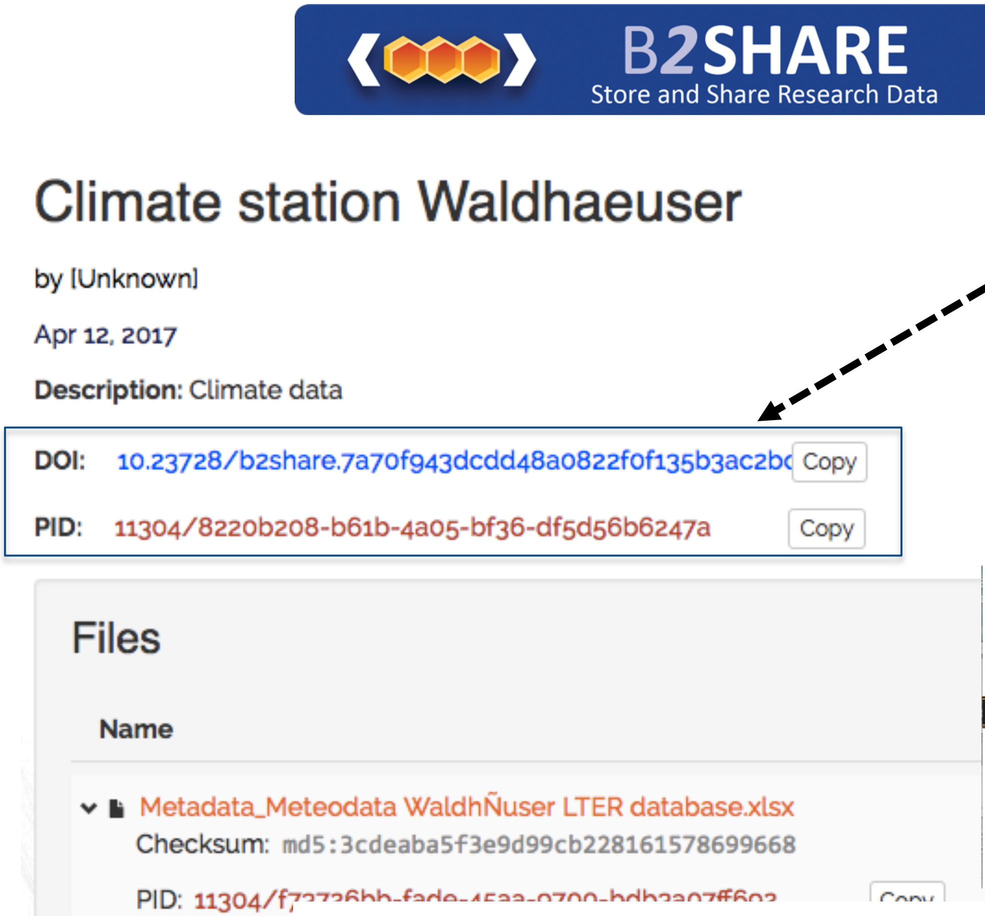
- PIDs point to landing page of the digital repository showing metadata
- “Real” data can be downloaded from this page with another link
- E.g. B2SHARE, FigShare, Zenodo, …
- PID

http://hdl.handle.net/11304/3265434c-4b34-11e4-81ac-dcbd1b51435e

- resolves to landing page

https://b2share.eudat.eu/records/feafb12e810c489b9e878949c6c35345

The persistent identifier for the **collection**



Climate station Waldhaeuser

by [Unknown]

Apr 12, 2017

Description: Climate data

DOI: 10.23728/b2share.7a70f943dcdd48a0822f0f135b3ac2bc Copy

PID: 11304/8220b208-b61b-4a05-bf36-df5d56b6247a Copy

Files

Name

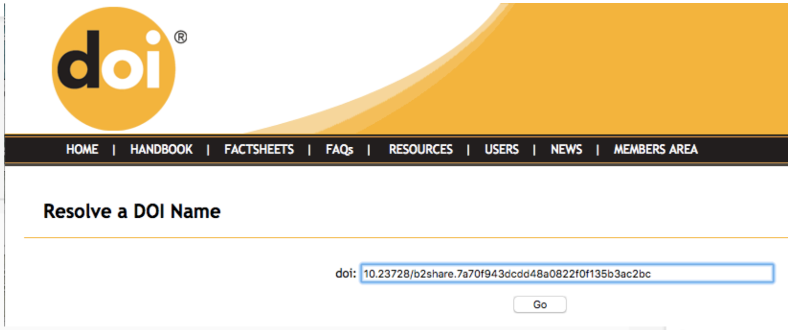
Metadata_Meteodata WaldhÑuser LTER database.xlsx

Checksum: `md5:3cdeaba5f3e9d99cb228161578699668`

PID: 11304/f73736bb-fade-45aa-9700-bdb3a07ff603 Copy

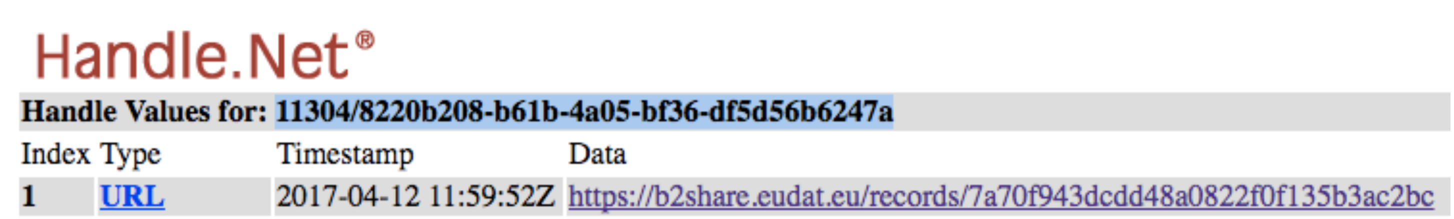
LTER

LTER


doi®
HOME | HANDBOOK | FACTSHEETS | FAQs | RESOURCES | USERS | NEWS | MEMBERS AREA

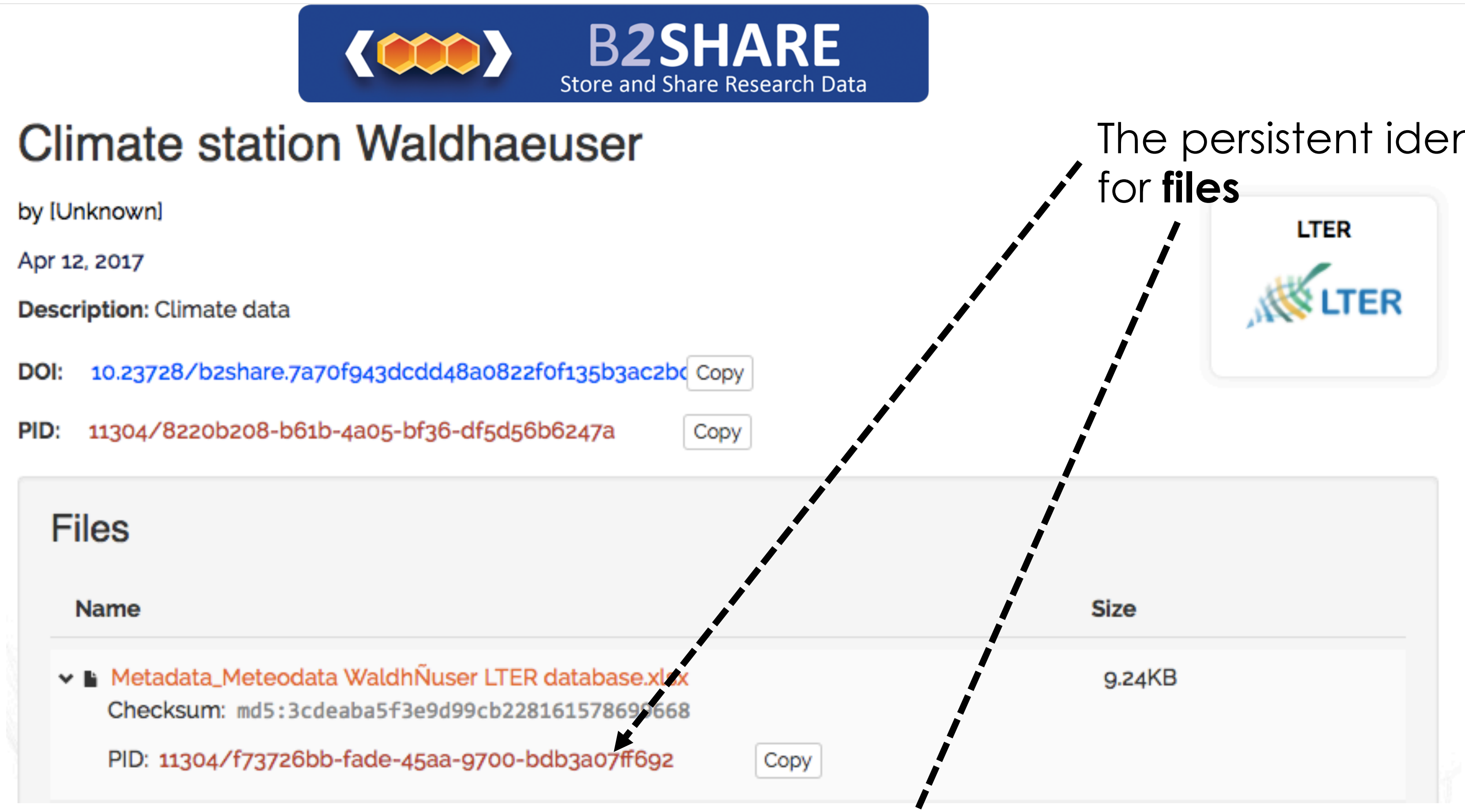
Resolve a DOI Name

doi: 10.23728/b2share.7a70f943dcdd48a0822f0f135b3ac2bc



Handle Values for: 11304/8220b208-b61b-4a05-bf36-df5d56b6247a

Index	Type	Timestamp	Data
1	URL	2017-04-12 11:59:52Z	https://b2share.eudat.eu/records/7a70f943dcdd48a0822f0f135b3ac2bc



B2SHARE
Store and Share Research Data

Climate station Waldhaeuser

by [Unknown]

Apr 12, 2017

Description: Climate data

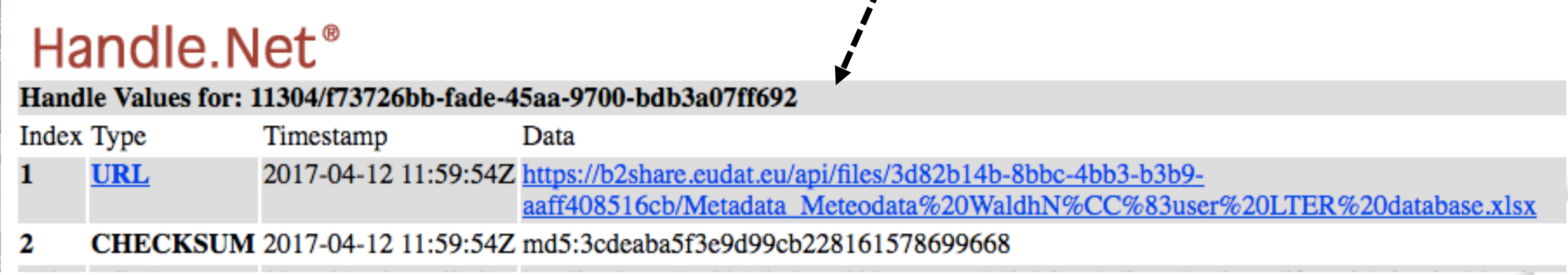
DOI: 10.23728/b2share.7a70f943dcdd48a0822f0f135b3ac2bc Copy

PID: 11304/8220b208-b61b-4a05-bf36-df5d56b6247a Copy

Files

Name	Size
Metadata_Meteodata WaldhÑuser LTER database.xlsx Checksum: md5:3cdeaba5f3e9d99cb228161578699668 PID: 11304/f73726bb-fade-45aa-9700-bdb3a07ff692 Copy	9.24KB

The persistent identifier for **files**



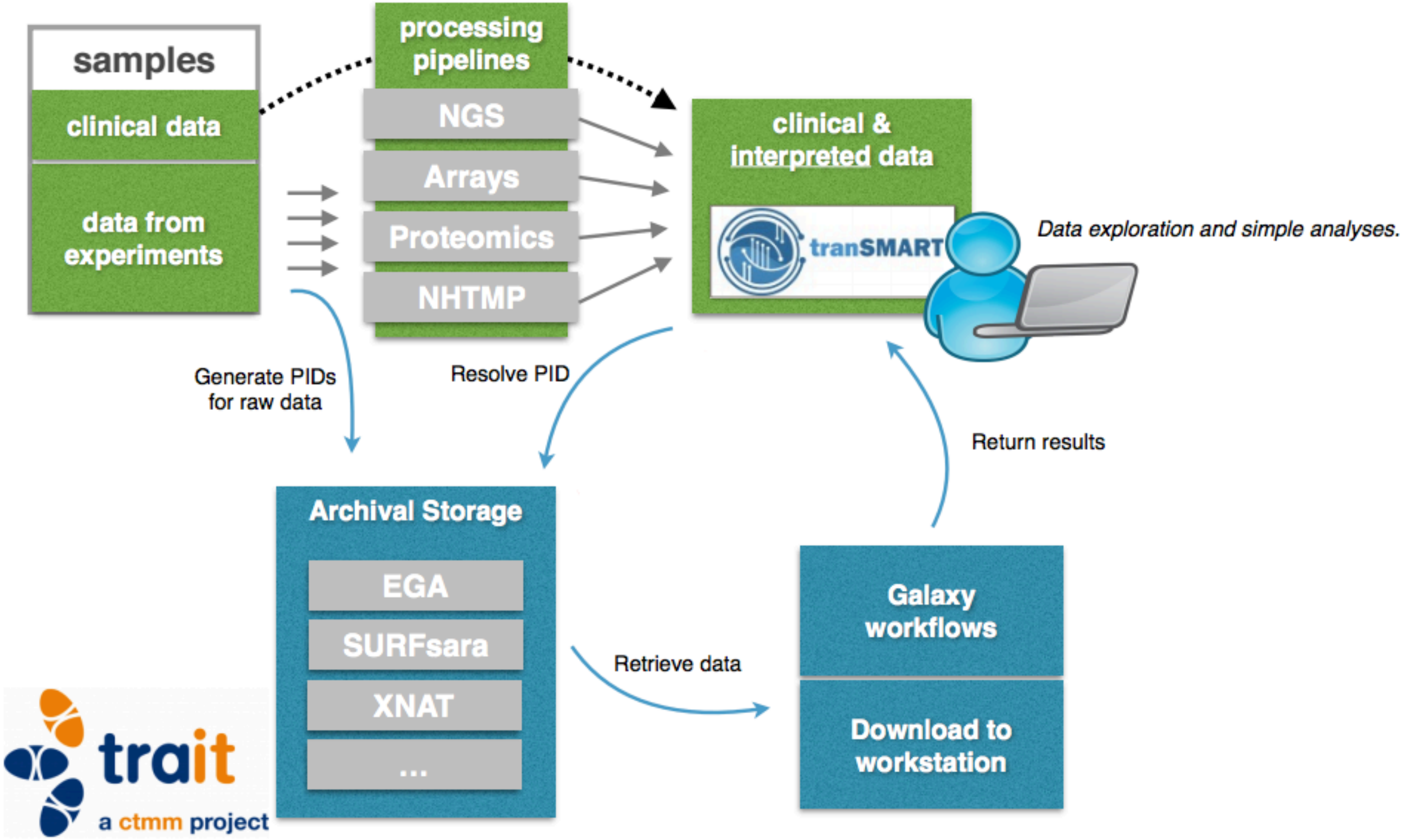
Handle.Net®

Handle Values for: 11304/f73726bb-fade-45aa-9700-bdb3a07ff692

Index	Type	Timestamp	Data
1	URL	2017-04-12 11:59:54Z	https://b2share.eudat.eu/api/files/3d82b14b-8bbc-4bb3-b3b9-aaff408516cb/Metadata_Meteodata%20WaldhN%CC%83user%20LTER%20database.xlsx
2	CHECKSUM	2017-04-12 11:59:54Z	md5:3cdeaba5f3e9d99cb228161578699668

Use Case 2: Enabling compute workflows

Molecular profiling dataflow in TraIT



samples

clinical data

data from experiments

processing pipelines

NGS

Arrays

Proteomics

NHTMP

clinical & interpreted data

tranSMART

Data exploration and simple analyses.

Generate PIDs for raw data

Resolve PID

Return results

Archival Storage

EGA

SURFsara

XNAT

...

Retrieve data

Galaxy workflows

Download to workstation

trait

a ctmm project

Use Case 2: Enabling compute workflows

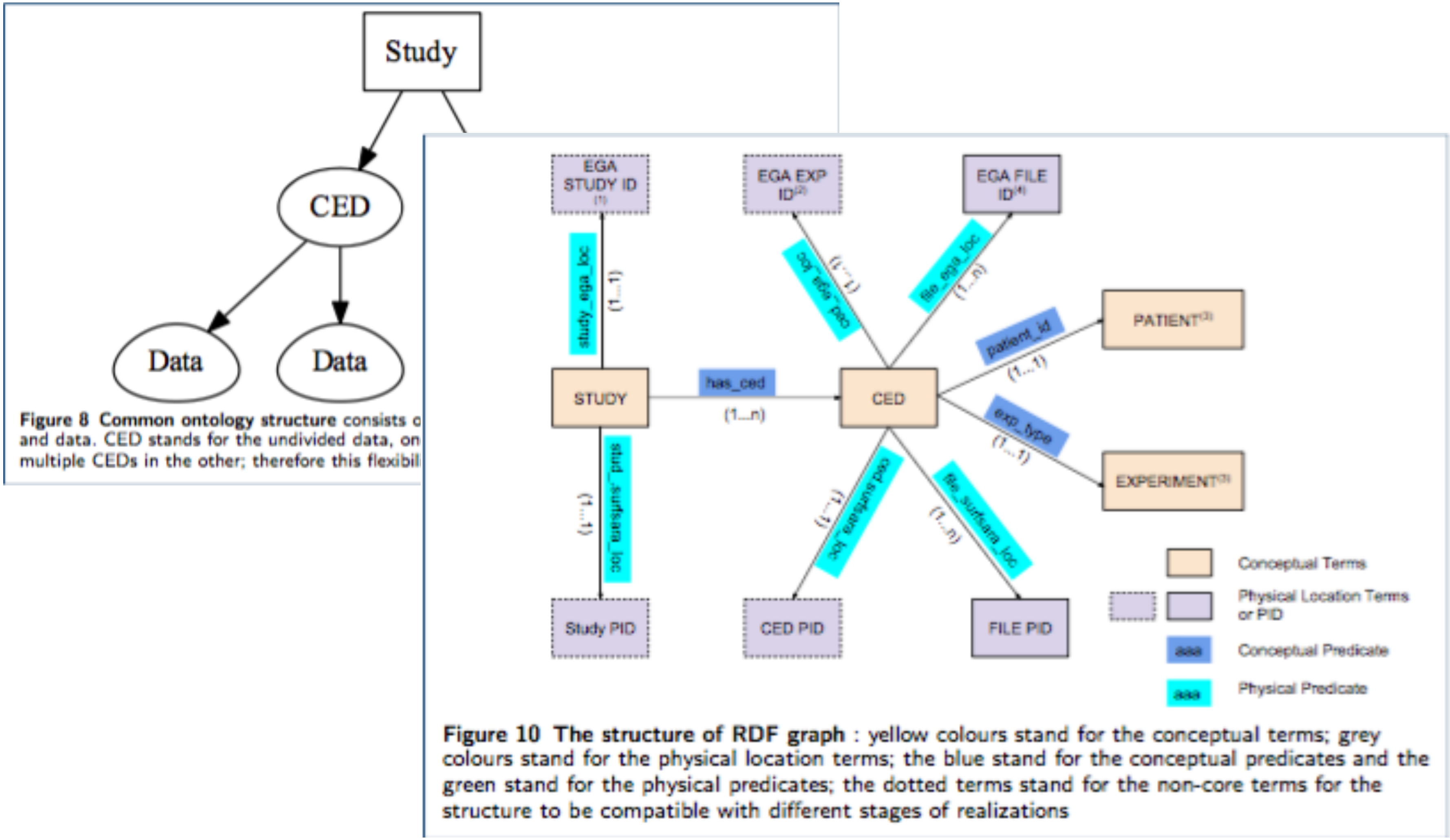


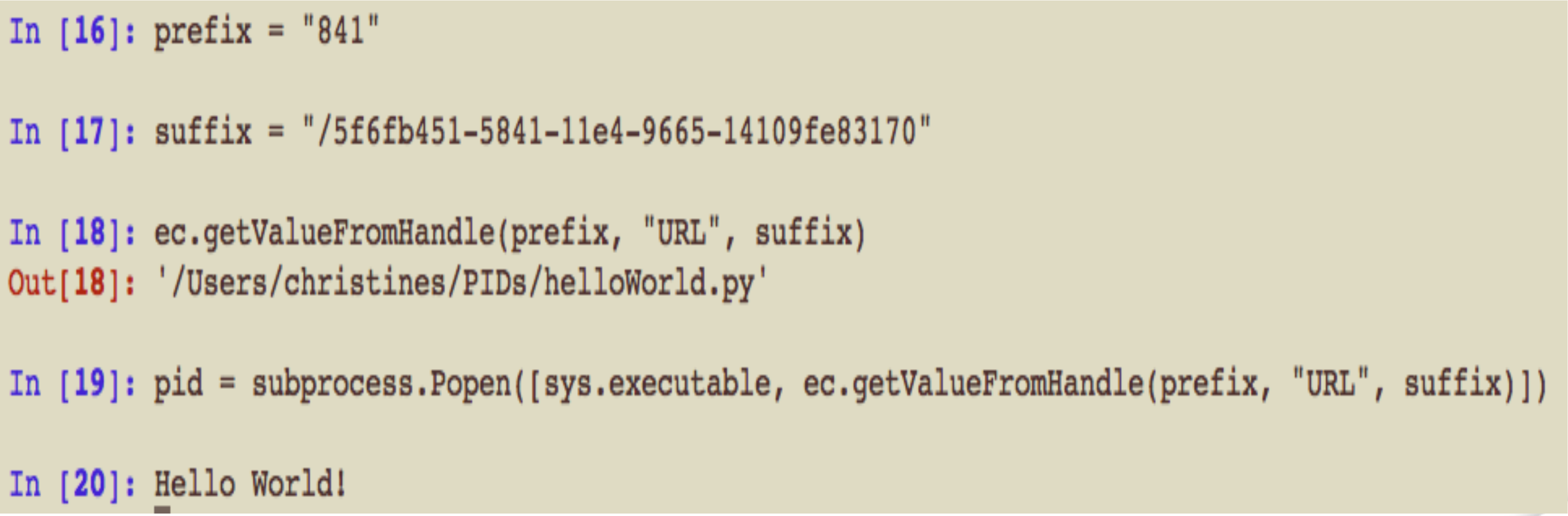
Figure 8 **Common ontology structure** consists o and data. CED stands for the undivided data, on multiple CEDs in the other; therefore this flexibil

Figure 10 **The structure of RDF graph** : yellow colours stand for the conceptual terms; grey colours stand for the physical location terms; the blue stand for the conceptual predicates and the green stand for the physical predicates; the dotted terms stand for the non-core terms for the structure to be compatible with different stages of realizations

Zhang, Abeln, Bijlard, Staiger: https://dx.doi.org/10.12688/f1000research.12168.1

Use Case 3: Labelling code

- Execute program hidden behind a PID
- A way to refer to workflows → reproducibility
- Example: Identification and resolving:



```
In [16]: prefix = "841"

In [17]: suffix = "/5f6fb451-5841-11e4-9665-14109fe83170"

In [18]: ec.getValueFromHandle(prefix, "URL", suffix)
Out[18]: '/Users/christines/PIDs/helloWorld.py'

In [19]: pid = subprocess.Popen([sys.executable, ec.getValueFromHandle(prefix, "URL", suffix)])

In [20]: Hello World!
```

PID systems

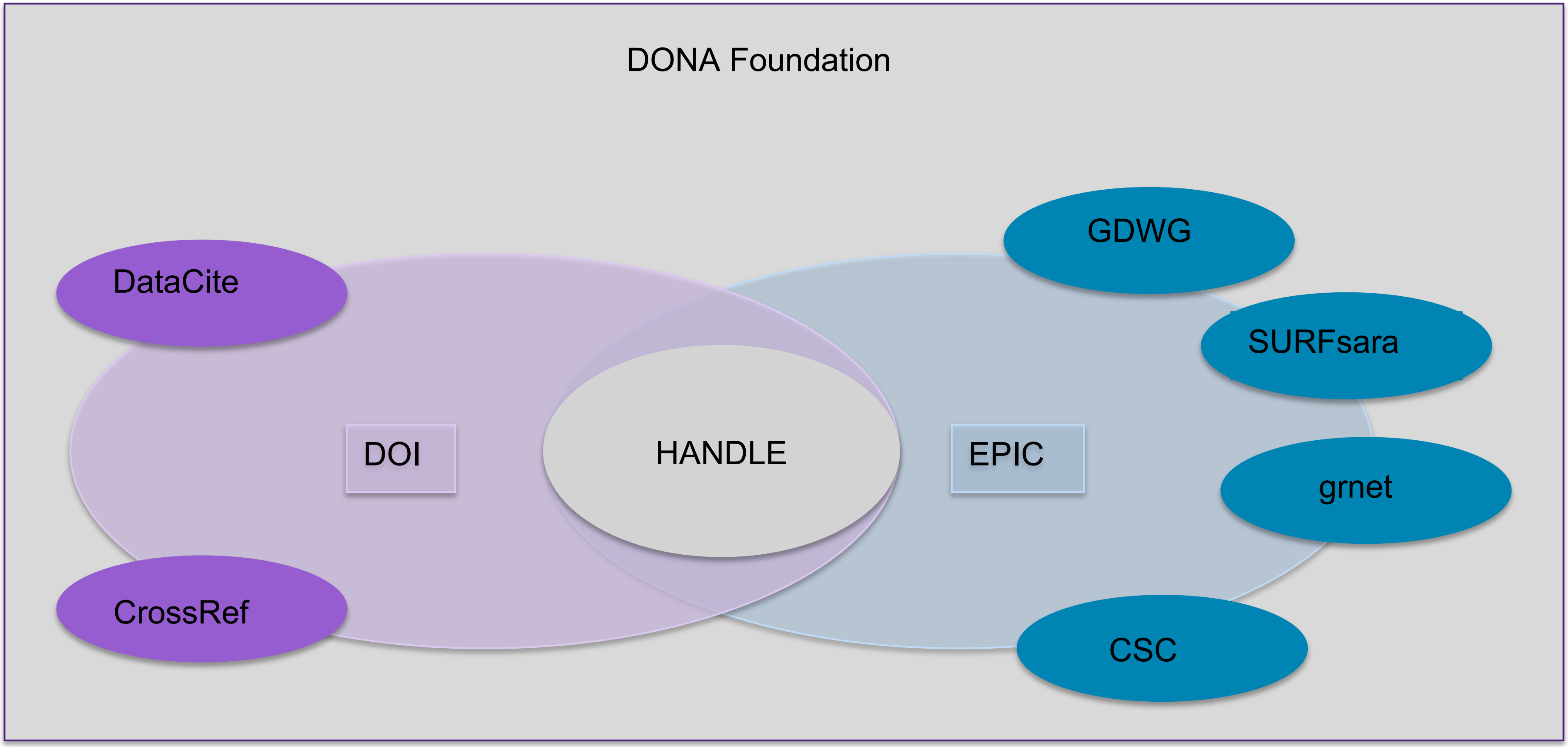
Hands-on: Resolution

- Resolve the PIDs
- What happens if you resolve a PID with a foreign resolver?

http://hdl.handle.net/21.T12995/PID-training

Exercise: Warming up!

PID systems and issuing authorities



PID systems and issuing authorities

- URN:NBN
  - Policies: PID is persistent and the data it is dereferenced to
  - Wants to be independent from transfer protocols
  - Currently all identifiers are compliant with http

- DOI
  - Policies: PID is persistent, data not
  - Based on the handle system
  - Datacite, Crossref are prefix issuing authorities
  - Requires extra metadata, stored in another database

- Both:
  - PIDs point to a landing page, not the file itself

    →Taylored towards data citation
  - User needs to provide a **minimum set of metadata** (Dublin Core)

PID systems and issuing authorities

- ePIC (European PID consortium)
  - Policies: PID is persistent, data is not
  - PIDs can point to anything
  - Based on the handle system
  - Taylored towards data identification and resolving

ePIC Persistent Identifiers for eResearch

- DONA foundation (www.dona.net)
  - Maintains global handle registry
  - Partners:
    - CNRI (developer of the handle system)
    - GDWG (main partner in ePIC)
    - International DOI foundation (IDF)

DONA

The Handle system – For whom?

- Metadata: You can create your own keyword-value pairs and store them with the PID

- PIDs allow to make a **distinction between data users and data managers**

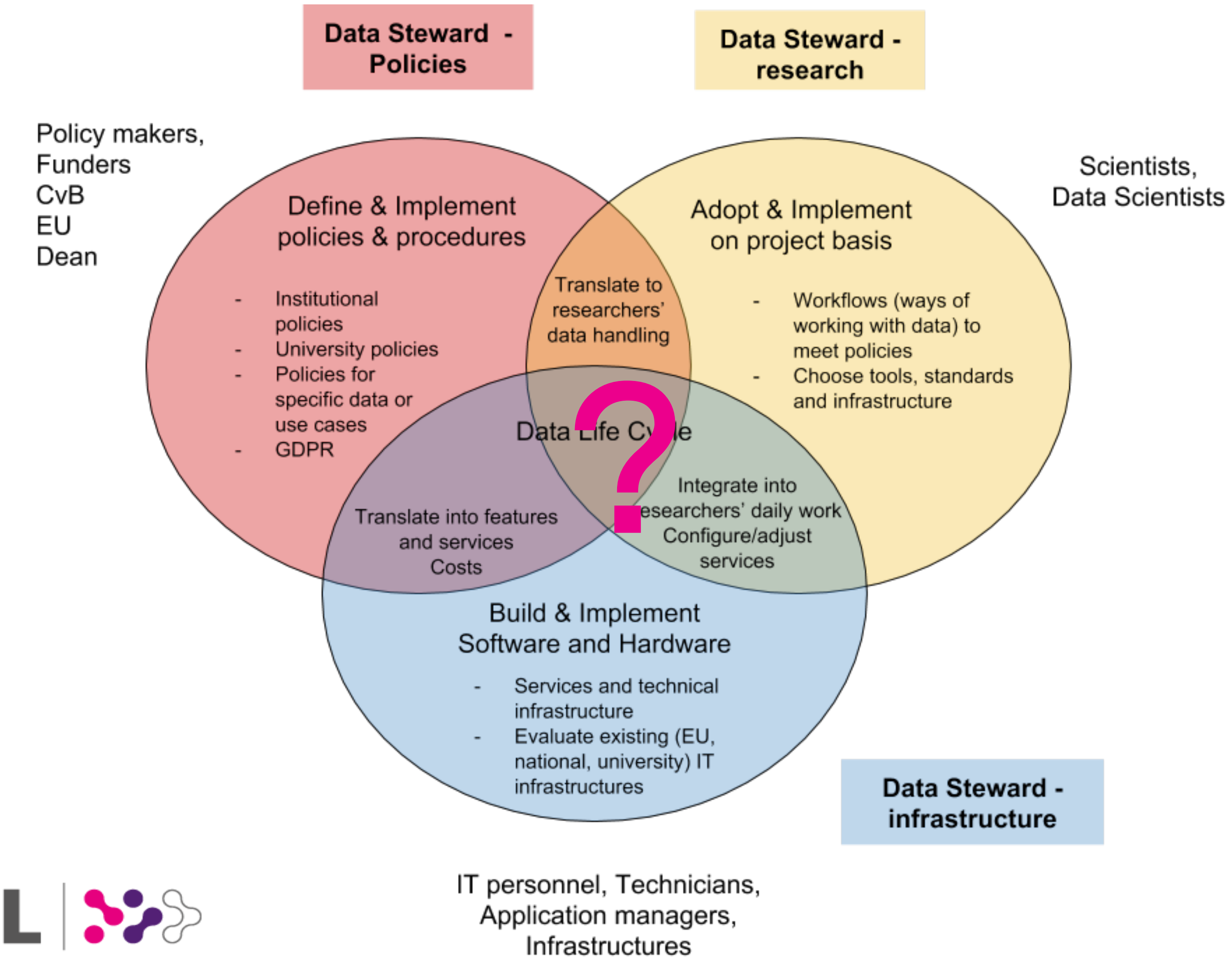
- **Data users get a PID** and can directly access the data, or the metadata stored with the PID
- Pipelines can programmatically access the metadata and start specific applications

- **Requires some serious thoughts** about data organisation and developing the **code to put data policies into practice**, including code maintenance

→ For **bigger research groups or consortia** working in a distributed data environment
→ For **repositories** who are in need of a host for their PIDs

Demo: Step-by-Step minting PIDs

- Register data with a Handle
- GET the details of a Handle
- Modify a Handle record
- Link two files on PID level
- Reverse look-up



Data Steward - Policies

Data Steward - research

Policy makers,
Funders
CvB
EU
Dean

Scientists,
Data Scientists

Define & Implement policies & procedures

- Institutional policies
- University policies
- Policies for specific data or use cases
- GDPR

Adopt & Implement on project basis

- Workflows (ways of working with data) to meet policies
- Choose tools, standards and infrastructure

Translate to researchers' data handling

Data Life Cycle

?

Translate into features and services
Costs

Integrate into researchers' daily work
Configure/adjust services

Build & Implement Software and Hardware

- Services and technical infrastructure
- Evaluate existing (EU, national, university) IT infrastructures

Data Steward - infrastructure

IT personnel, Technicians,
Application managers,
Infrastructures