Archiving and Preservation

Giacomo Cannizzaro

RDM Training – SURF – April 2024

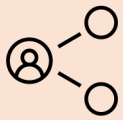
What do we have

A dataset

A set of metadata

What do we want

To preserve data



To make it available to other researchers

To follow guidelines from institutions and funders

Why?

- ❑ Open Science policies (national, institutional, funders etc.)
- ❑ Reproducibility (e.g., unique observations)
- ❑ Saving resources (e.g., expensive experiments)
- ❑ **Your own necessities**

How?

- ➢ Archives and Repositories
- ➢ Appropriate Metadata
- ➢ Persistent Identifiers

SURF

Long term preservation - archiving

Data that is not reproducible or difficult to reproduce

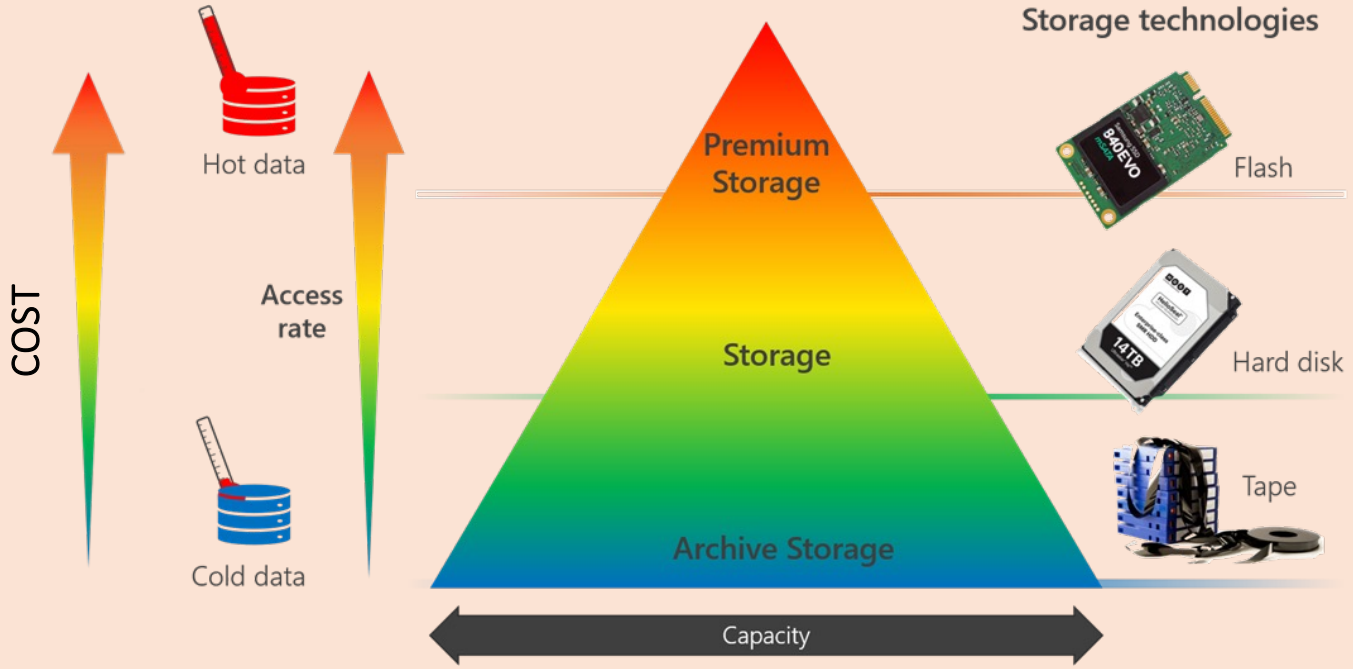
Astronomy, meteorology, also medical and SSH

Data that is expensive to reproduce

Particle or Material Physics, Satellite data

How to choose storage?

- Policies
- Size of data
- Frequency of access
- Sharing (internal/external)
- Resources available



https://tools.uu.nl/storagefinder/

SURF Data Archive

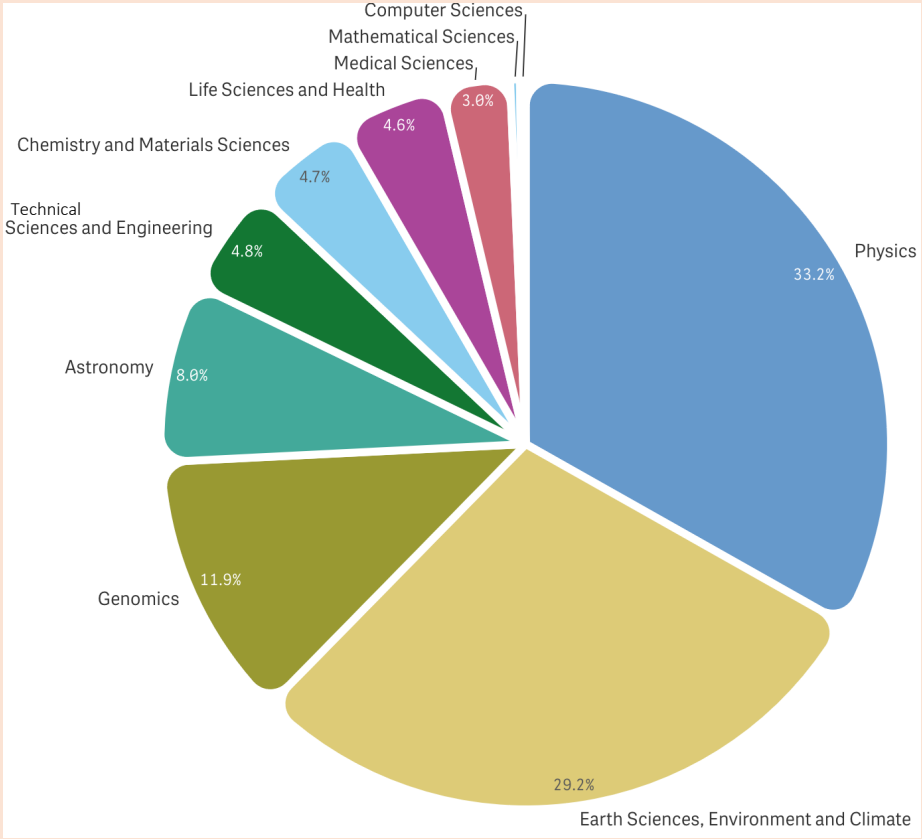
Large datasets! 10s Tb – Pb

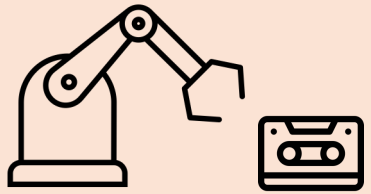
- Astronomy
- Meteorology
- Particle Physics

Double copy possible, cheap

Only for preservation

Offline, cold storage, physical downtime





https://ams17cam1.storage.surfsara.nl/

Long term preservation - Repositories

The Data Archive is not ”forward-facing”, it is only for long-term storage of data for a single user, research group or institution.

All metadata are for internal use only

If we need to make our data public, we use a **Repository**

SURF REPOSITORY

Same infrastructure of the Data Archive, but forward facing: the metadata are findable.

Ideal for (very) large datasets

SURF

Repositories – How to choose?

- Community best practices
- Type of content and field
- Size
- Availability of resources (contracts, institutional repositories, yoda)
- Policies and regulations
- Openness (e.g., metadata available, dataset no)

https://www.re3data.org/

https://fairsharing.org/

Metadata

Now your data is public, is it enough?

You need **metadata**: “data about data”

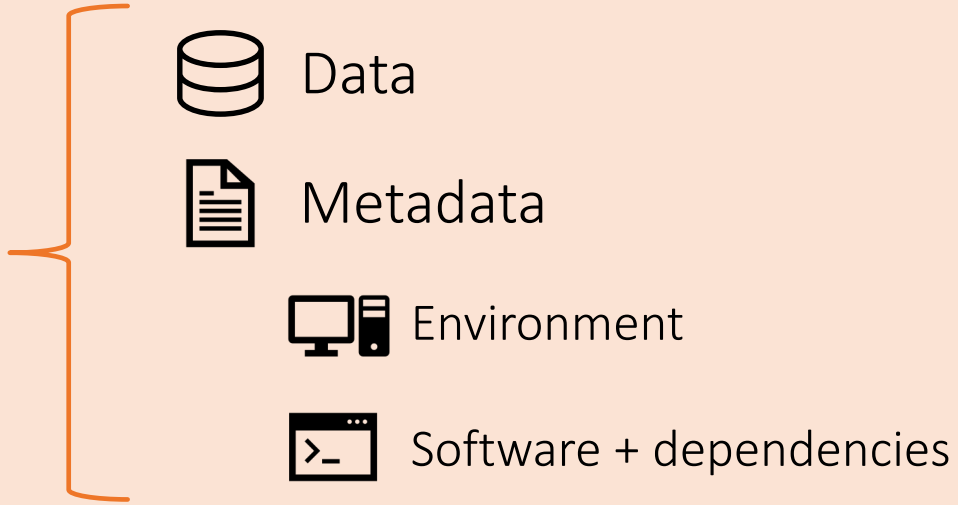
Metadata is what makes the dataset findable and accessible, without metadata, the dataset is only a group of bytes.

Aliens?

Metadata

Easy initial best practice: what metadata will I need in 1-5 years to understand how to use the dataset?

Repository data → Reproduction Package



Standards (community, funders, publishers)

https://fairsharing.org/

SURF

Persistent Identifiers

PIDs guarantee the long-term findability of digital resources (unlinke URLs that may break) through a long lasting, immutable reference.

<resolver service> / <prefix> / <suffix>

Resolver service: database to get information from
Prefix: identifies assigning body
Suffix: identifies resource

PID	Example
Handle	http://hdl.handle.net/2381/12775
DOI - Digital Object Identifier	http://doi.org/10.1186/2041-1480-3-9
ARK - Archival Resource	http://example.org/ark:/13030/tf5p30086k

SURF

Persistent Identifiers

<resolver service> / <prefix> / <suffix>

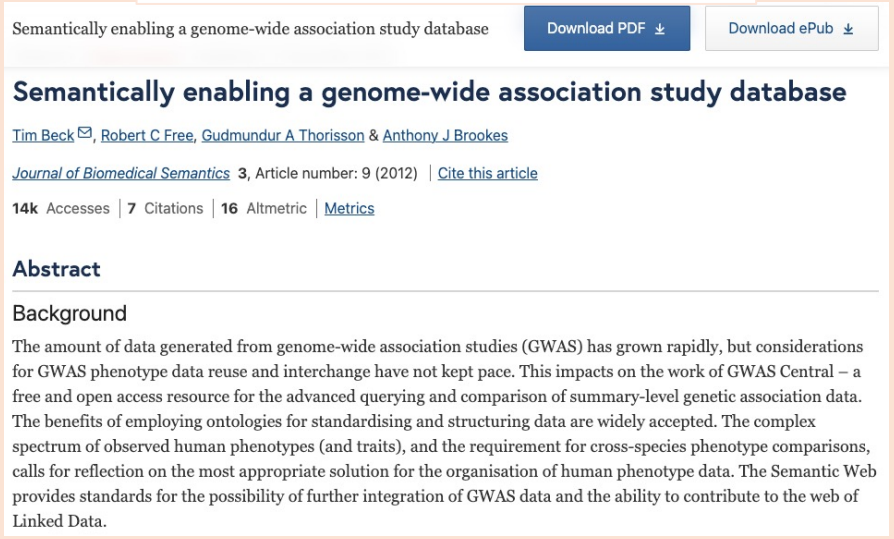
Landing page

- Metadata
- Resource

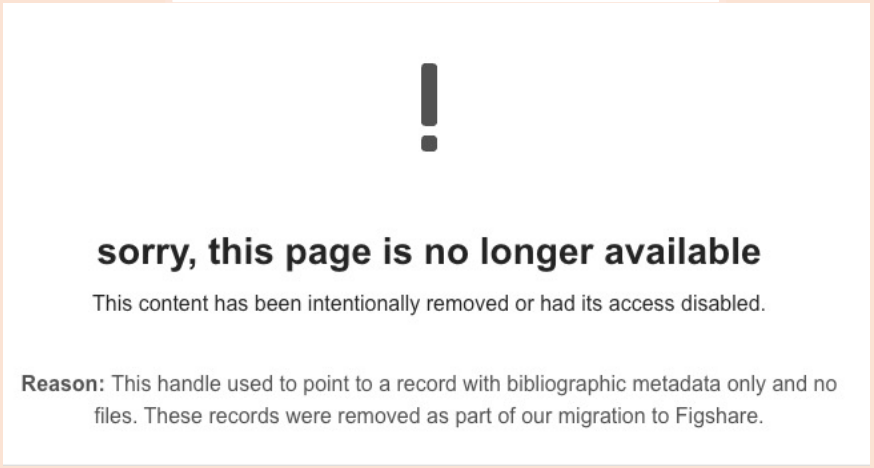
OR

- Tombstoning page

http://doi.org/10.1186/2041-1480-3-9

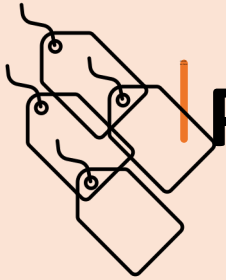


http://hdl.handle.net/2381/12775



Exercise: is this a good tombstone page?

Persistent Identifiers – Landscape



doi Foundation

DataCite

Crossref

ARK Alliance

Home of the Archival Resource Key (ARK)

B2INST

ORCID

Connecting research and researchers

SURF

ePIC

Persistent Identifiers for eResearch

ROR

SURF

Persistent Identifiers – How to choose

- Community standards: PIDs work when uptake is high, sensible to use what your community is using
- Flexibility in MD schema
  - Expertise level – tradeoff between flexibility and technical knowledge needed for setup
- Amount of PIDs to be minted (cost)
- Repository used

PIDwijzer - https://www.pidwijzer.nl/

FREYA (H2020 program) - 10.5281/zenodo.4192174

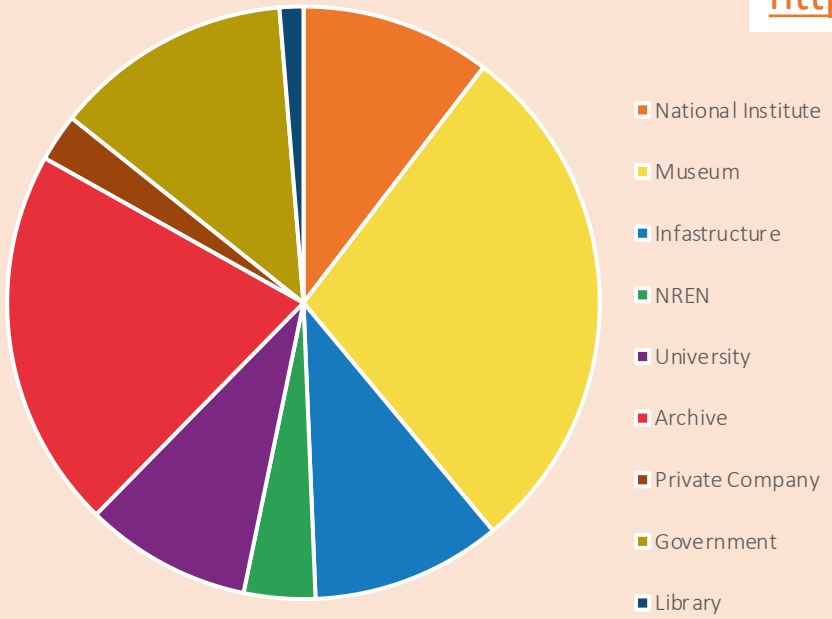
https://pidforum.org/

Persistent Identifiers – SURF ePIC PID

Part of European Persistent Identifier Consortium, Handle-based

- Affordable
- Scalable
- Flexible

- iRODs
- Yoda
- SURF Data Repository



http://hdl.handle.net/10934/RM0001.COLLECT.5216

Research Activity Identifier

https://raid.org/

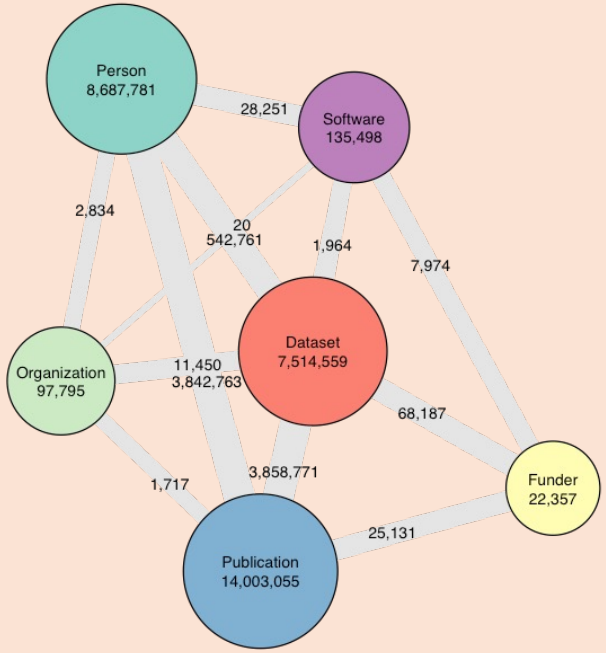
- PID for research project
- Connecting all elements of a project through the whole research lifecycle
  - People, grants, inputs, outputs…
- Being onboarded in SURF
- Project history/versioning
- Single source of truth

Thank you for your attention!

Persistent Identifiers - Graphs

PIDs guarantee the findability of digital resources, and likewise for their metadata.

Finding connected research/resources becomes easier, through graphs



DataCite

OpenAIRE Graph

PIDGraph
EOSC PID Graph

Relational metadata!

SURF