Collaborative
EUDAT Data Infrastructure

PRACE

Data Preservation 1

The Why, The What and the How

Shaun de Witt, UKAEA

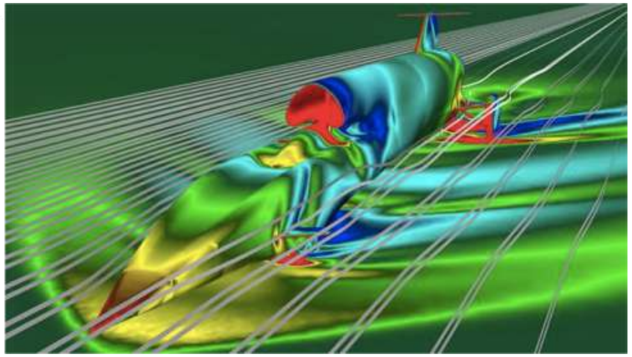
Agenda

- What
  - Types of Data
  - Considerations
- Why?
  - Scientific Benefit
  - Social Benefit
  - Economic Benefit
- FAIR and Open Data
  - FAIR Principles
  - Difference between Open and FAIR
- How
  - Planning your data management

Collaborative
EUDAT Data Infrastructure

PRACE

What is Data

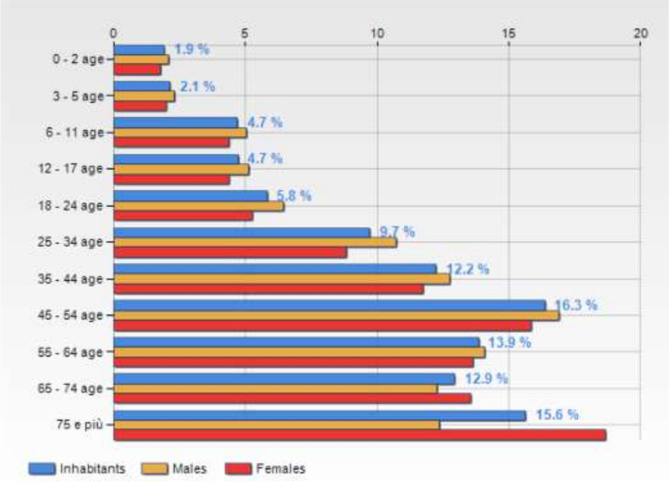


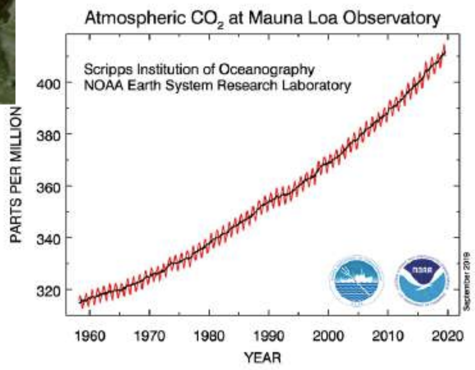
CFD Model of Bloodhound
Unstructured, derived, simulation

- Data can be split in many ways
  - **Structured** (databases), **semi-structured** (HDF files), **unstructured** (e-mails)
  - **Primary** (data taken by observer), **secondary** (data taken by someone else), **derived** (data after processing)
  - **Synoptic** (point in time, satellite image), **temporal** (analysis over time, seismic data), **geo-spatial** (data in a specific area, e.g. sea surface temperature maps)
  - **Observational**, **Experimental**, **Survey**, **Simulation**

Satellite Data
Secondary, synoptic, observational
, Semi-structured

Population Data
Derived, synoptic, survey
, Un-structured





CO2 Observations
Primary, temporal, observational
structured

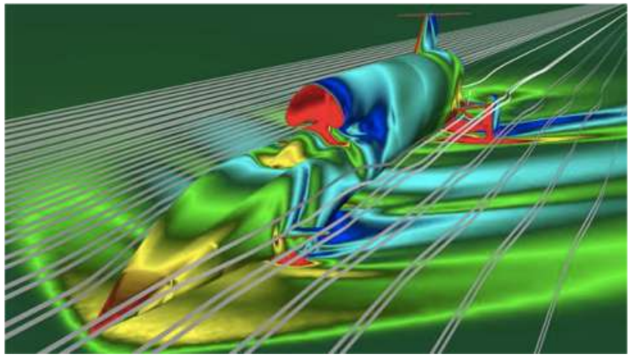
Data Sources…

Models

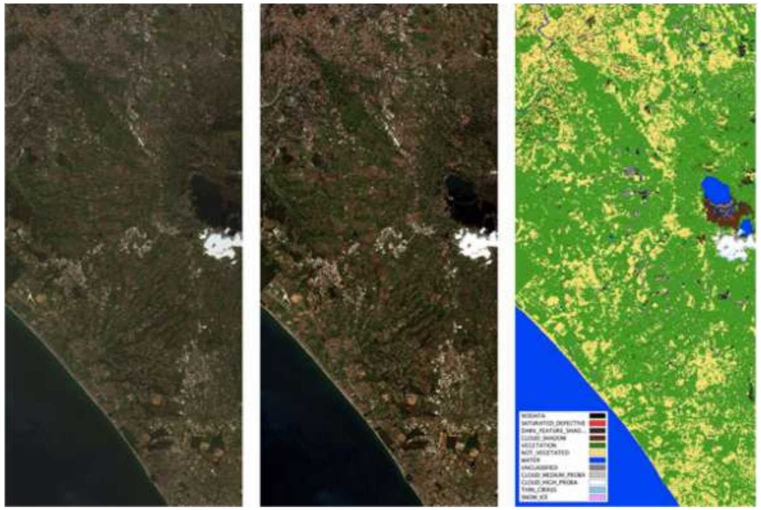
Batch Script

Software

CFD Model of Bloodhound Unstructured, derived, simulation

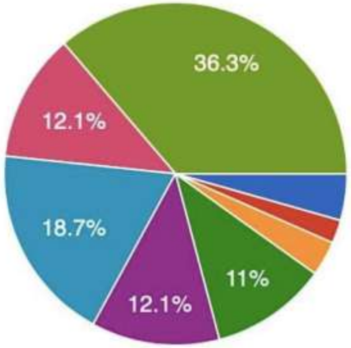


Observations



(1) Sentinel-2 Level-1C TOA reflectance input image, (2) the atmospherically corrected Level-2A BOA reflectance image, (3) the output scene classification of the Level-1C product.
(c) ESA

Surveys



The Data Pyramid

Increasing Volume

Increasing Sharing

Published Data

Registered Data

Working Data

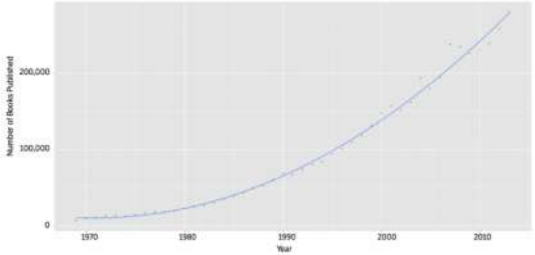
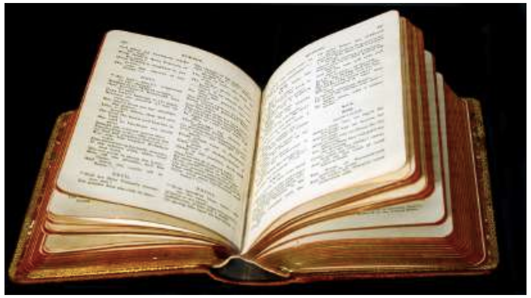


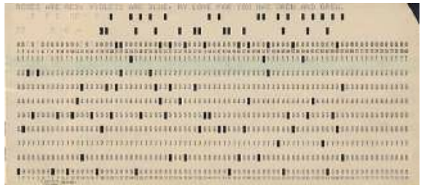
Figure 1: Number of individually titled English language books classified as single works of fiction published per year from 1969 to 2014. Information derived from Bowker's *Books in Print* database.

So What is Data Storage?

Pre-Digital



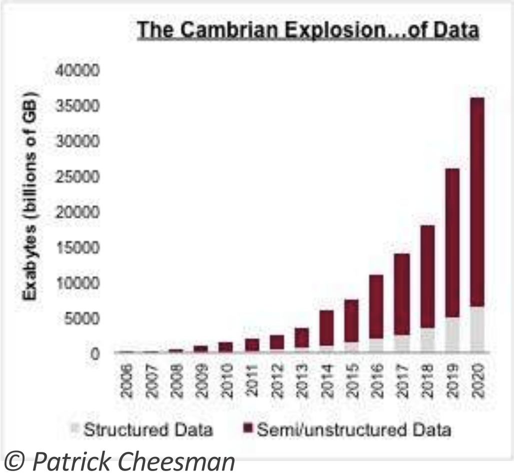
Digital



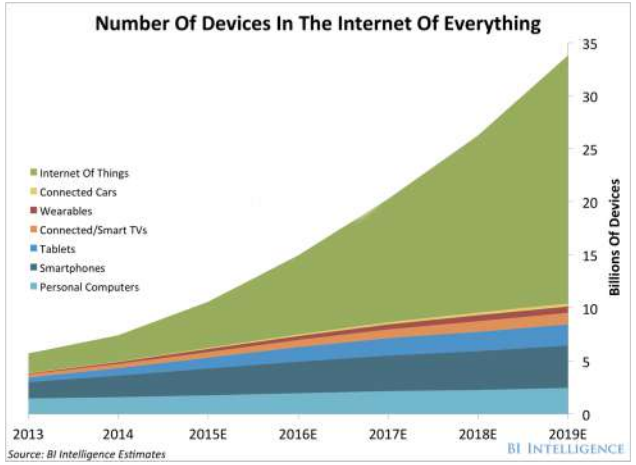
900 bits

What is the problem with Storage?

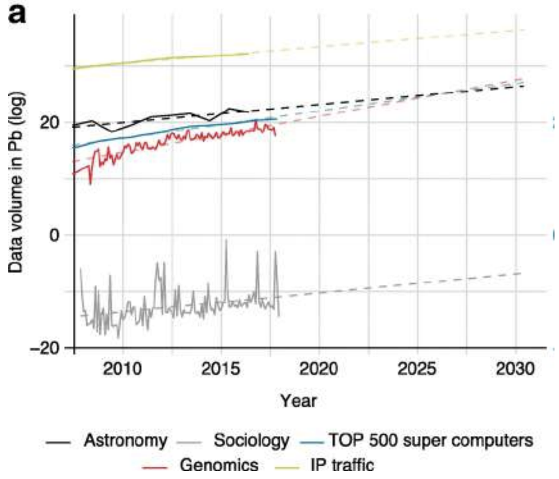
Meaningless Graphs



© Patrick Cheesman



Source: BI Intelligence Estimates



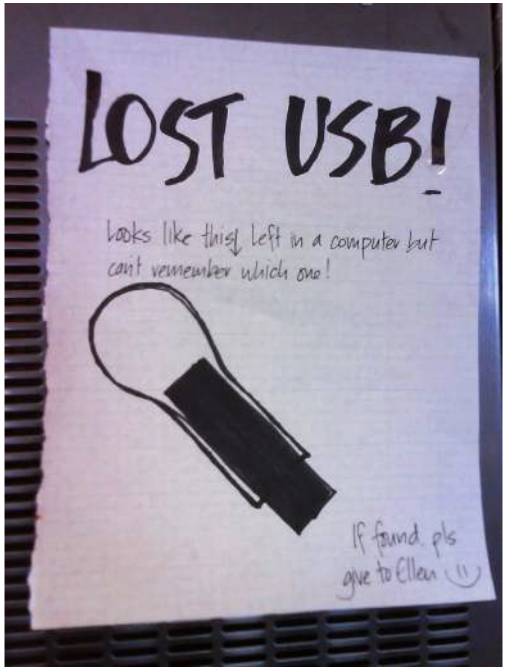
Fábio C. P. Navarro, Hussein Mohsen, Chengfei Yan, Shantao Li, Mengting Gu, William Meyerson & Mark Gerstein Genome Biologyvolume 20, Article number: 109 (2019)

Real Examples…

- *In 1 minute (2019):*
  - *350 000 tweets are twitted*
  - *208 million emails are sent (mostly to me)*
  - *500 hours of video are uploaded to youtube*
  - *3.3 million facebook posts are posted*
  - *66000 photos and videos are shared on Instagram*
- *And that is JUST social media!*

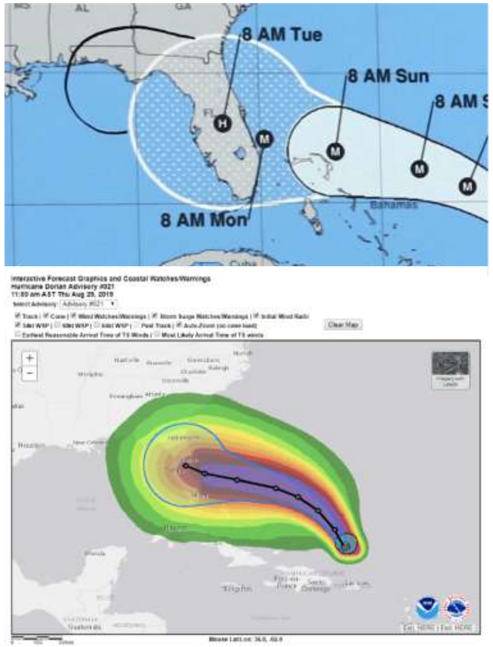
Is Data Storage the Same as Data Preservation?

- No!
- Storage is subject to many issues
  - Media failure
  - 'bit rot'
  - Natural disaster
  - Format obsolescence
  - Human deletion – accidental or malicious
  - Media Loss
  - Loss of funding
  - Link rot and reference rot
- Data Preservation is ensuring data is available for re-use into future
  - **Minimising** the risk of data loss


LOST USB!
Looks like this left in a computer but can't remember which one!
If found pls give to Ellen :)

Why Preserve Your Data?

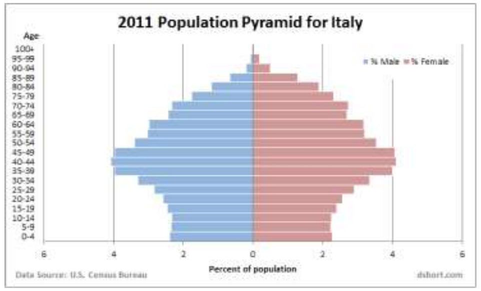
Reproducability



Interactive Forecast Graphics and Coastal Watches/Warnings
Hurricane Dorian Advisory #021
11:00 am AST Thu Aug 29, 2019
Select Advisory: Advisory #021 ▼
☑ Track | ☑ Cone | ☑ Wind Watches/Warnings | ☑ Storm Surge Watches/Warnings | ☑ Initial Wind Radii
☑ 34kt WSP | ☐ 50kt WSP | ☐ 64kt WSP | ☐ Past Track | ☑ Auto-Zoom (on cone load)
☐ Earliest Reasonable Arrival Time of TS Winds | ☐ Most Likely Arrival Time of TS winds

Alternative Facts

Uniqueness



[citation needed]
[citation needed]
[citation needed] [citation needed]
[citation needed] [citation needed]
[citation needed] [citation needed] [citation needed]
[citation needed] [citation needed] [citation needed]

Recognition

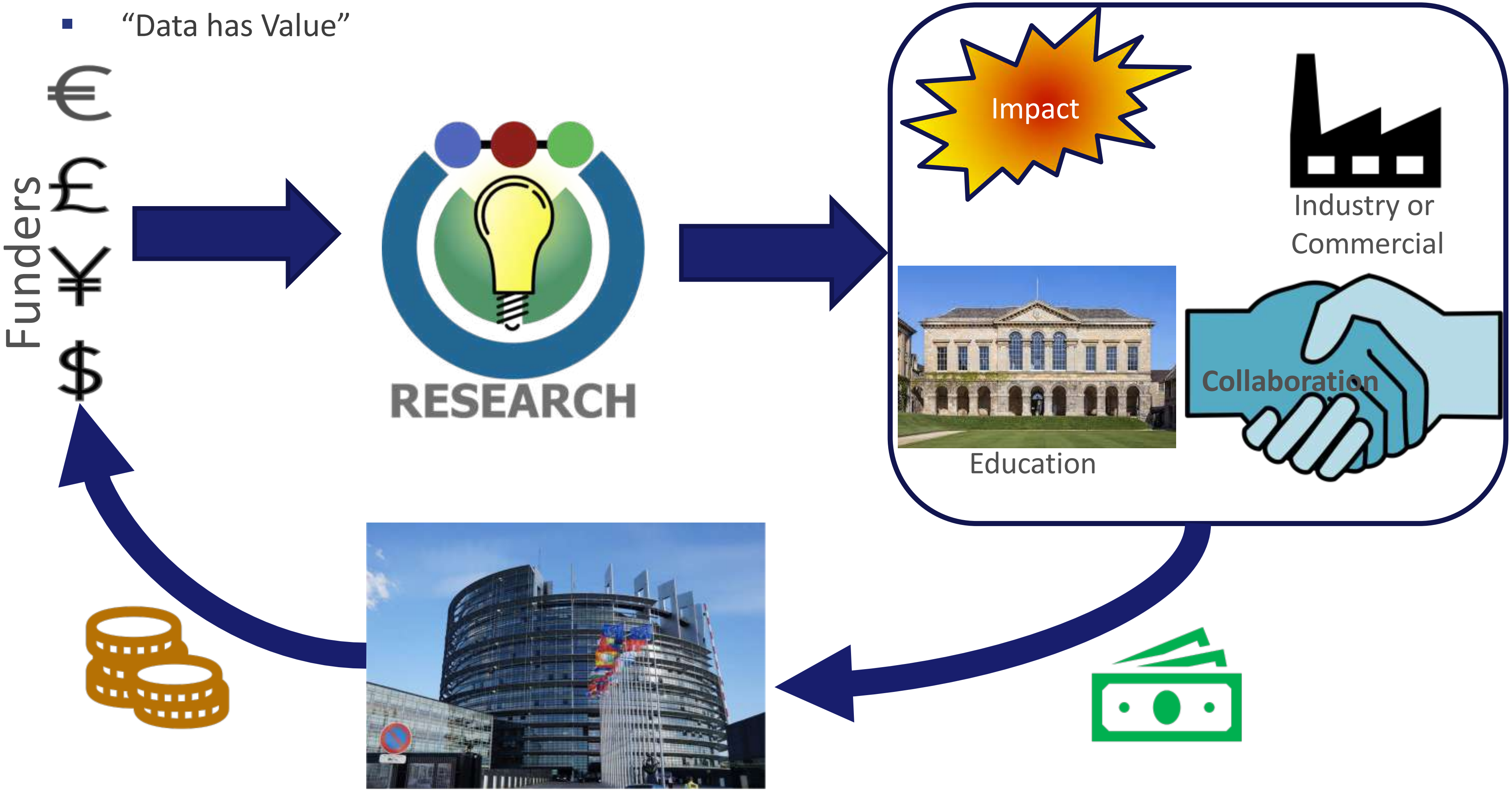
Cost

Legal

The ***real*** *reason? Your funding agency says you have to!*

Why do Funders Want you to Preserve your data?

- "Data has Value"

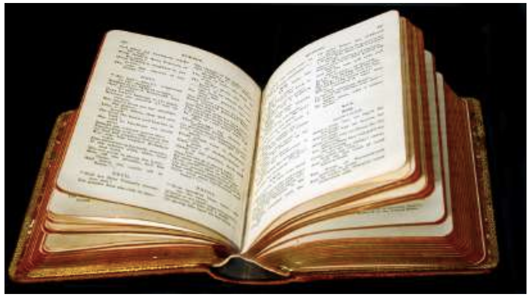
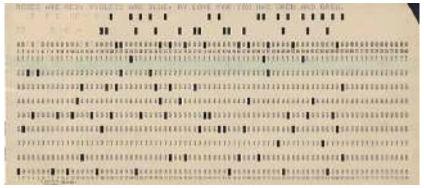


Preservation vs Curation

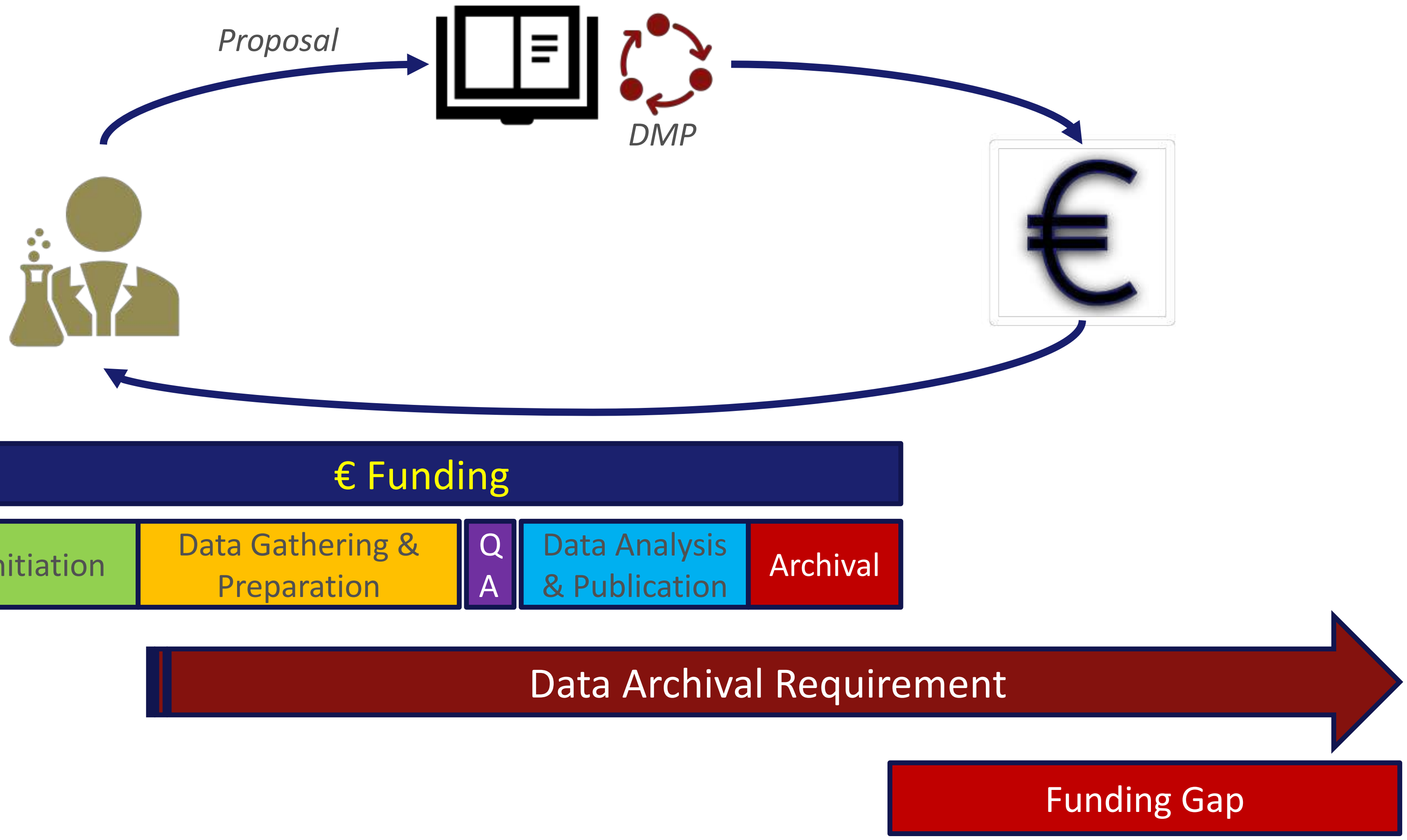
- **Definitions (lexico.com)**
- **<u>Preserve</u>**
  - Maintain (something) in its original or existing state.
- **<u>Curate</u>**
  - Select, organize, and look after items (in a collection or exhibition)

- In the digital era the concepts are similar:
  - Preserve: Make sure the **bits** you store stay the same over time
  - Curate: Make sure the **information** you store is discoverable and usable over time
- Lets look at some examples…

Preservation vs Curation - QUIZ

	Preservation 👍	*Curation* 👎
	*Preservation* 👍	*Curation* ?
	*Preservation* 👎	*Curation* 👎
	Preservation ?	*Curation* ?

Preservation – The Funding Problem



€ Funding

Initiation | Data Gathering & Preparation | QA | Data Analysis & Publication | Archival

Data Archival Requirement

Funding Gap

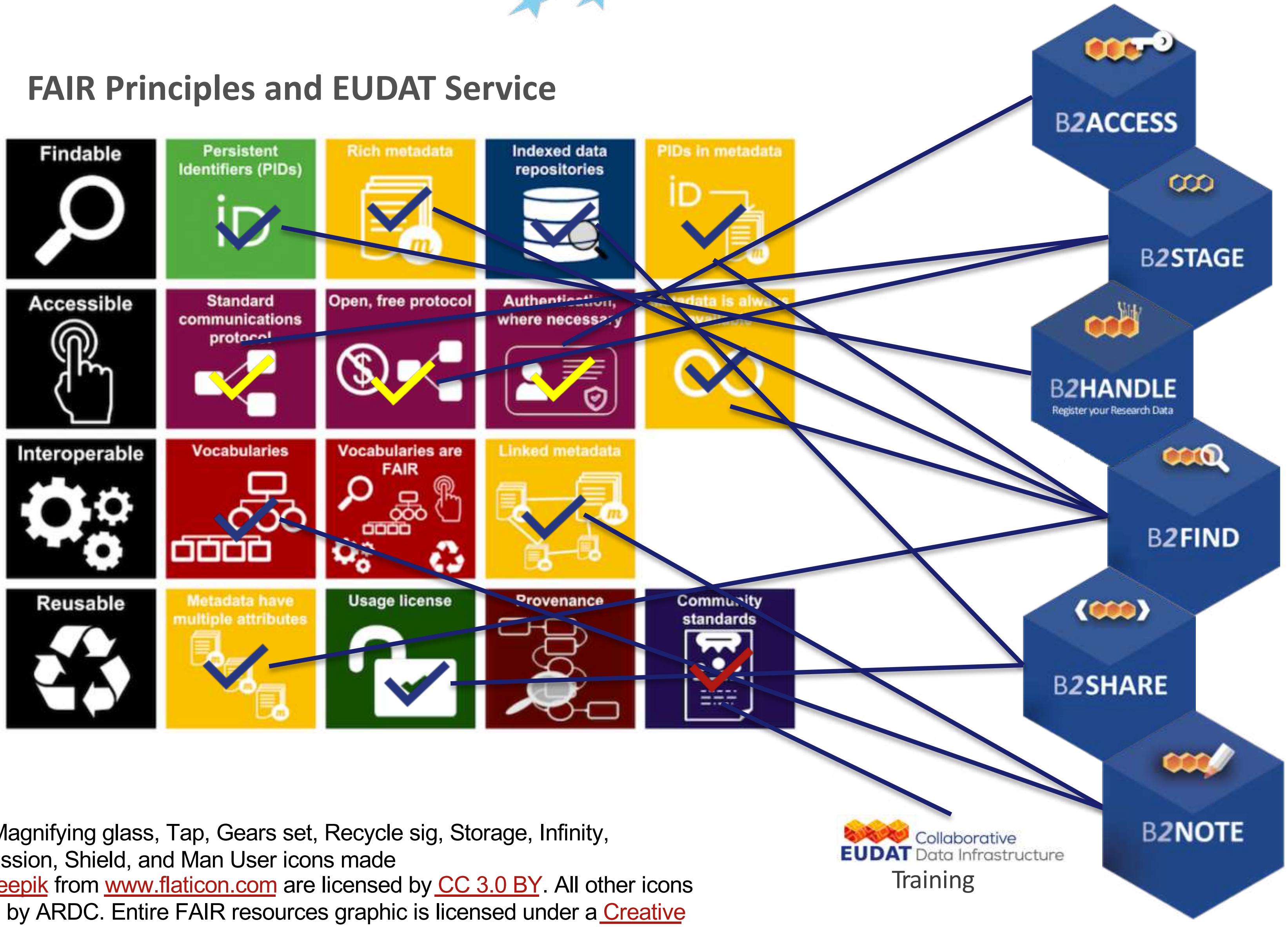
FAIR and Open Data

- Definition of **open** data:
  "Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike." (Open Knowledge Foundation)
  - Full definition at https://opendefinition.org/od/2.1/en/
  - E.g. licensed under CC0 (https://creativecommons.org/share-your-work/public-domain/cc0/)
- Definition of **FAIR** data
  "…a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable…" (https://www.force11.org/group/fairgroup/fairprinciples)
- So what is the difference?
  - Open data **need not** be findable
  - FAIR data **need not** be free – you may need to register and/or pay to access
  - FAIR data **is** interoperable and reusable, open data **can be** reused
- H2020 Guidelines on this:
  "Data should be as open as possible, as closed as is necessary"

Collaborative EUDAT Data Infrastructure

PRACE

FAIR Principles and EUDAT Service

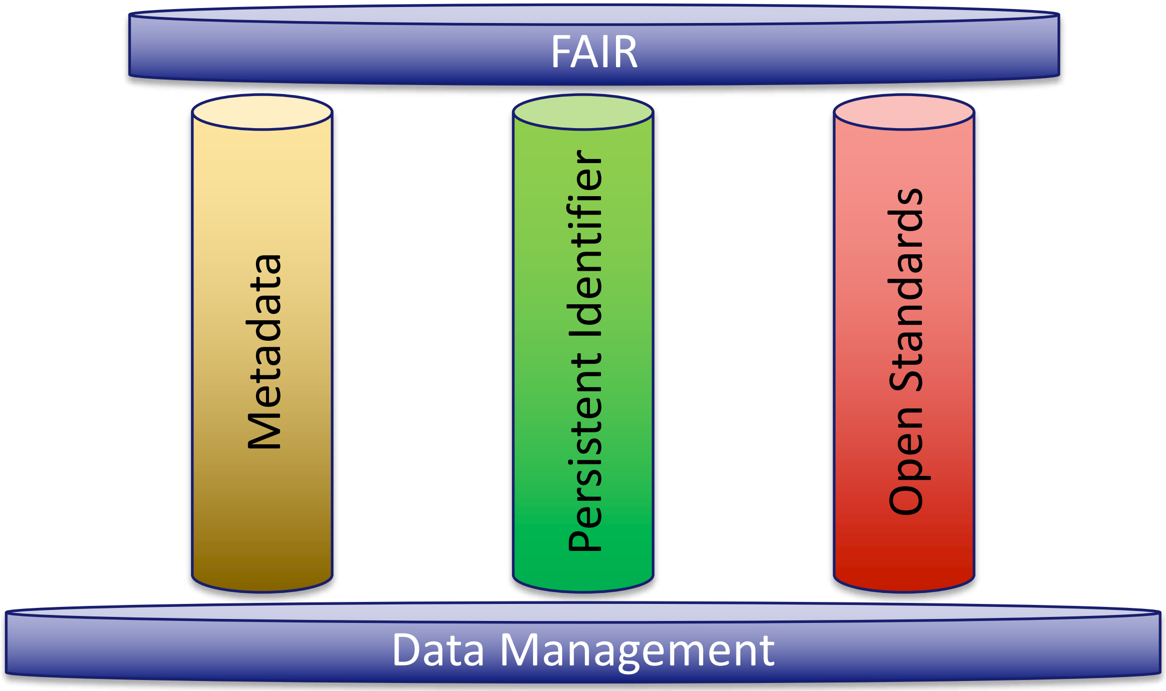


Findable	Persistent Identifiers (PIDs)	Rich metadata	Indexed data repositories	PIDs in metadata
Accessible	Standard communications protocol	Open, free protocol	Authentication, where necessary	Metadata is always available
Interoperable	Vocabularies	Vocabularies are FAIR	Linked metadata	
Reusable	Metadata have multiple attributes	Usage license	Provenance	Community standards

B2ACCESS
B2STAGE
B2HANDLE
Register your Research Data
B2FIND
B2SHARE
B2NOTE
Collaborative EUDAT Data Infrastructure
Training

The Magnifying glass, Tap, Gears set, Recycle sig, Storage, Infinity, Discussion, Shield, and Man User icons made by Freepik from www.flaticon.com are licensed by CC 3.0 BY. All other icons made by ARDC. Entire FAIR resources graphic is licensed under a Creative Commons Attribution 4.0 International License

The Pillars of FAIR



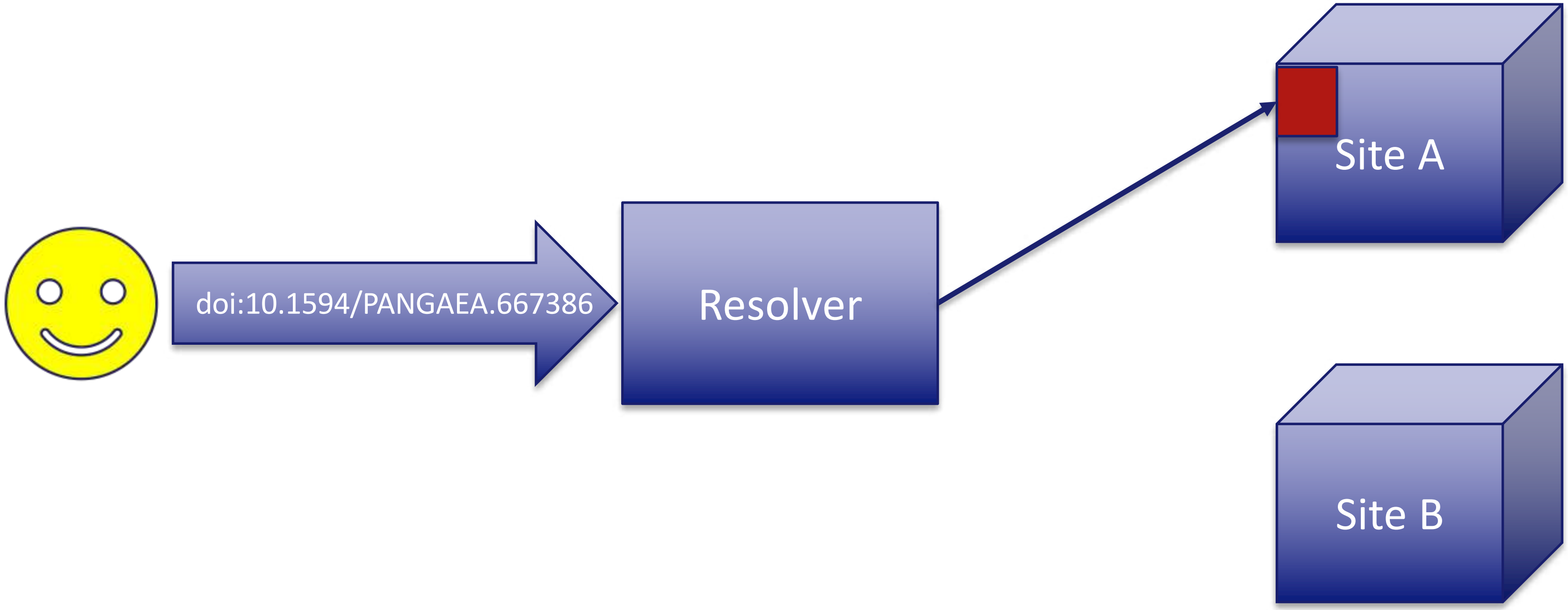
Four Horsemen of Metadata

Image by Emma Honeybun, CC BY-NC-SA 3.0

- **Descriptive** Metadata
  - Information about the data
  - Aimed at both expert and non expert users
  - Some Common Standards exist (e.g. Dublin Core, DataCite, DIF, METS)
- **Community** Metadata
  - Many hundreds of community standards (http://www.dcc.ac.uk/resources/metadata-standards/list)
- **Structural** Metadata
  - How to read the data, how the data is organized within the file
  - Some formats (e.g netcdf, HDF) can be self describing
- **Administrative** Metadata
  - Copyright and licensing information

Persistent Identifiers

- Resolvable links independent of data locations
- Common technologies:
  - DOI, PURL, URNs, ARKs,…



Preservation - Planning

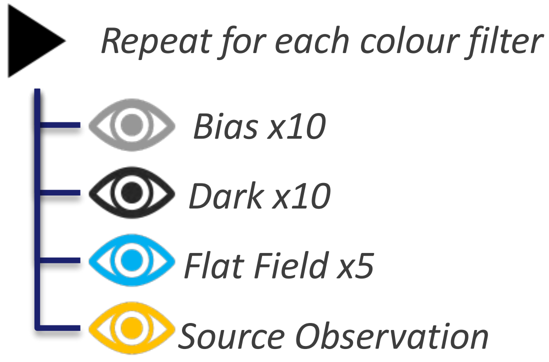
- Key is PLANNING – Writing a Data Management Plan
- Online tools:
  - easyDMP (https://easydmp.eudat.eu/)
  - DMPTool (https://dmptool.org/)
  - DMPOnline (https://dmponline.dcc.ac.uk/)
  - Data Stewardship Wizard (https://ds-wizard.org/)
- And various templates

DMP – The Essentials for H2020

- Within Horizon2020, the following should be addressed
  - Data Summary
    - Purpose, formats, reusing data?, source, volume, reusability, processing
  - FAIRness
    - Metadata standard, persistent identifier, what data will be shared, how will it be made accessible, methods of access, embargo periods, duration, licensing…
  - Costs
    - How much, how long and who is responsible
  - Security
    - Is the data sensistive, what security measures need to be in place
  - Ethics
    - ethical or legal issues impacting data sharing, informed consent
  - Other issues
    - Use of other national/funder/sectorial/departmental procedures
- Lets look at a fictitious use case….but based on reality

Data Management Planning Scenario

- An astronomer is going to use the new FLT (Fascinatingly Large Telescope) to observe M51. We want to take images in three different colour bands to investigate the presence of LGM (Little Green Men) in the outer spiral arms. This images will be taken with the Narrow Field TLA device, a high resolution CCD sensor. A typical observing sequence might look like:



Repeat for each colour filter

Bias x10

Dark x10

Flat Field x5

Source Observation

- See https://easydmp.eudat.eu/plan/809/

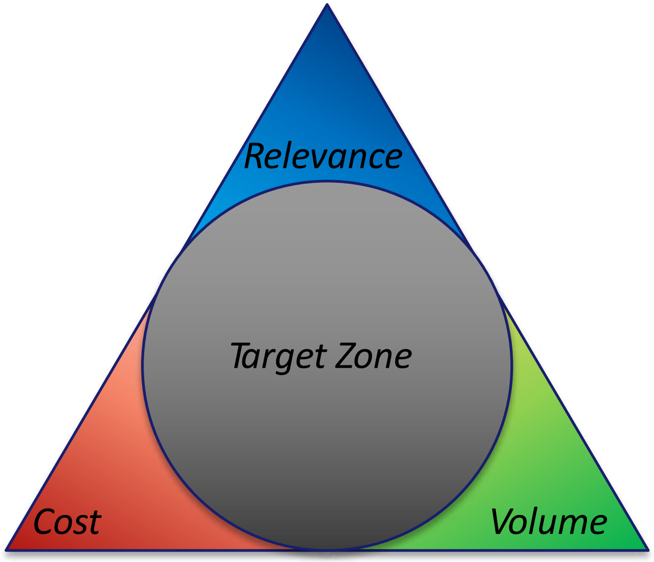
Raw Image *Processed Image*

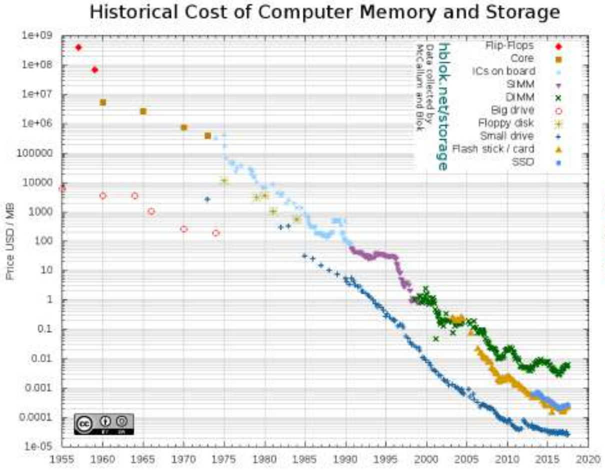
EXERCISE – Write Your Own DMP

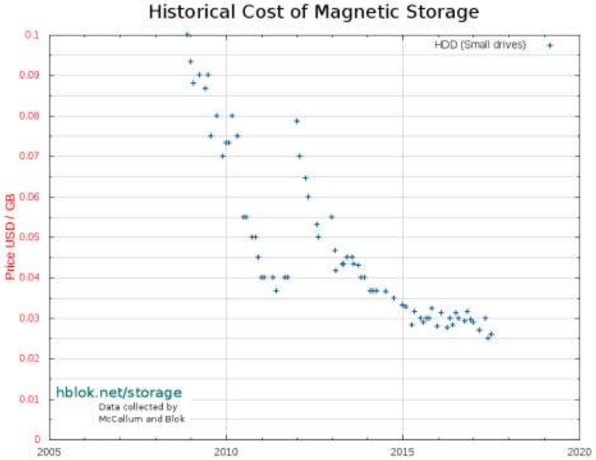
- Using easy.DMP (https://easydmp.eudat.eu/)
- Actions
  - Log in using B2ACCESS
  - Select simplified H2020 Template
  - You can either create a DMP **based on your own research** or use the following scenario: *”You need to write a data management plan for statistical analysis of census information from the Eurostat (*https://ec.europa.eu/eurostat/web/population-and-housing-census/census-data/database*) and the US Census Bureau (*https://www.census.gov/data.html*) to look at the difference in housing stock and population age distributions between the EI and US. The final output should be a report including graphics, however the graphics themselves should be accessible. You data should be deposited in an open depository such as B2SHARE and both the paper and any associated graphs should be be accessible using a doi”*
  - *When finished, you should be able to validate your plan…*
  - *Finally – share your data with me (*hungmung@gmail.com*)*
  - *We will select one at random to analyse…*

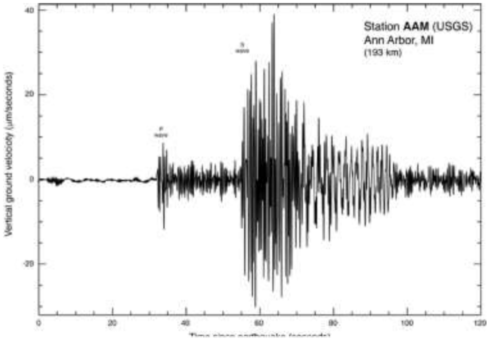
The Hard Part – Selecting What Data to Preserve

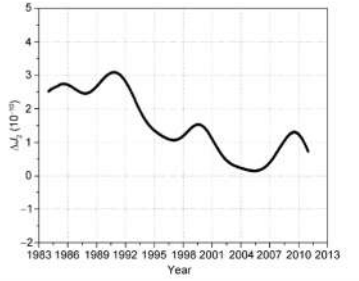
- *The more data you preserve, the more the cost, but the more reproducible and reusable it is*
- *The longer the data is relevant, the more the cost, but if someone has reused their data, their results may not be reproducible/reusable*
- *But each of these is hard to estimate….*







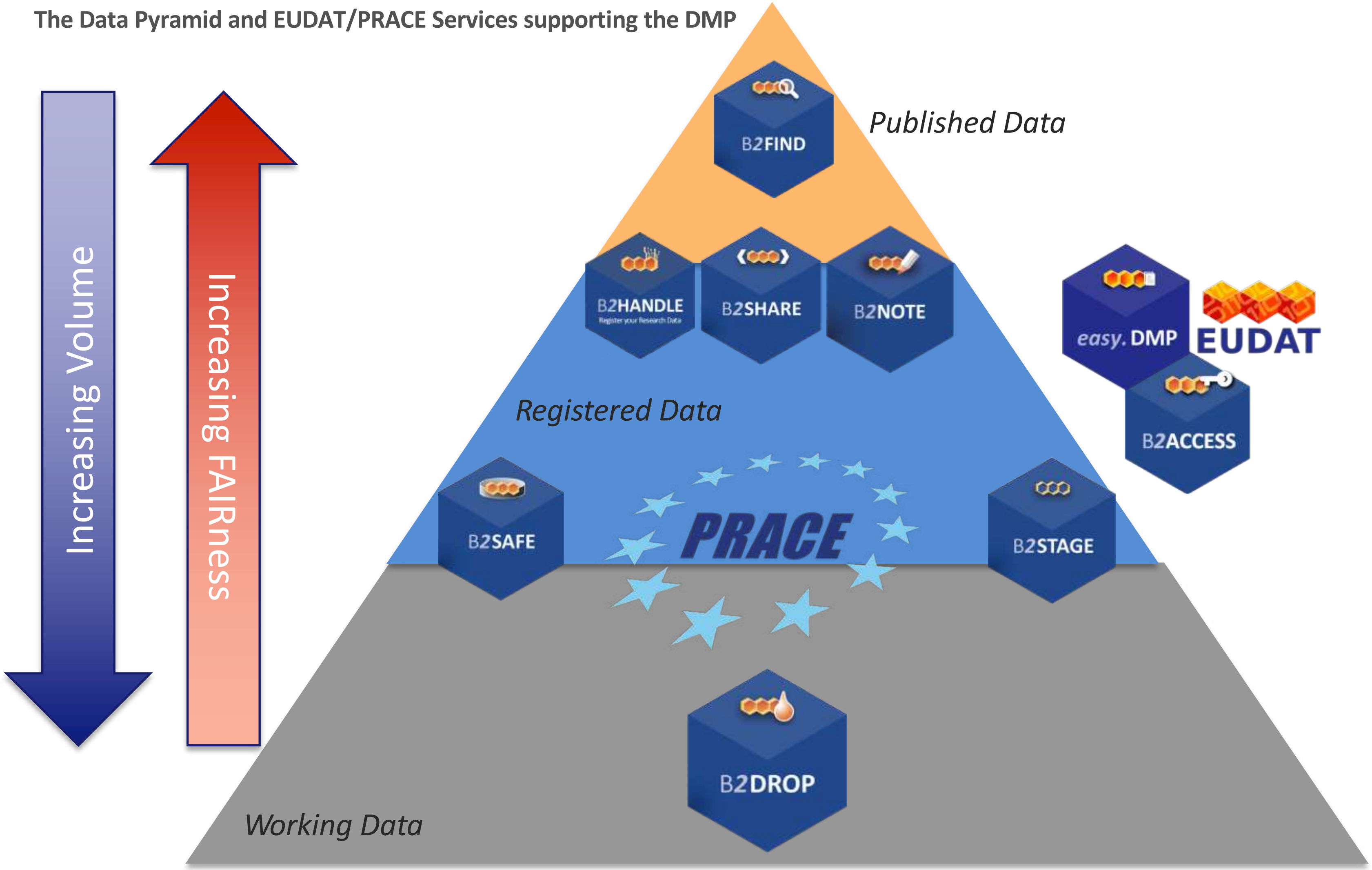




EUDAT Collaborative Data Infrastructure

PRACE

The Data Pyramid and EUDAT/PRACE Services supporting the DMP



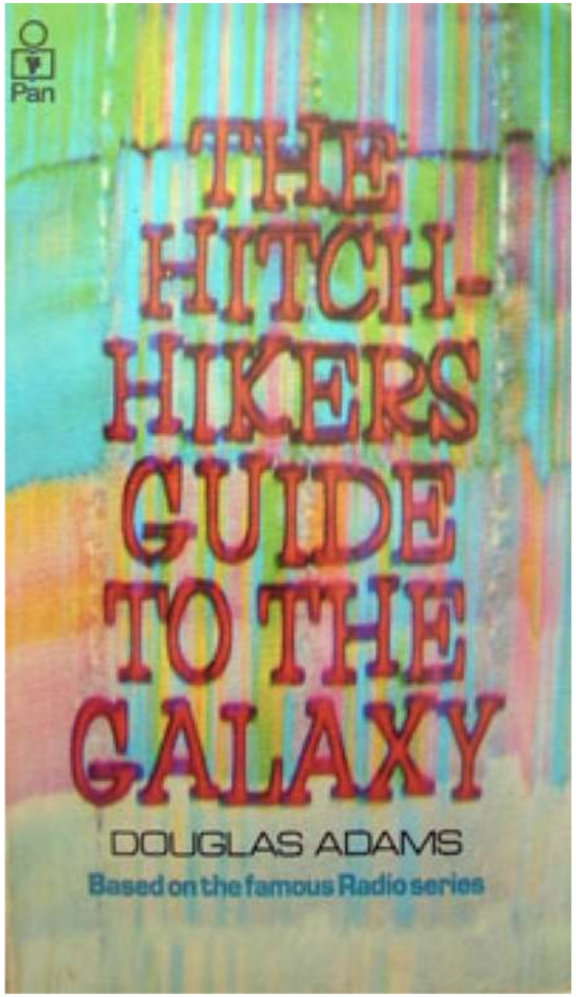
Summary

- There are a lot of reasons why you should preserve (or curate) your data
- Consider data preservation as long term
    - UK Research Councils require 10 years from date of last access – so it could be **very** long
- Always try and be FAIR
    - Consider **granularity** of access
    - Have a well defined **metadata schema**
    - Give your data a **Persistent Identifier**
        - **Cite** your data to your publications
    - Use well defined and common **formats**
        - Standard or community specific
    - Give enough **provenance** information to allow other users to trust your data
    - Put your data into a **trusted repository**
- Think carefully about what data you want to keep…

And Finally – A Hitchhiker Guide…

Mr. Prosser: **“But the data were accessible…”**
Arthur Dent: **“Accessible? I eventually had to go down to the cellar to find them.”**
Mr Prosser: **“That’s the repository.”**
Arthur Dent: **“With a flashlight.”**
Mr Prosser**: “Ah, well, the lights had probably gone.”**
Arthur Dent: **“So had the stairs.”**
Mr. Prosser: **“But look, you found the data, didn’t you?”**
Arthur Dent: “**Yes, yes I did. It was on on a 5 ¼ inch floppy disk in the bottom of a locked filing cabinet stuck in a disused lavatory with a sign on the door saying ‘Beware of the Leopard.’”**

(with Apologies to Douglas Adams)


Pan
THE HITCH-HIKERS GUIDE TO THE GALAXY
DOUGLAS ADAMS
Based on the famous Radio series