Marjan Grootveld, DANS

@MarjanGrootveld
@openaire_eu

Research data management and data sharing

“Open Access in practice”, January 12, 2018

Where do we go?

- Why manage your data?
- EC’s Open Research Data Policy
- Planning FAIR data management
- Concerns about data sharing
- References

Data Archiving and Networked Services

DANS

https://dans.knaw.nl/en

Institute of Dutch Academy and Research Funding Organisation (KNAW & NWO) since 2005

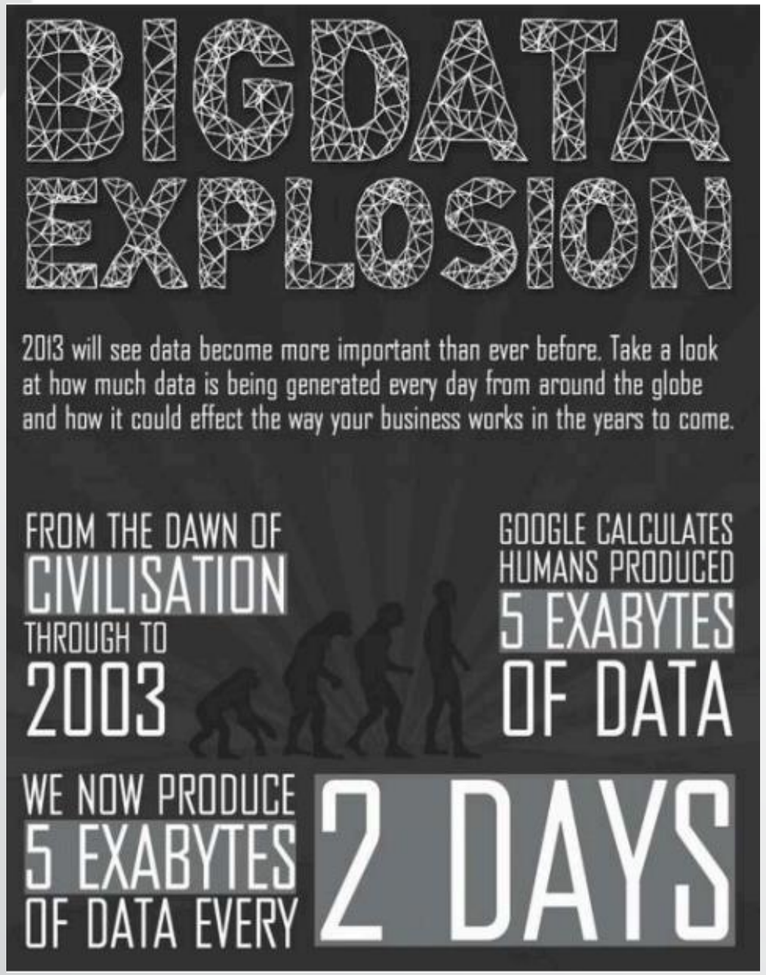
Mission: promote and provide permanent access to digital research information

First predecessor dates back to 1964 (Steinmetz Foundation), Historical Data Archive 1989

Why manage data?

Image CC-BY-NC-SA by Leo Reynolds www.flickr.com/photos/lwr/13442910354

Data explosion



- More and more data is being created
- Issue is not creating data, but being able to navigate and use it
- Data management is critical to make sure data are well-organised, understandable and reusable

Prevent data loss

Digital data are fragile and susceptible to loss for a wide variety of reasons

- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- Format obsolescence
- Human error
- Malicious attack
- Loss of staffing competencies
- Loss of institutional commitment
- Loss of financial stability
- Changes in user expectations

For your own sake

- Make your research easier
- Stop yourself drowning in irrelevant stuff
- Save data for later
- Avoid accusations of fraud or sloppy science
- Write a data paper, connect your nano publications
- Share your data for re-use & get them validated in real life
- Get credit for it

The data sharing advantage in astrophysics

S. B. F. Dorch, T. M. Drachen, O. Ellegaard

(Submitted on 8 Nov 2015)

We present here evidence for the existence of a citation advantage within astrophysics for papers that link to data. Using simple measures based on publication data from NASA Astrophysics Data System we find a citation advantage for papers with links to data receiving on the average significantly more citations per paper than papers without links to data. Furthermore, using INSPEC and Web of Science databases we investigate whether either papers of an experimental or theoretical nature display different citation behavior.

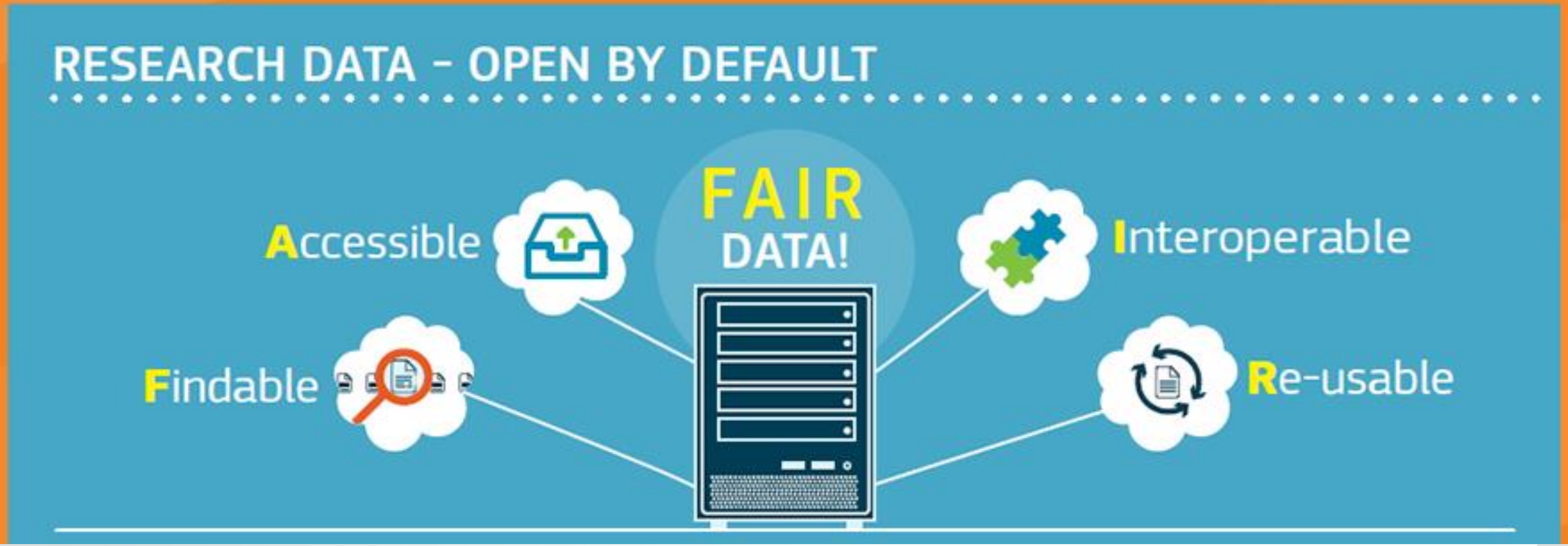
Comments: 4 pages, 2 figures, Conference proceedings of Focus Meeting 3 on Scholarly Publication in Astronomy, IAU GA 2015, Honolulu

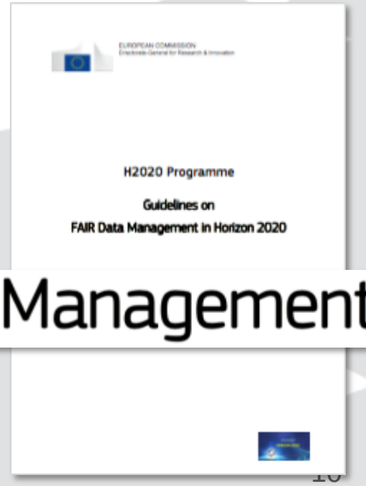
Subjects: **Instrumentation and Methods for Astrophysics (astro-ph.IM)**; Digital Libraries (cs.DL)

Cite as: arXiv:1511.02512 [astro-ph.IM]

(or arXiv:1511.02512v1 [astro-ph.IM] for this version)

EC’S OPEN RESEARCH DATA PILOT





FAIR Data Management

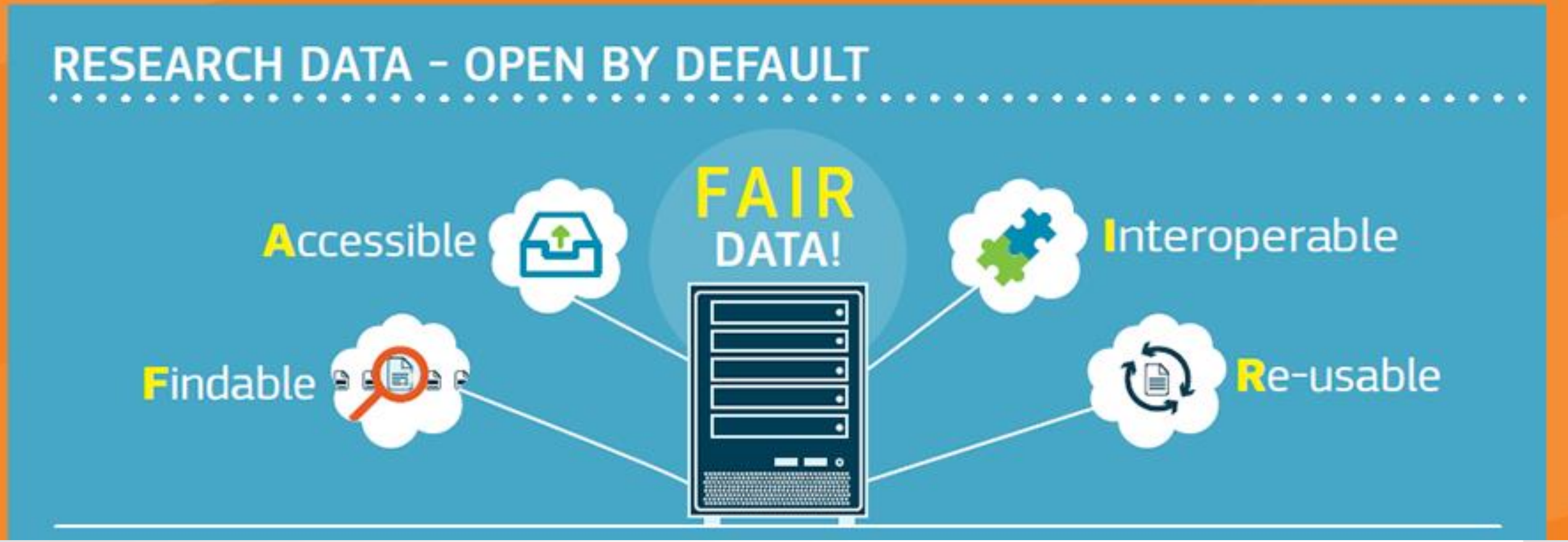
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Clarifying terminology...

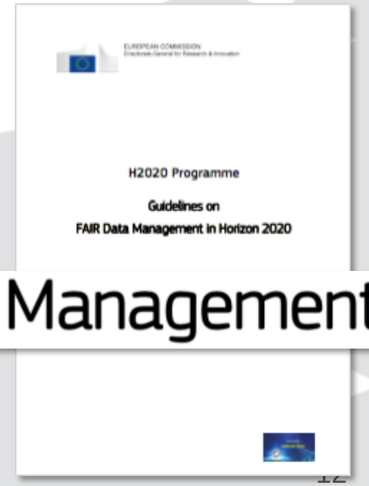
In the past our policy mainly addressed the 'accessibility' part of FAIR.

- Started off with 'open access to research data'
- Moved towards open (research) data with the ORD pilot (which also covered further aspects)
- We are now seeing openness as one component of FAIR data and aim to address all of the FAIR aspects in Horizon 2020

Source: Daniel Spichtinger, European Commission DG RTD, Unit A.6. – October 11, 2017



EC in the Guidelines: “This template is not intended as a strict technical implementation of the FAIR principles, it is rather inspired by FAIR as a general concept (…) without suggesting any specific technology, standard, or implementation solution”



FAIR Data Management

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

PLANNING
DATA MANAGEMENT

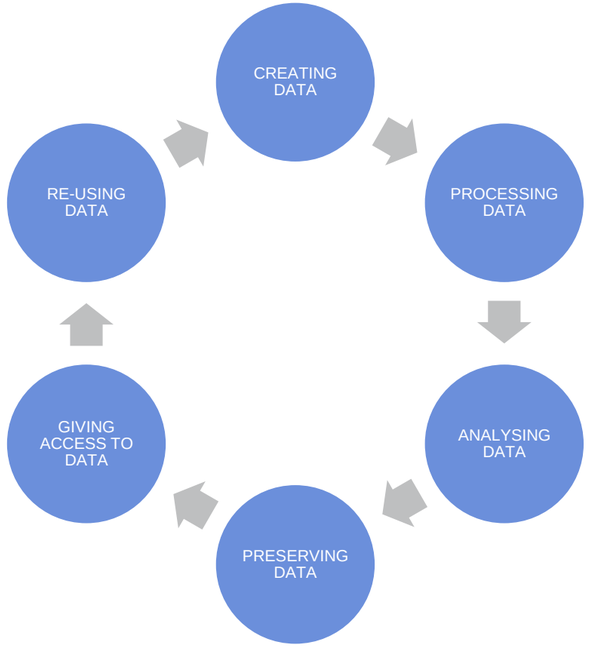
Examples of research data

- Text or Word documents, spreadsheets
- Statistics
- Results of experiments
- Measurements
- Observations resulting from fieldwork
- Survey results
- Interview recordings: audiotapes, videotapes
- Images
- Laboratory notebooks
- Database contents
- Models, algorithms, scripts

Data Management Plans

A DMP is a brief plan to define:

- how the data will be created or re-used
- how it will be documented
- who can access it
- where it will be stored
- whether it will be shared
- where it will be preserved

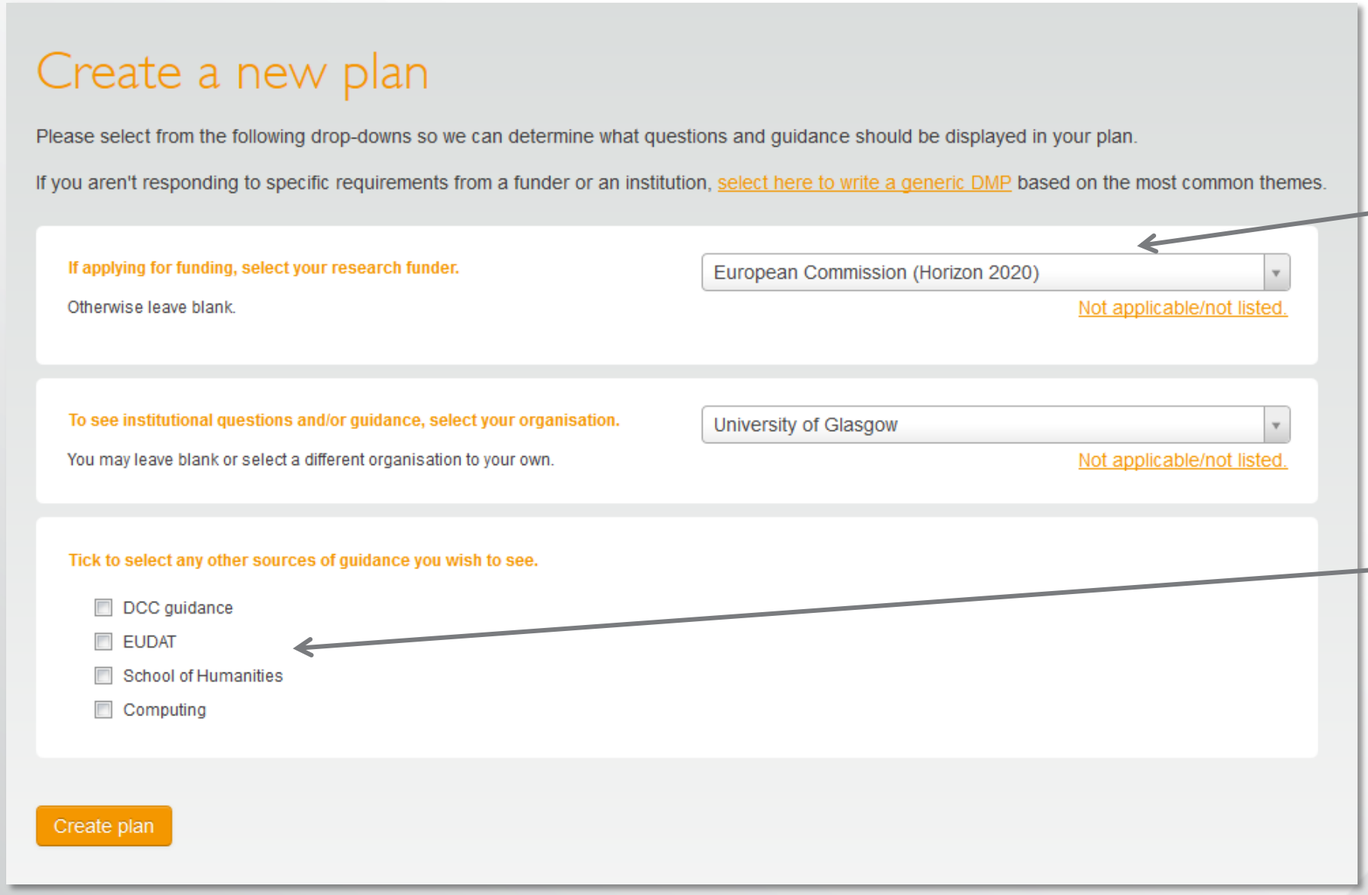


DMPs are sometimes submitted as part of grant applications, sometimes afterwards, but they are useful whenever researchers are creating data.

DMPonline

A web-based tool to help researchers write DMPs

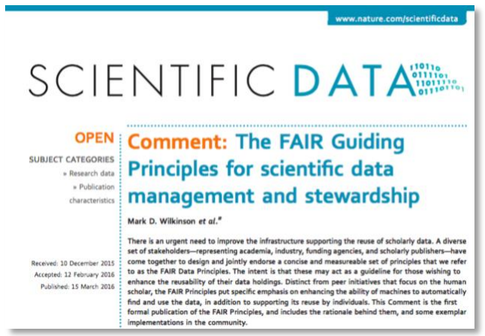
https://dmponline.dcc.ac.uk



Choose your funder to get their specific template

Choose any additional optional guidance

Making data FAIR



- **Findable**
  - Assign persistent IDs, provide rich metadata, register in a searchable resource, ...
- **Accessible**
  - Retrievable by their ID using a standard protocol, metadata remain accessible even if data aren't...
- **Interoperable**
  - Use formal, broadly applicable languages, use standard vocabularies, qualified references...
- **Reusable**
  - Rich, accurate metadata, clear licences, provenance, use of community standards...

www.force11.org/group/fairgroup/fairprinciples and http://www.nature.com/articles/sdata201618

Some “F” questions

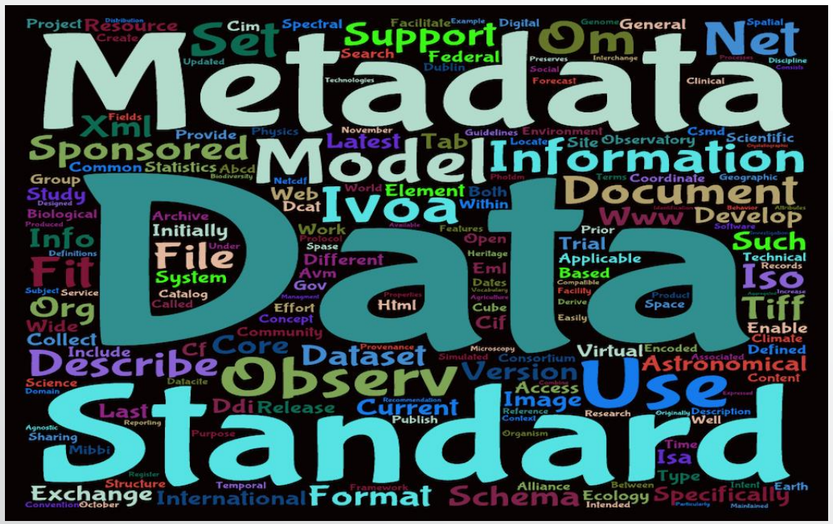
§2.1 Making data findable, including provisions for metadata

- Use metadata and specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how.
- Use search keywords
- Persistent and unique identifiers such as DOI
- File and folder naming conventions: see OpenAIRE-EUDAT July 2016 webinar
- Versioning of the datasets and clear version numbers

Metadata standards

Metadata Standards Directory

- Broad, disciplinary listing of standards and tools
- Maintained by RDA group



http://rd-alliance.github.io/ metadata-directory

Biosharing

- A portal of data standards, databases, and policies
- Focused on life, environmental and biomedical sciences

https://biosharing.org

Documentation?

- Code book explaining the variables
- Study design
- Lab journal
- iPython or Jupyter notebook
- Statistical queries
- Software or instruments to understand or reproduce the
- Machine configurations
- Consent information
- Data usage licence
- …

In short: document and preserve everything that is needed to reproduce the study – ideally following the standard in your discipline

Some “A” questions

§ 2.2 Making data openly accessible:

- Explain which data can’t be shared openly, if any
- Specify how access will be provided in case of restrictions, e.g. through a data committee, a license, or arranged with the repository.
- Will methods or software tools needed to access the data (if any) be included or documented?
- Deposit the data and associated metadata, documentation and code preferably in certified repositories which support Open Access.

Core Trust Seal
Data Seal of Approval
ICSU World Data System
nestor seal
ISO 16363

Where to find a repository?

More information: https://www.openaire.eu/opendatapilot-repository

Zenodo: http://www.zenodo.org Re3data.org: http://www.re3data.org

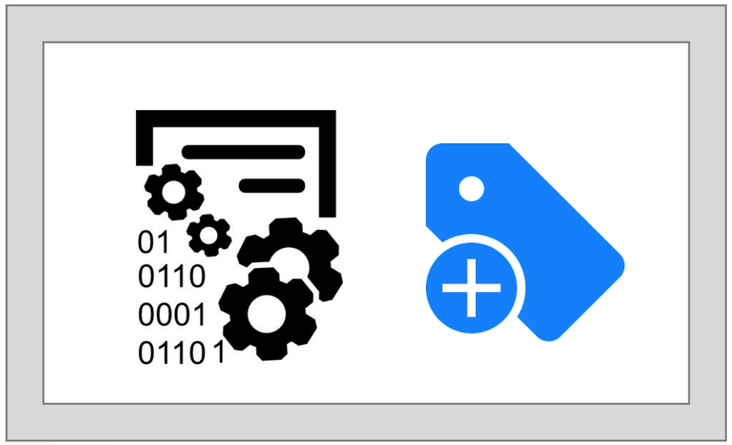
Storing or archiving? Both!

Storing and backing up files while research is active

Likely to be on a networked filestore or hard drive

Easy to change or delete

Archiving or preserving data in the long-term



Likely to be deposited in a digital repository

Safeguarded and preserved

Before clocks were invented, people kept time using different instruments to observe the Sun's zenith at noon. Towns and cities set clocks based on sunsets and sunrises. Time calculation became a serious problem for people travelling by train, sometimes hundreds of miles in a day. UTC is the **World's Time Standard**.

Interoperable

In the aftermath of the French Revolution (1789), the traditional units of measure used in the Ancien Régime were replaced. The livre monetary unit was replaced by the decimal franc, and a new unit of length was introduced which became known as the metre. **The metre gained adoption in continental Europe** during the mid nineteenth century, particularly in scientific usage, and was officially established as an international measurement unit by the Metre Convention of 1875.

A440, which has a frequency of 440 Hz, is the musica above middle C and serves as a **general tuning stand musical pitch**. Prior to the standardization on 440 Hz, countries and organizations followed the Austrian government's 1885 recommendation of 435 Hz. In the period instrument movement, a **consensus** has arisen around a modern *baroque pitch* of 415 Hz (A♭ of A440), baroque for some special church music (*Chorton pitch*) at 466 Hz (A♯ of A440), and *classical pitch* at 430 Hz.

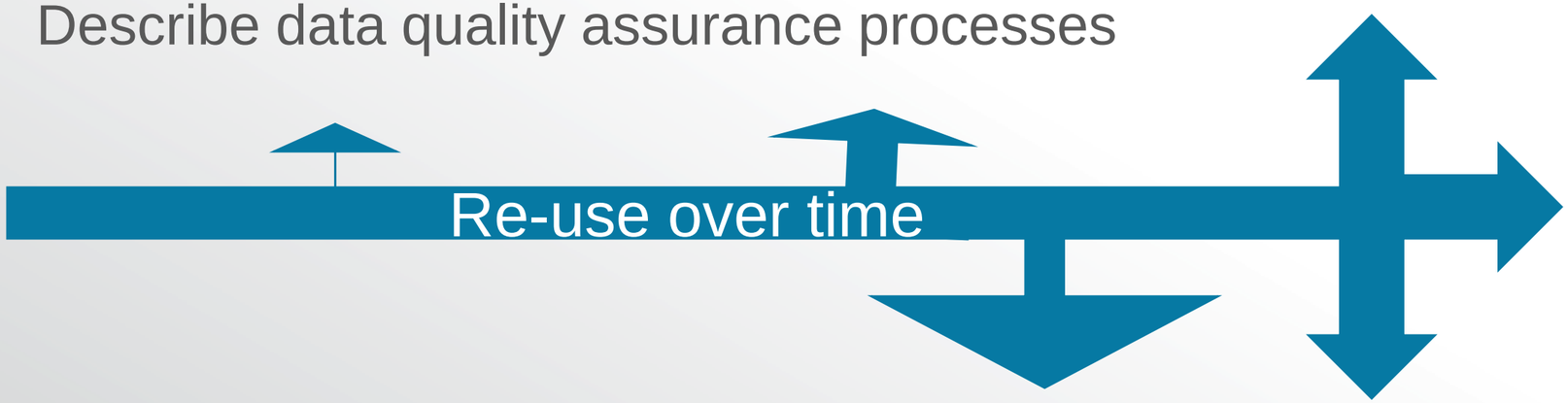
Medical classification is the process of transforming descriptions of medical diagnoses and procedures into universal medical code numbers. SNOMED Clinical Terms (SNOMED CT) is intended to provid a set of concepts and relationships that offers a **common reference point for comparison and aggregation of data about the health care process**. SNOMED-CT is designed to be managed by computer.

European Commission

Some “R” questions

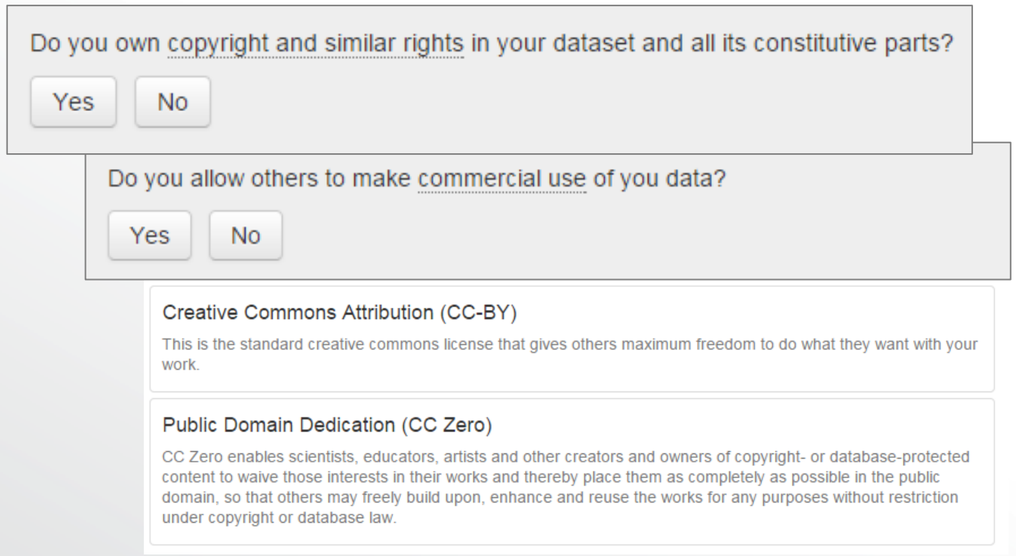
§ 2.4 Increase data re-use (through clarifying licences)

- License the data to permit the widest reuse possible
- Specify a data embargo, if this is needed
- How long will the data remain reusable?
- Describe data quality assurance processes



Licensing research data and software

EUDAT licensing wizard helps you pick licences for data & software



Horizon 2020 Open Access guidelines point to:

or

You should also license Open Access data, or waive rights.

http://ufal.github.io/public-license-selector/

Sharing data: what is meant?

With collaborators while research is active

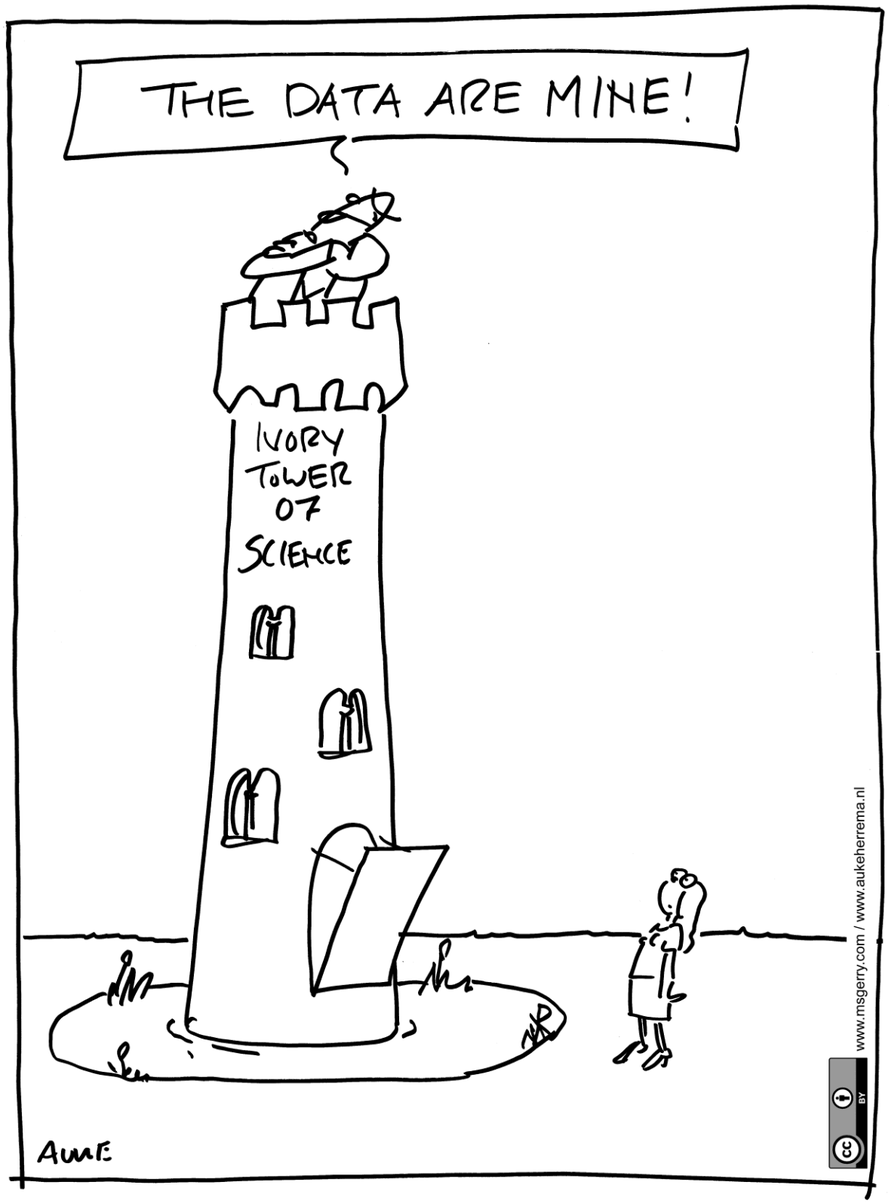
Data are mutable

Dataverse

Dropbox

(Open) data sharing

Data are stable, searchable, citable, clearly licensed

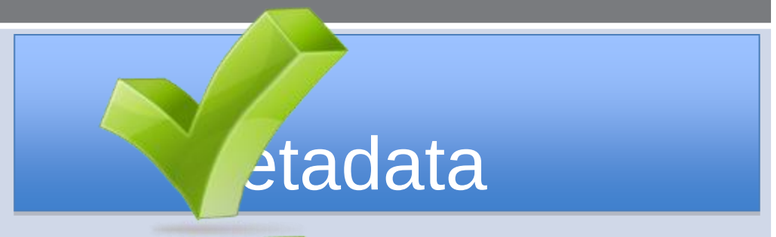


SCENE FROM THE PAST?

Concerns about data sharing

Concern	Solution
inappropriate use due to misunderstanding of research purpose or parameters	
security and confidentiality of sensitive data	
lack of acknowledgement / credit	

Concerns about data sharing

Concern	Solution
inappropriate use due to misunderstanding of research purpose or parameters	
security and confidentiality of sensitive data	
lack of acknowledgement / credit	

Concerns about data sharing

Concern	Solution
inappropriate use due to misunderstanding of research purpose or parameters	provide rich *Abstract, Purpose, Use Constraints* and *Supplemental Information* where needed
security and confidentiality of sensitive data	• the metadata does NOT contain the data • *Use Constraints* specify who may access the data and how
lack of acknowledgement / credit	specify an obligatory data citation within the *Use Constraints* and the licence

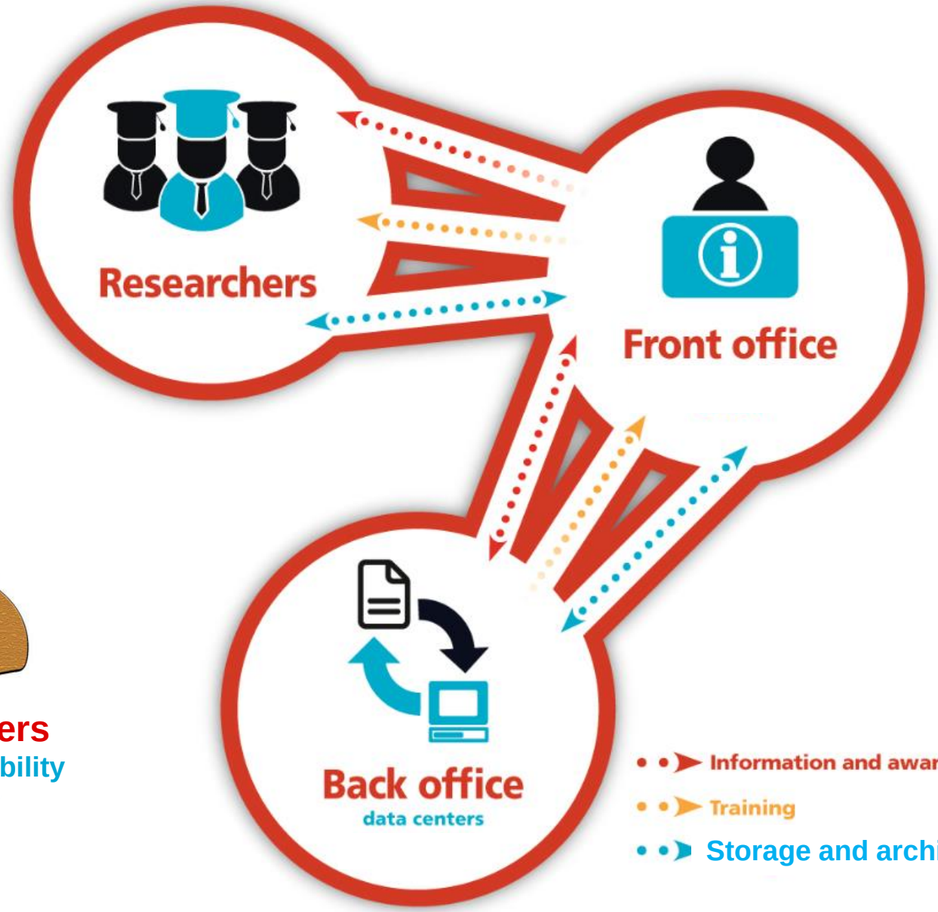
Overwhelmed?

A DMP is also a communication instrument!

RDM stakeholders

Commercial partners

Publishers
Data Availability
Policy



Institution
RDM policy
Facilities

MALTA

KEVIN ELLUL

Research funders

Data Management in H2020 - summary

- Research data should be as open as possible, as closed as necessary.
- Manage and document all data FAIRly, whether they will be open or not.
- A Data Management Plan (DMP) is due by month 6. It is a regular project deliverable.
- A DMP is a living document: to be used, updated and shared. You can use the Horizon 2020 template in DMPonline.
- Deposit the data* in a research data repository for sharing and preserving the data long term.
- "Sharing" means "outside the project consortium".

* at least the data underlying publications, but ideally all data that have value

The EC Open Research Data policy

Key sources of information

- Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
- Guidelines on FAIR Data Management in Horizon 2020 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- Annotated model grant agreement, clause 29.3 http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_en.pdf
- Multi-beneficiary General Model Grant Agreement, changes in clause 29.3 and references to research integrity ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_en.pdf
- Infographic summarising key policy points http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf

OpenAIRE support materials

https://www.openaire.eu/what-is-the-open-research-data-pilot

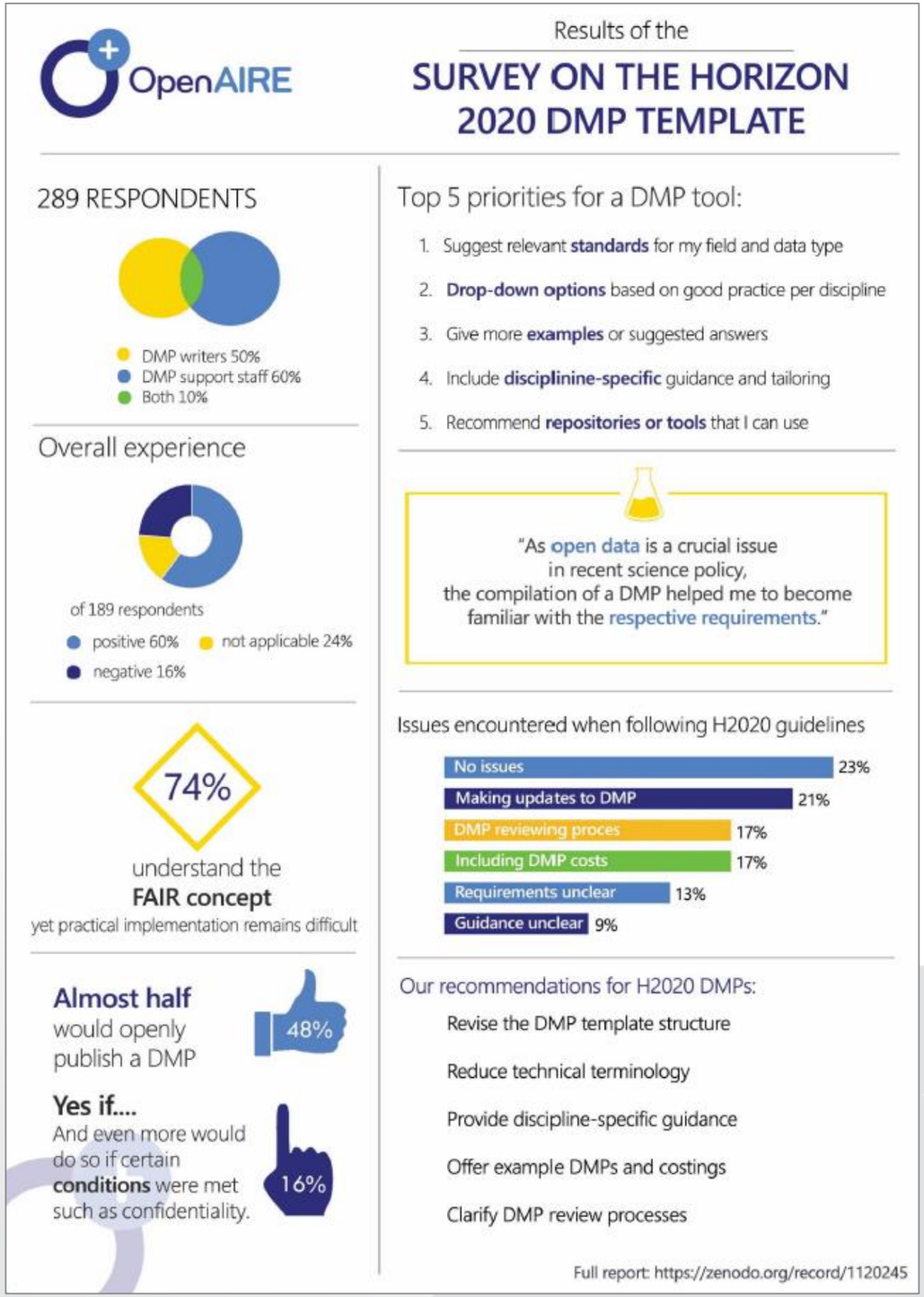
https://www.openaire.eu/support

- Briefing papers, factsheets, webinars, workshops, FAQs

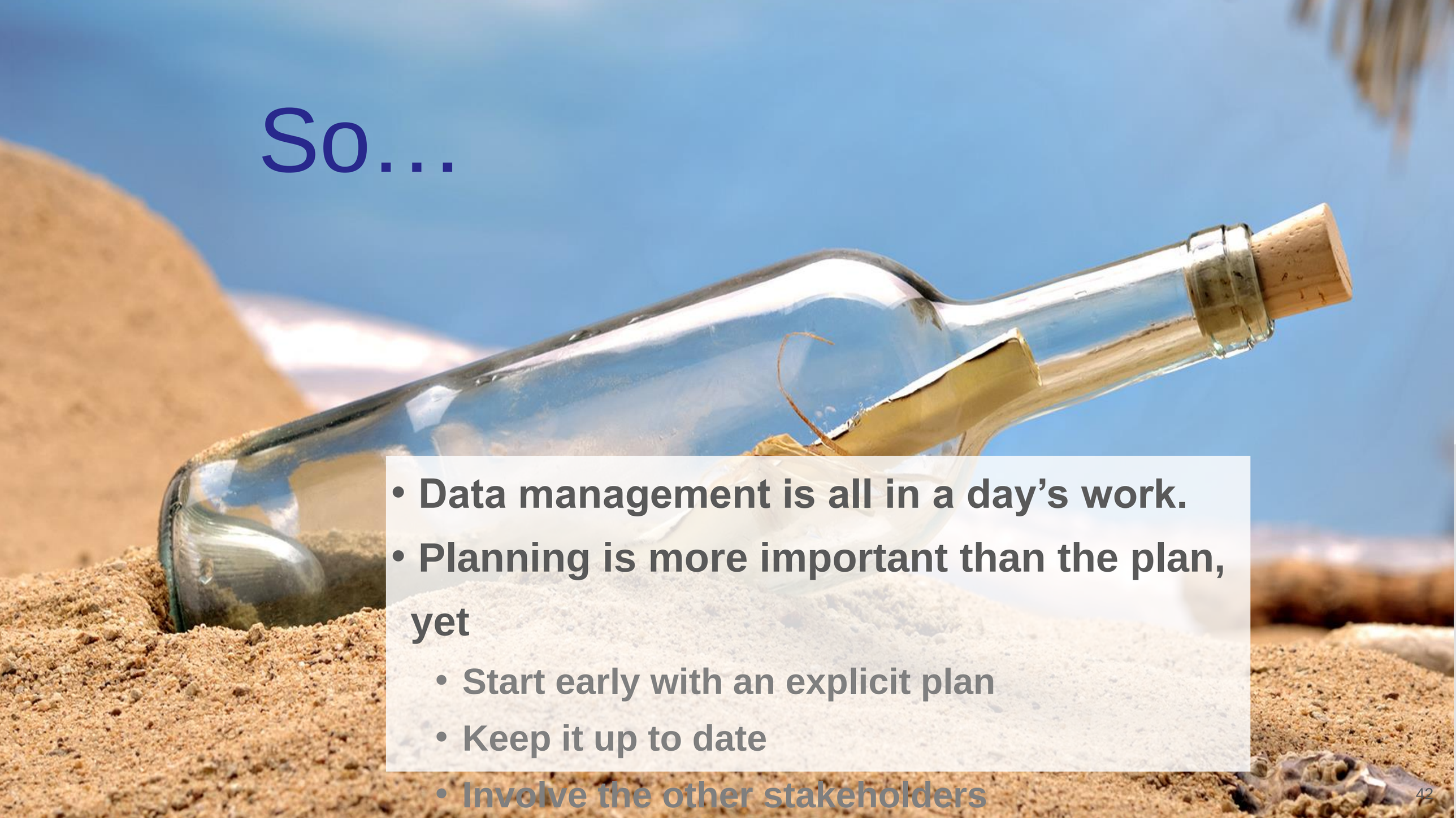
- Information on:
  - Open Research Data Policy
  - Creating a data management plan
  - Selecting a data repository
  - Personal data

Survey

- OpenAIRE and FAIR Data Expert Group ran a survey about the H2020 DMP template
- Findings and recommendations available via: https://zenodo.org/record/1120245



So…



- **Data management is all in a day's work.**
- **Planning is more important than the plan, yet**
  - **Start early with an explicit plan**
  - **Keep it up to date**
  - **Involve the other stakeholders**

Questions?

www.openaire.eu

@openaire_eu

facebook.com/groups/openaire

linkedin.com/groups/OpenAIRE-3893548

marjan.grootveld@dans.knaw.nl