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

Long term preservation - archiving

Data that is not reproducible or difficult to reproduce

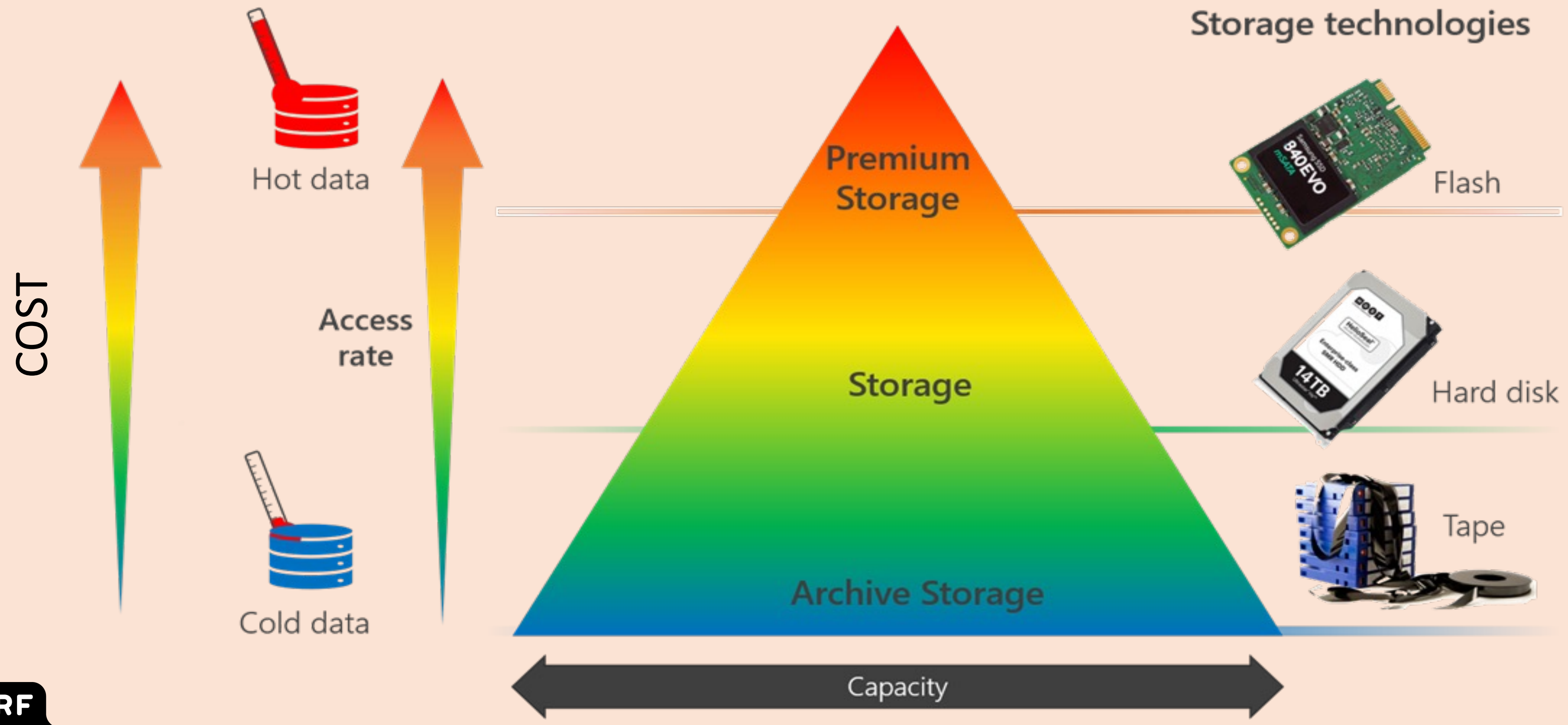
Astronomy, meteorology, also medical and SSH

Data that is expensive to reproduce

Particle or Material Physics, Satellite data

SURF

Storage types

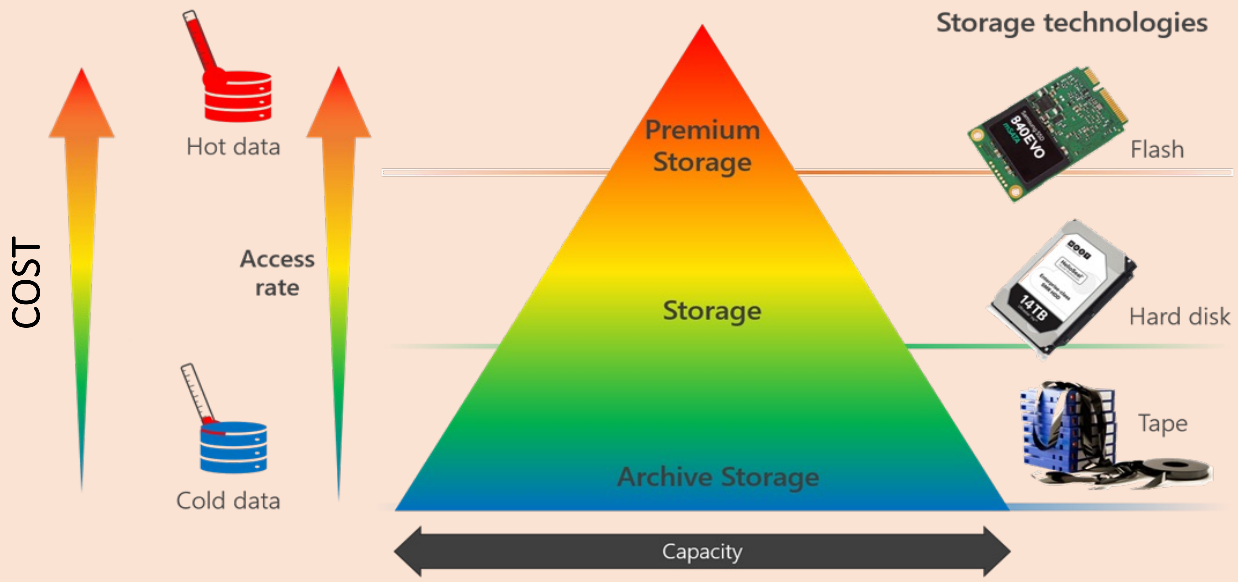


SURF

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

Repositories – How to choose?

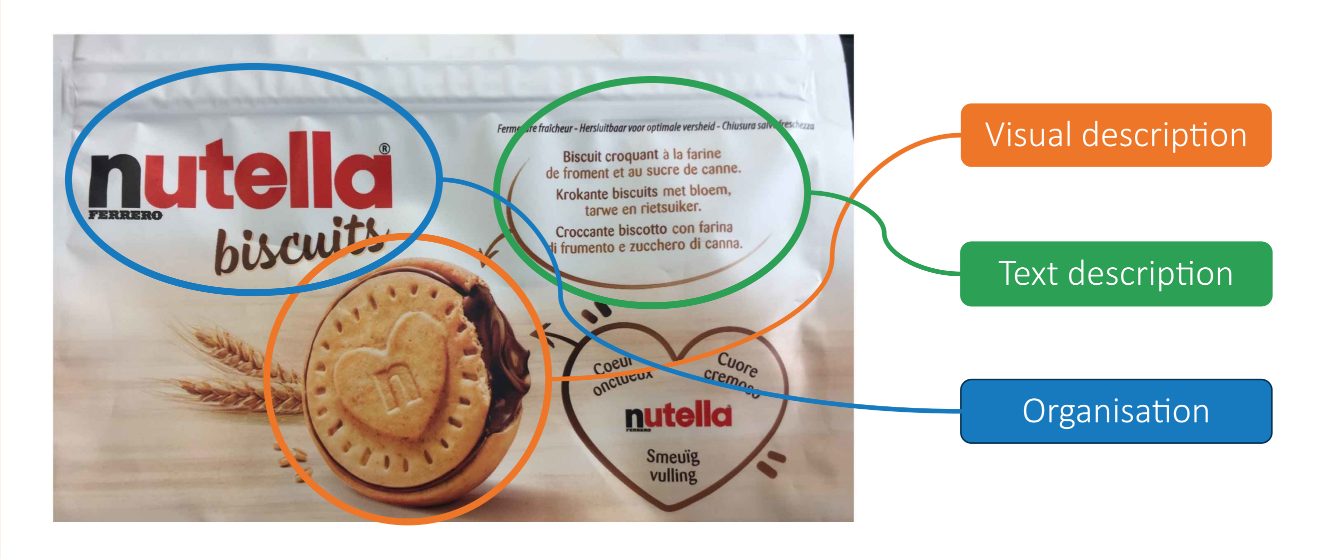
- Community best practices
- Type of content and field
- Size
- Availability of resources (contracts, institutional repositories)
- Policies and regulations

SURF REPOSITORY
figshare
DRYAD
OSF
HARVARD Dataverse
DataverseNL
Data Archiving and Networked
DANS
zenodo
4TU.ResearchData
SCIENCE • ENGINEERING • DESIGN
Research Software Directory

https://www.re3data.org/

https://fairsharing.org/

Metadata


nutella
FERRERO
biscuits
Fermeture fraîcheur - Hersluitbaar voor optimale versheid - Chiusura salva freschezza
Biscuit croquant à la farine de froment et au sucre de canne.
Krokante biscuits met bloem, tarwe en rietsuiker.
Croccante biscotto con farina di frumento e zucchero di canna.
Coeur onctueux
Cuore cremoso
nutella
Smeuïg vulling

Visual description

Text description

Organisation

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.

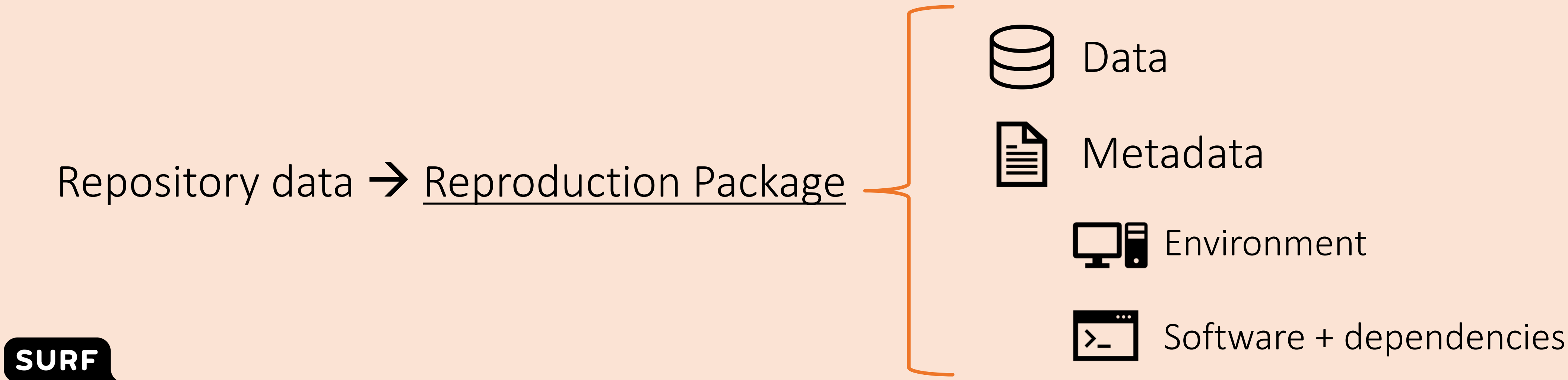
Aliens?

Metadata

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

Standards (community, funders, publishers)

https://fairsharing.org/



Repository data → Reproduction Package

Data

Metadata

Environment

Software + dependencies

Persistent Identifiers

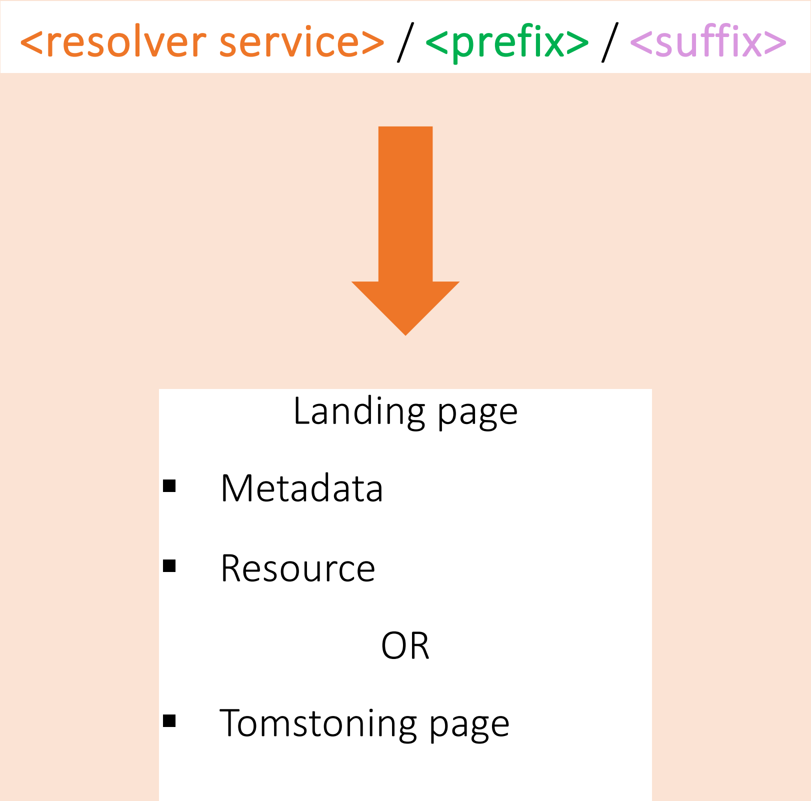
PIDs guarantee the long-term findability of digital resources (unlinke URLs that may break) through a long lasting, immutable reference.

<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

Persistent Identifiers



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

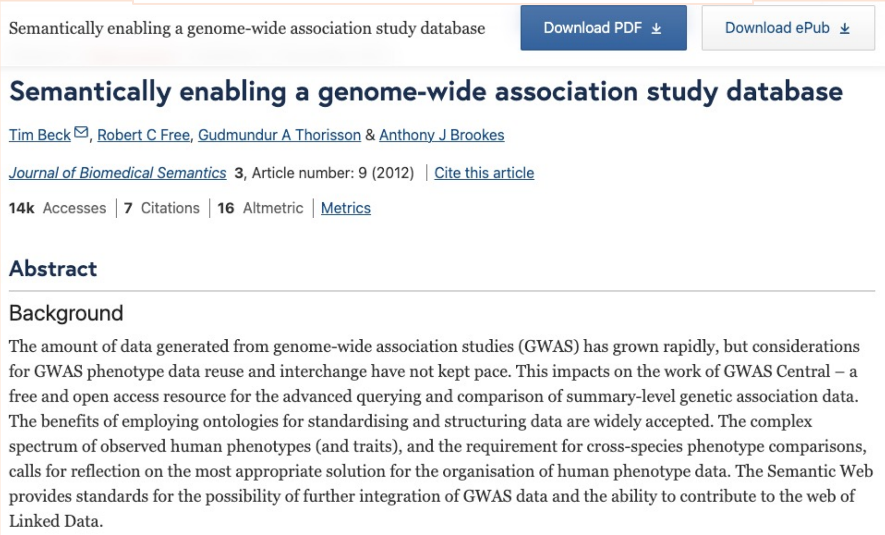
Landing page

- Metadata
- Resource

OR

- Tomstoning page

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



Semantically enabling a genome-wide association study database

Semantically enabling a genome-wide association study database

Tim Beck, Robert C Free, Gudmundur A Thorisson & Anthony J Brookes

Journal of Biomedical Semantics 3, Article number: 9 (2012) | Cite this article

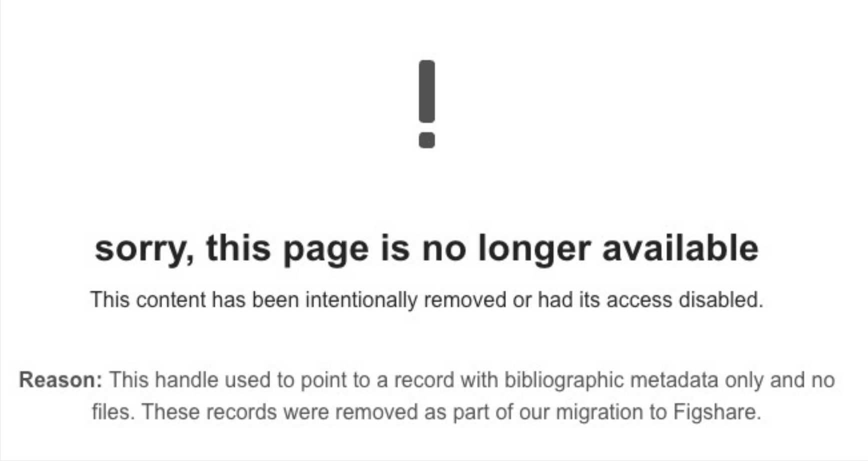
14k Accesses | 7 Citations | 16 Altmetric | Metrics

Abstract

Background

The amount of data generated from genome-wide association studies (GWAS) has grown rapidly, but considerations for GWAS phenotype data reuse and interchange have not kept pace. This impacts on the work of GWAS Central – a free and open access resource for the advanced querying and comparison of summary-level genetic association data. The benefits of employing ontologies for standardising and structuring data are widely accepted. The complex spectrum of observed human phenotypes (and traits), and the requirement for cross-species phenotype comparisons, calls for reflection on the most appropriate solution for the organisation of human phenotype data. The Semantic Web provides standards for the possibility of further integration of GWAS data and the ability to contribute to the web of Linked Data.

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



!

sorry, this page is no longer available

This content has been intentionally removed or had its access disabled.

Reason: This handle used to point to a record with bibliographic metadata only and no files. These records were removed as part of our migration to Figshare.

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, good to use what your community is using
- 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/

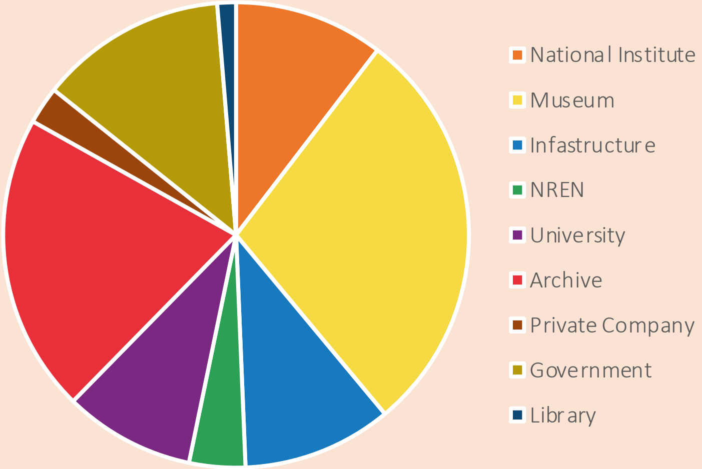
FREYA (H2020 program) - 10.5281/zenodo.4192174

https://pidforum.org/

Persistent Identifiers – SURF ePIC PID

Part of European Persistent Identifier Consortium, Handle-based

- Affordable
- Scalable
- Flexible



- iRODs
- Yoda
- SURF Data Repository

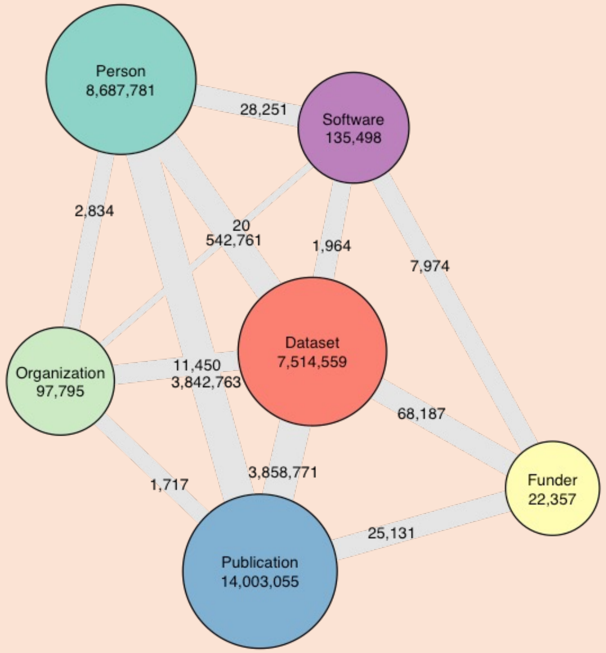
http://hdl.handle.net/10934/RM0001.COLLECT.5216

SURF

Persistent Identifiers - Graphs

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

Finding connected research/resources becomes easier, through graphs



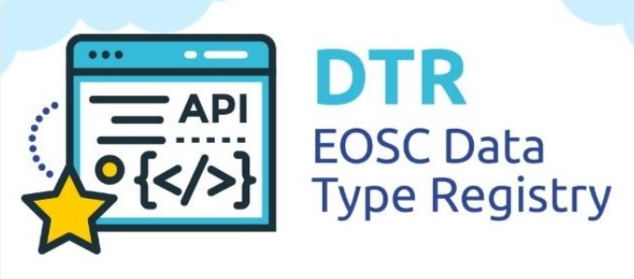
DataCite

OpenAIRE Graph

PIDGraph
EOSC PID
Graph

Relational metadata!

Data Type Registry & Metadata Schema Crosswalk Registry



- Effort to standardise Data Types with a registry
- Data type? Also a PID metadata element
- Machine actionable standardisation

- Metadata Interoperability
- Create and register schemas and crosswalks
- Uses DTR information

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!

SURF