Data Archiving and Networked Services

DANS

Data management planning: AMID PROFS

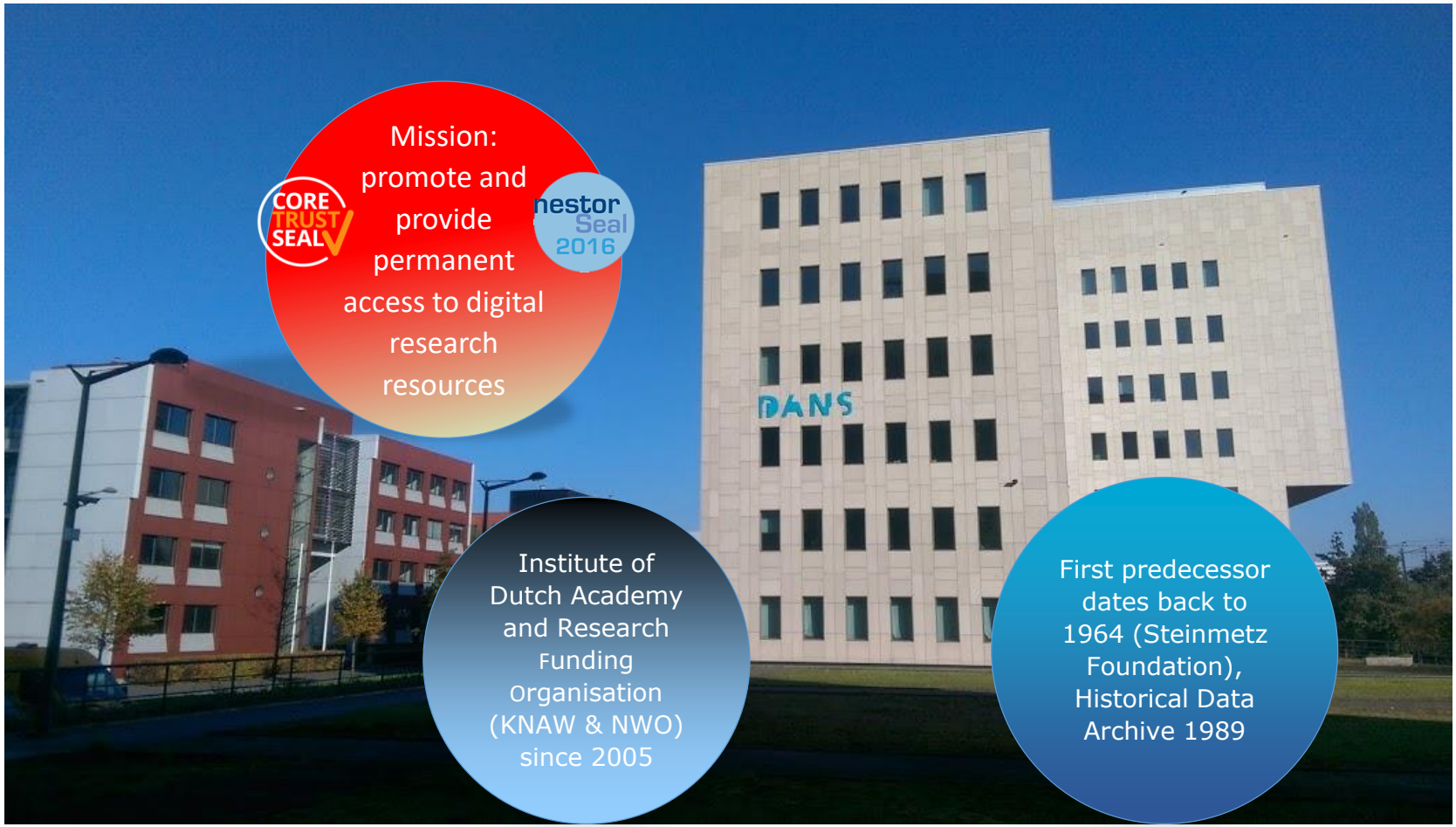
Marjan Grootveld

February 6, 2019

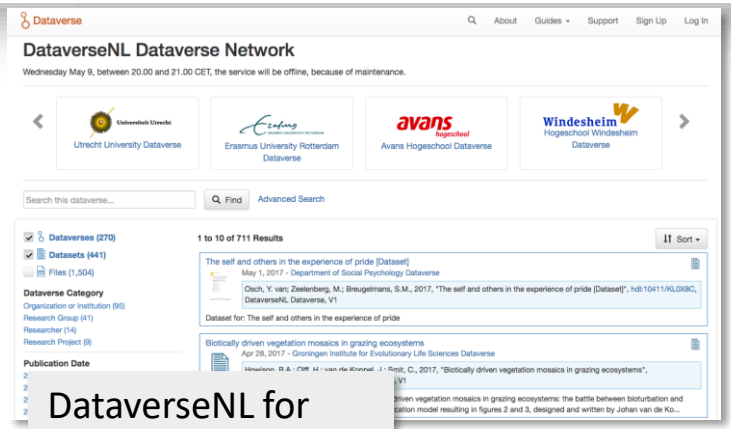
1. What is DANS?
2. What is data?
3. All data want to be cared for
4. ~~Exercise: reviewing DMP text samples~~
5. What a DMP should cover
6. Tools for writing a DMP
7. Exercise: assessing the FAIRness of existing data

https://pixabay.com/en/number-wood-7-spring-the-vine-2893866/

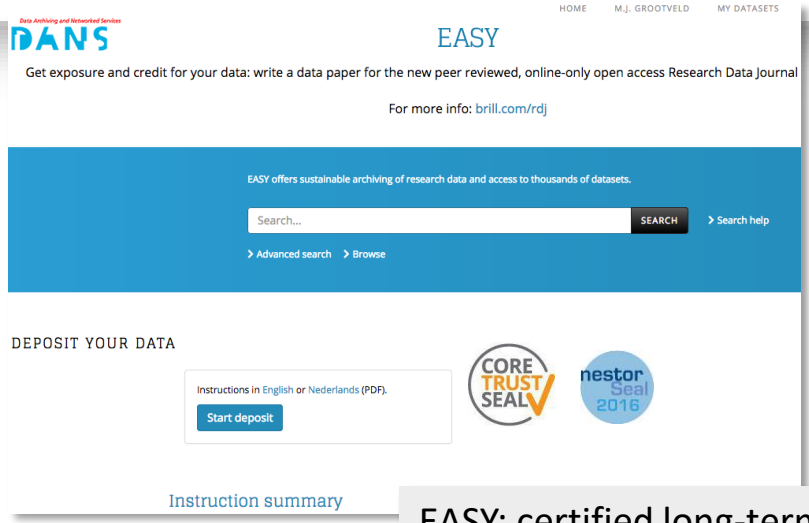
DANS https://dans.knaw.nl/nl



Core DANS services



DataverseNL for short- and mid-term data storage

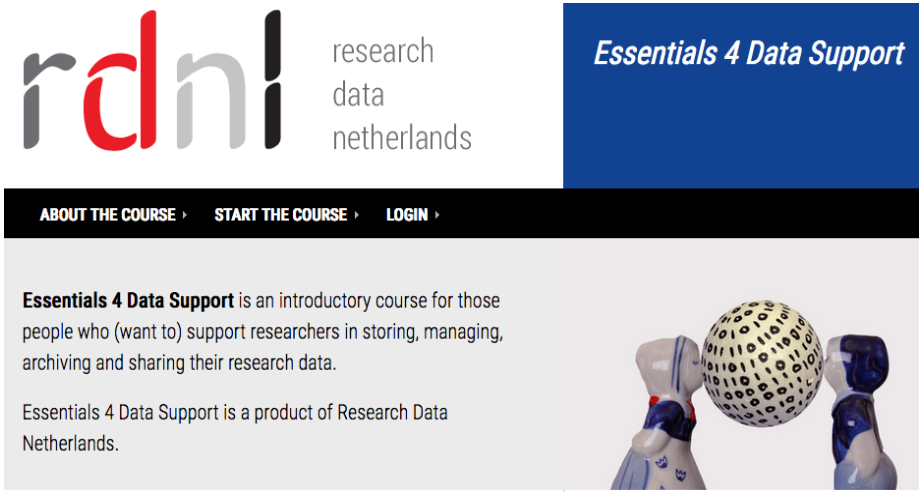


EASY: certified long-term Electronic Archiving System for self-deposit

NARCIS: Gateway to scholarly information in the Netherlands

Data management support training

DANS gives RDM training in projects and in Research Data Netherlands collaboration



https://datasupport.researchdata.nl/en/

EUDAT

EOSCpilot
The European Open Science Cloud for Research Pilot Project

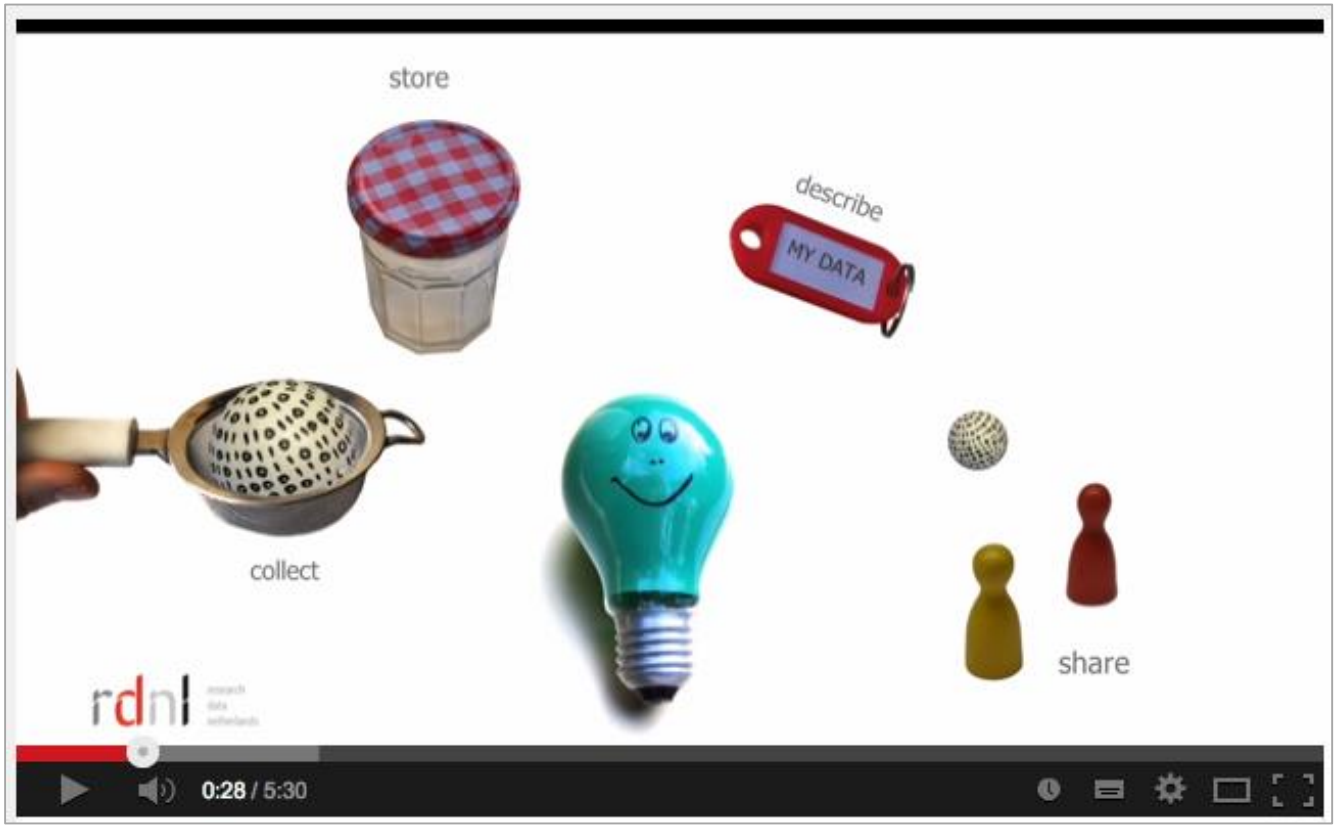
EOSC-hub

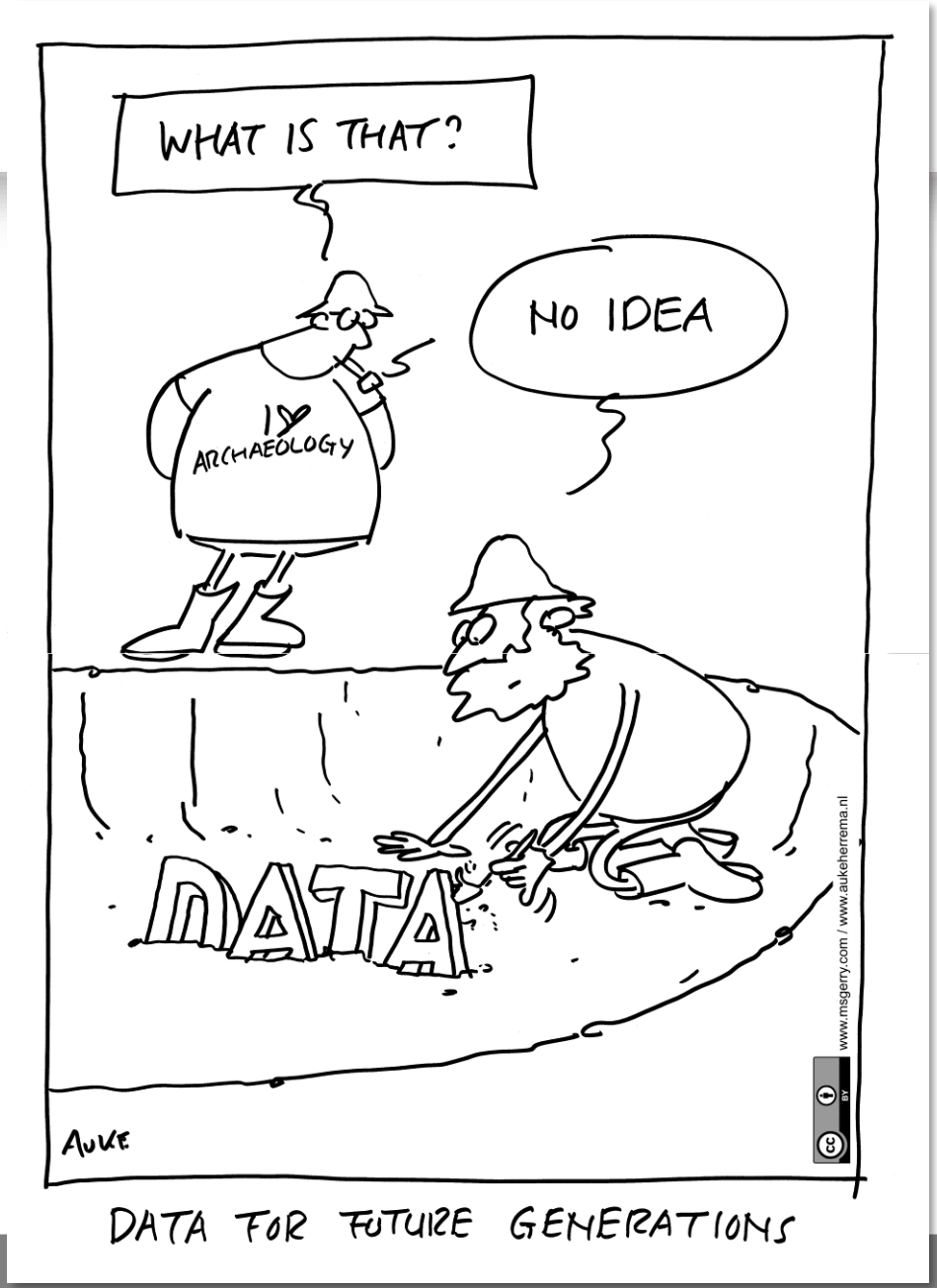
OpenAIRE ADVANCE

https://eudat.eu/
https://eoscpilot.eu/
https://www.eosc-hub.eu/
https://www.openaire.eu/

The what, why and how of data management planning

From the training *Essentials 4 Data Support*

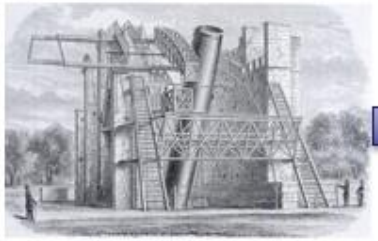


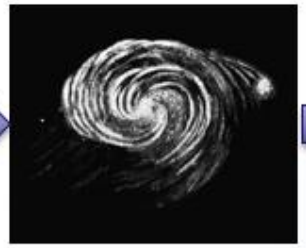


DATA FOR FUTURE GENERATIONS

What is Data, and how is it managed

Instrumentation





Then

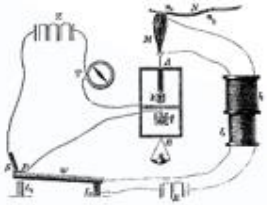
Observation

Results & Data

Ralf Roletschek CC-BY-3.0

Results & Data

Experimentation

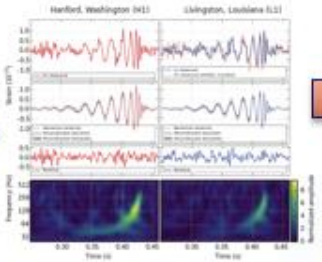


Slide credits: Shaun de Witt (CCFE) for EOSC-hub

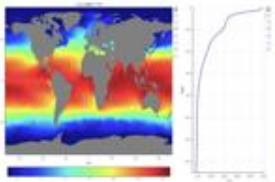
What is Data, and how is it managed

Instrumentation

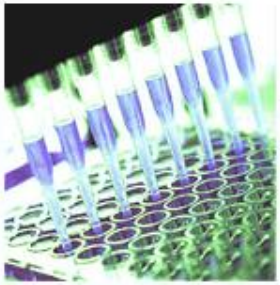
The Virgo collaboration/CCO 1.0

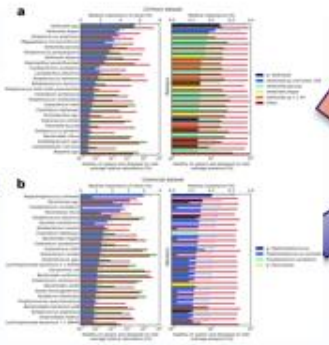


Observation



Experimentation





Results

Data

Ralf Roletschek CC-BY-3.0

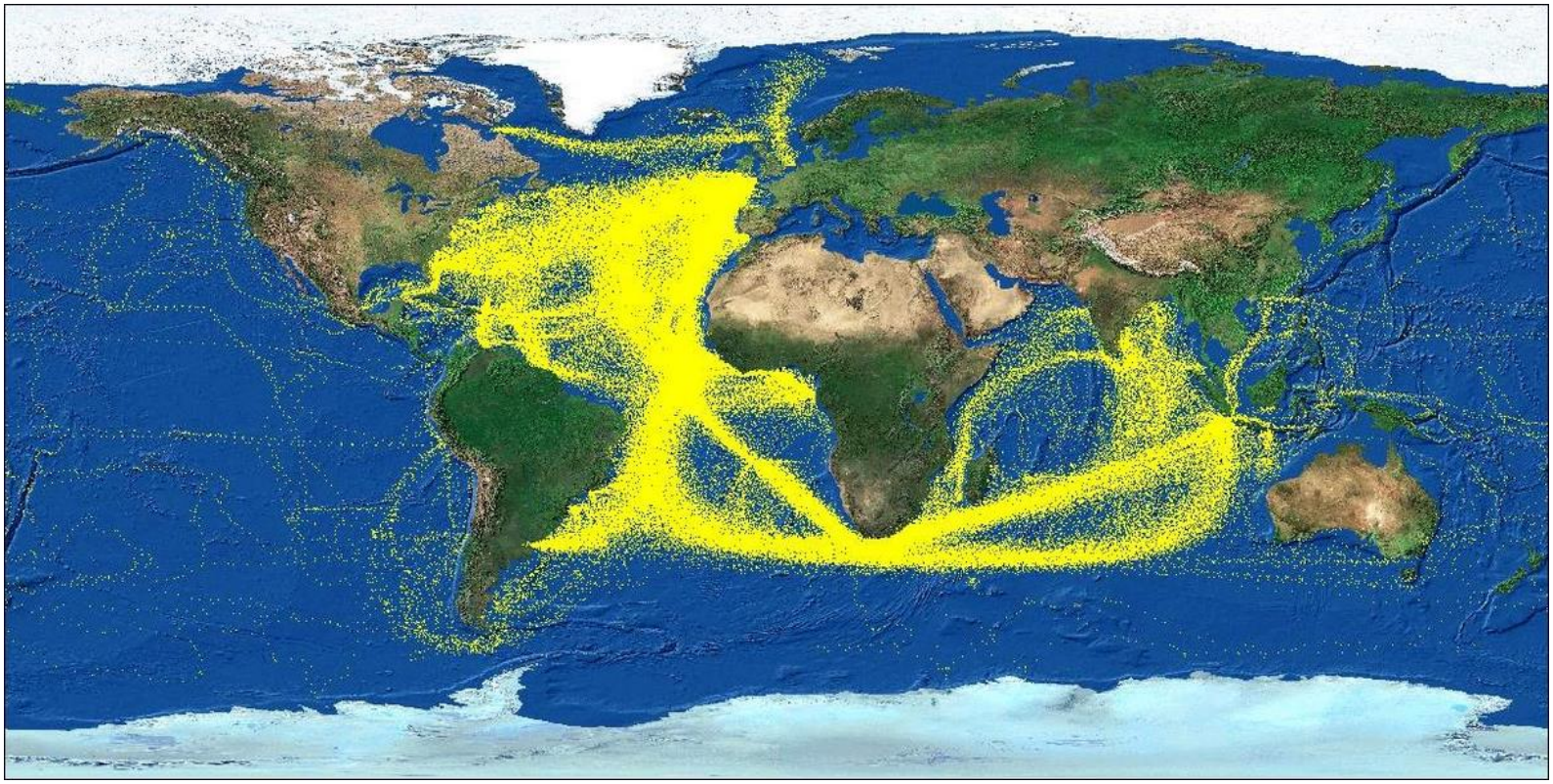
Now

Slide credits: Shaun de Witt (CCFE) for EOSC-hub

Or more generic:

An introduction to the basics of research data
https://www.youtube.com/watch?v=q2aiDJzJPuw

From “real life” to research: CLIWOC - climatological database for the world’s oceans



Every yellow dot represents a ship report.

Image copied from https://www.knmi.nl/kennis-en-datacentrum/achtergrond/cliwoc

Project web site: http://pendientedemigracion.ucm.es/info/cliwoc/

From research into real life

- https://www.greenpeace.org/archive-africa/en/Press-Centre-Hub/New-satellite-data-reveals-the-worlds-largest-air-pollution-hotspot-is-Mpumalanga---South-Africa/
- http://data.enel.com/blog/ebola-open-data
- https://www.nature.com/articles/533469b

https://pixabay.com/en/hand-taking-care-of-launchy-295702/

Managing and sharing data for your own sake

Image: https://www.flickr.com/photos/dmh650/4031607067/in/gallery-wlef70-72157633022909105/

arXiv.org > astro-ph > arXiv:1511.02512 Search or Ar

Astrophysics > Instrumentation and Methods for Astrophysics

The data sharing advantage in astrophysics

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

(Submitted on 8 Nov 2015)

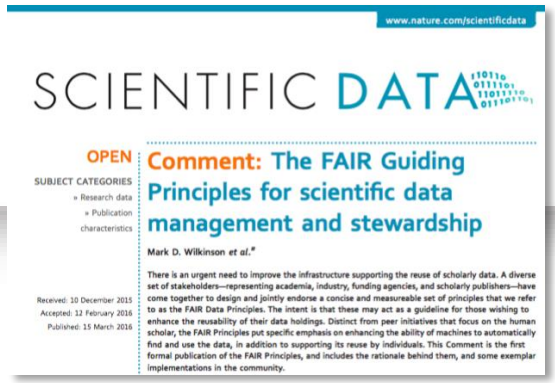
We present here evidence for the existence of a citation advantage within astrophysics for papers that link to data. Using simple measures based on publication data from NASA Astrophysics Data System we find a citation advantage for papers with links to data receiving on the average significantly more 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)

Open? ✓

FAIR? ????

FAIR data principles



Findable

- Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets;
- Assign persistent IDs, provide rich metadata, register in a searchable resource, ...

Accessible

- Stored for long term such that they can be easily accessed and/or downloaded with well-defined licence and access conditions (Open Access *when possible*), whether at the level of metadata, or at the level of the actual data content;
- Retrievable by their ID using a standard protocol, metadata remain accessible even if data aren't...

Interoperable

- Ready to be combined with other datasets by humans as well as computer systems
- Use formal, broadly applicable languages, use standard vocabularies, qualified references...

Reusable

- Ready to be used for future research and to be processed further using computational methods.
- Rich, accurate metadata, clear licences, provenance, use of community standards...

http://www.nature.com/articles/sdata201618
http://www.dtls.nl/fair-data/
www.force11.org/group/fairgroup/fairprinciples

Mons, B. et al. (2017) *Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud* https://dx.doi.org/10.3233/ISU-170824

Horizon2020: Open and FAIR

European Commission

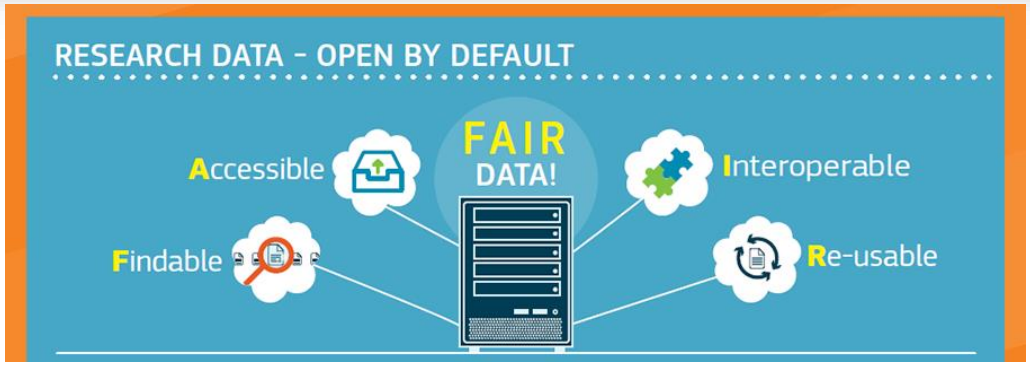
Clarifying terminology...

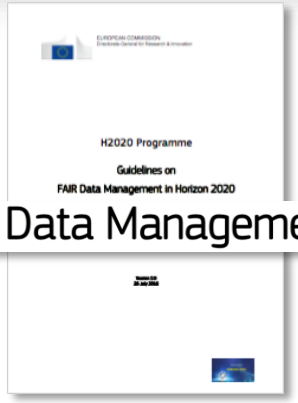
In the past our policy mainly addressed the 'accessibility' part of FAIR.

- **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

Principles =/= practice





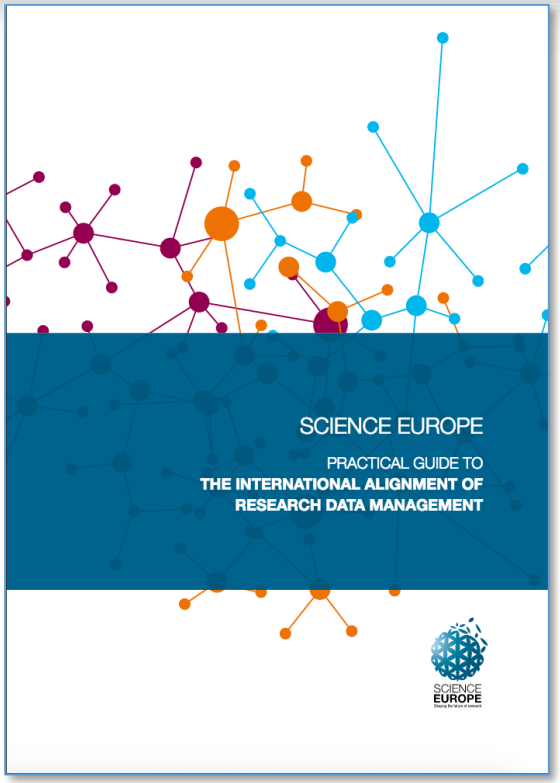
H2020 DMP Guidelines: "This template is inspired by FAIR as a general concept." Meaning: find your own (disciplinary) practice.

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

GO FAIR: initiative towards the internet of FAIR data and services. Started in Europe, but reaches out wide.
https://www.dtls.nl/fair-data/go-fair/

Infographic EC: http://ec.europa.eu/research/images/infographics/policy/open-data-2016-w920.png

“Core DMP requirements” published last week



Research funding organisations, research organisations,[1] and individual researchers have different needs and requirements when it comes to research data management (RDM). The core requirements for data management plans (DMPs) and criteria for the selection of trustworthy repositories presented in this guide provide organisations and communities with a common basis from which they can develop RDM policies. These should be considered as minimum requirements that can be supplemented with more specific ones according to the needs of the community or organisation in question.

http://scieur.org/rdm-guide

Simplified research data lifecycle

CREATING DATA: designing research, DMPs, planning consent, locate existing data, data collection and management, capturing and creating metadata

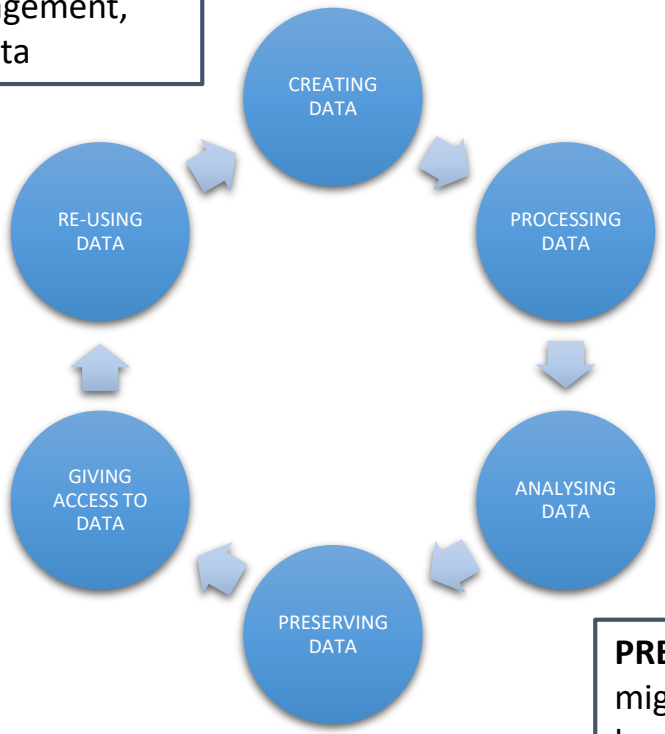
PROCESSING DATA: entering, transcribing, checking, validating and cleaning data, anonymising data, describing data, manage, store, back-up data

RE-USING DATA: follow-up research, new research, undertake research reviews, scrutinising findings, teaching & learning

ANALYSING DATA: interpreting, & deriving data, producing outputs, authoring publications, preparing for sharing

ACCESS TO DATA: distributing data, sharing data, controlling access, establishing copyright, promoting data

PRESERVING DATA: data archiving, migrating to best format & medium for long term, creating metadata and documentation

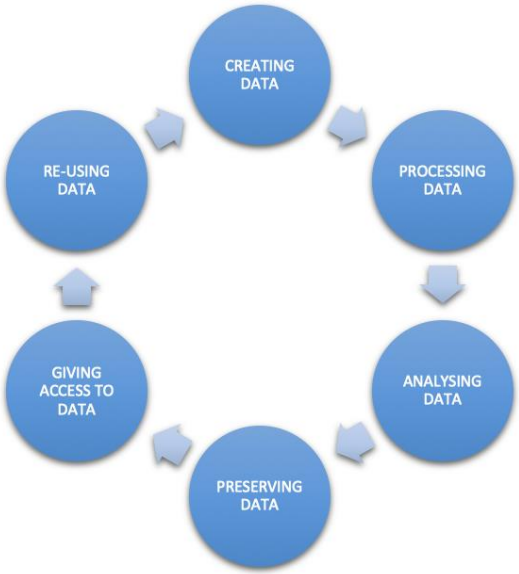


Based on UK Data Archive lifecycle: https://www.ukdataservice.ac.uk/manage-data/lifecycle
Used in OpenAIRE RDM briefing paper: https://www.openaire.eu/briefpaper-rdm-infonoads

Data management plans

A DMP is a brief plan to define:

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

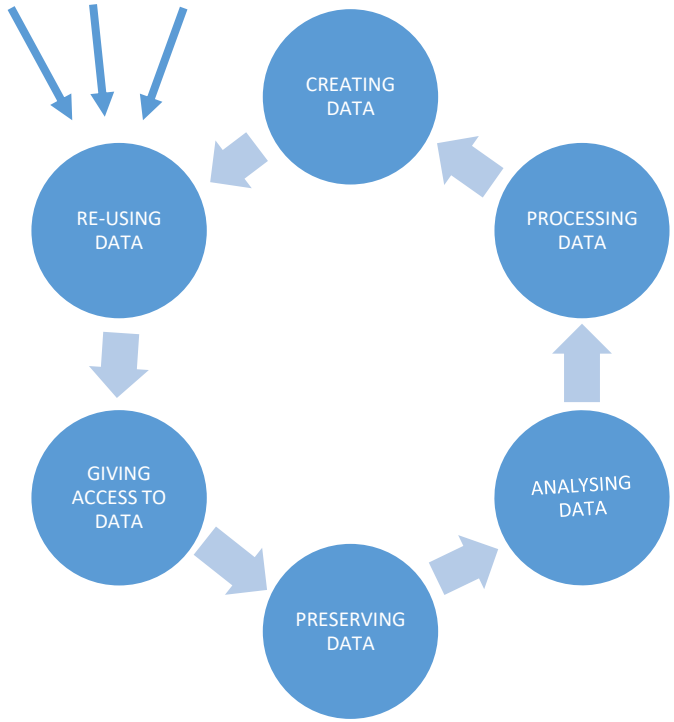


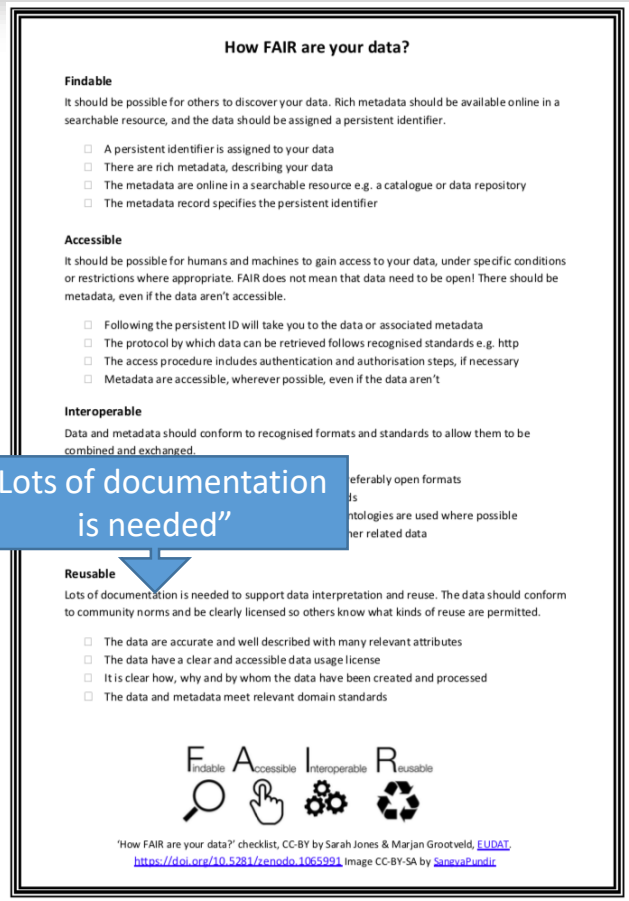
DMPs are sometimes submitted as part of grant applications, sometimes afterwards, but they are useful whenever researchers are creating data.

https://pixabay.com/en/hand-business-plan-business-3190204/

Planning trick 1: think backwards

What would a re-user need?





How FAIR are your data?

Findable

It should be possible for others to discover your data. Rich metadata should be available online in a searchable resource, and the data should be assigned a persistent identifier.

- ☐ A persistent identifier is assigned to your data
- ☐ There are rich metadata, describing your data
- ☐ The metadata are online in a searchable resource e.g. a catalogue or data repository
- ☐ The metadata record specifies the persistent identifier

Accessible

It should be possible for humans and machines to gain access to your data, under specific conditions or restrictions where appropriate. FAIR does not mean that data need to be open! There should be metadata, even if the data aren't accessible.

- ☐ Following the persistent ID will take you to the data or associated metadata
- ☐ The protocol by which data can be retrieved follows recognised standards e.g. http
- ☐ The access procedure includes authentication and authorisation steps, if necessary
- ☐ Metadata are accessible, wherever possible, even if the data aren't

Interoperable

Data and metadata should conform to recognised formats and standards to allow them to be combined and exchanged.

- ☐ eferably open formats
- ☐ ds
- ☐ ntologies are used where possible
- ☐ her related data

Reusable

Lots of documentation is needed to support data interpretation and reuse. The data should conform to community norms and be clearly licensed so others know what kinds of reuse are permitted.

- ☐ The data are accurate and well described with many relevant attributes
- ☐ The data have a clear and accessible data usage license
- ☐ It is clear how, why and by whom the data have been created and processed
- ☐ The data and metadata meet relevant domain standards

Findable Accessible Interoperable Reusable

'How FAIR are your data?' checklist, CC-BY by Sarah Jones & Marjan Grootveld, EUDAT, https://doi.org/10.5281/zenodo.1065991 Image CC-BY-SA by SangyaPundir

EUDAT FAIR checklist - CC-BY Sarah Jones & Marjan Grootveld, EUDAT. https://doi.org/10.5281/zenodo.1065991

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 keywords to support searching
- Persistent and unique identifiers such as DOI
- Versioning of the datasets and clear version numbers

Metadata

Findable

Metadata (persistent identifier included) is needed to locate research data and get a first idea of the content.

Use relevant standards to enable interoperability.

- Arts and humanities
- Engineering
- Life sciences
- Physical sciences and mathematics
- Social and behavioral sciences
- General research data: e.g. Dublin Core and DataCite

Check which standards the long-term repository supports or expects.

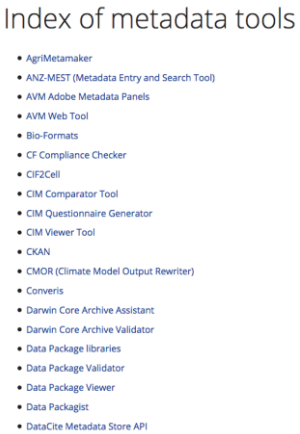
FAIRsharing.org
standards, databases, policies

https://fairsharing.org/

RDA
RESEARCH DATA ALLIANCE

DCC

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



https://rdamsc.dcc.ac.uk/

Extra: metadata tools:
https://rdamsc.dcc.ac.uk/tool-index

Documentation?

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

In short: document and preserve everything that is needed to replicate 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.

CoreTrustSeal
Data Seal of Approval
ICSU World Data System
nestor seal
ISO 16363

Sharing data: what is meant?

With collaborators while research is active

Dropbox

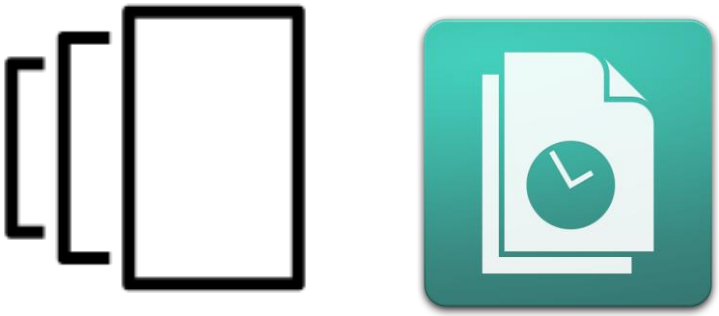
Data are mutable

(Open) data sharing

Data are stable, searchable, citable, clearly licensed

Storing versus archiving

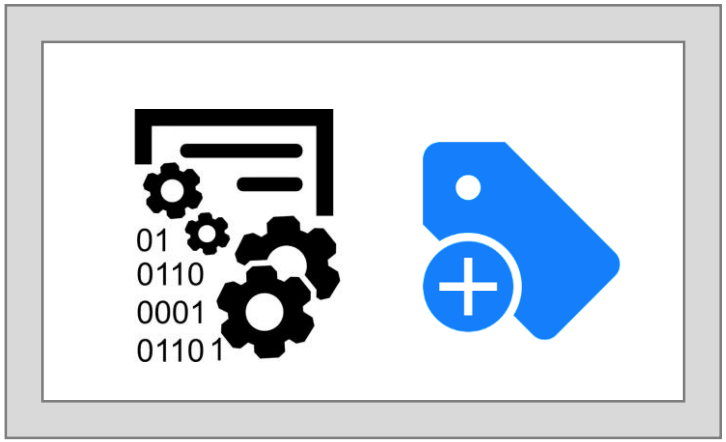
Storing and backing up files while research is active



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

Archiving, repositories, ehm?

Horizon 2020 demands pilot participants to "deposit your data in a research data repository": a digital archive collecting and displaying datasets and their metadata.

Select a data repository that will preserve your data, metadata and possibly tools in the long term.

Repositories may offer guidelines for sustainable data formats and metadata standards, as well as support for dealing with sensitive data and licensing. Plus: they provide persistent identifiers.

It is advisable to contact the repository of your choice when writing the first version of your DMP.

Where to find a repository?

More information: https://www.openaire.eu/find-trustworthy-data-repository

Zenodo: http://www.zenodo.org

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

Keep everything? For always?

- When regenerating data would be cheaper than archiving, don't archive. Select what data you'll need and want to retain.
- 10 years is often stated in data policies and academic codes, but data can be valuable for ages, in climatology, sociology, health sciences, astronomy, linguistics, … Look beyond minimal retention periods where relevant.
- Explain your selection criteria in the DMP.

RDNL Selection criteria: http://www.researchdata.nl/en/services/data-management/selecting-research-data/

DCC How-to guide: http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

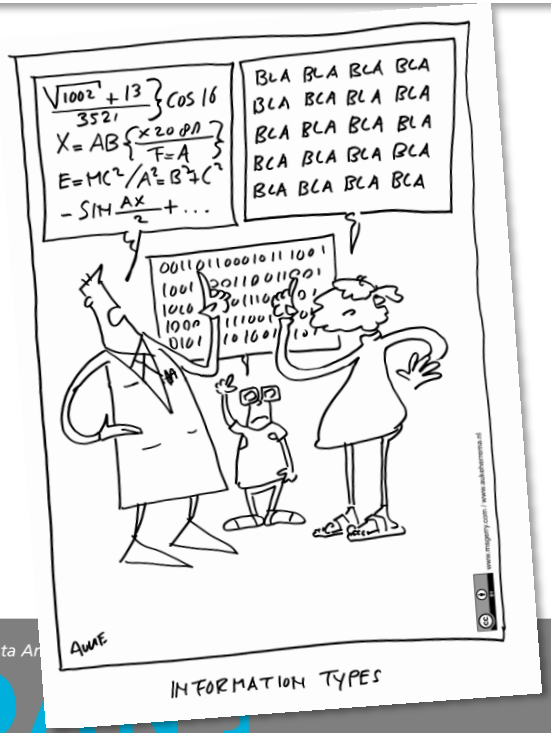

D2 A1 T1 A1

Interoperability

Interoperable

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**.

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.

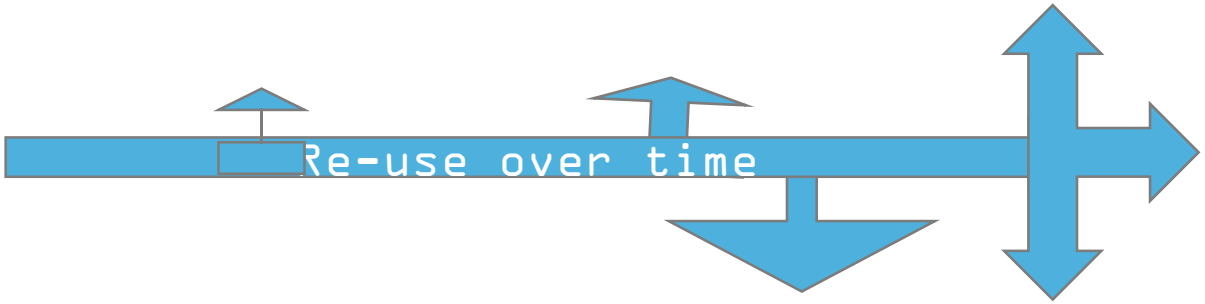


Medical classification is the process of transforming descriptions of medical diagnoses and procedures into universal medical code numbers. SNOMED Clinical Terms (SNOMED CT) is intended to provide a set of concepts and relationships that offers a **common reference point for comparison and aggregation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

Some "R" questions

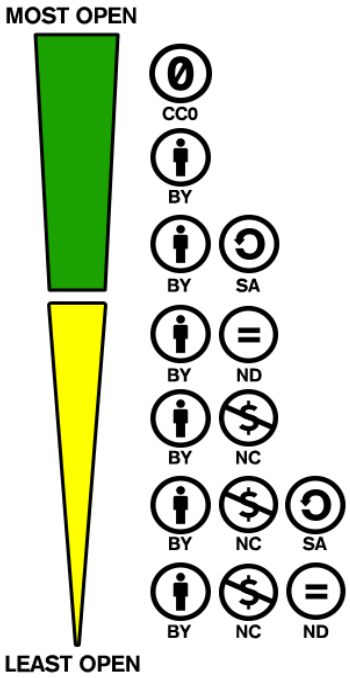
§ 2.4 Increase data re-use (through clarifying licences)

- 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 1/2

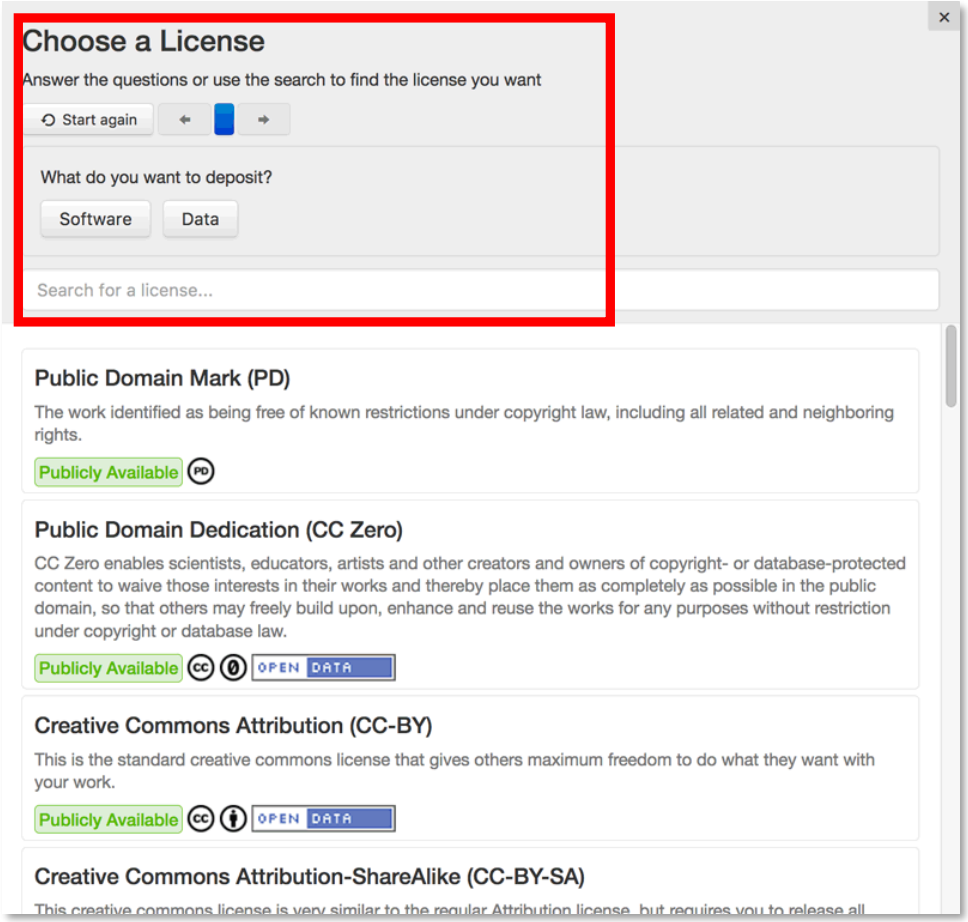
Horizon 2020 guidelines point to CC0 “waiver” or CC-BY licence



https://creativecommons.org/share-your-work/public-domain/freeworks/

Licensing research data 2/2

EUDAT/B2SHARE licensing wizard helps you pick a licence for data & software



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

Remaining sections in H2020 template

§ 3 Allocation of resources:

- Estimate cost for making data FAIR. How do you cover them?
- Who is responsible for data management in the project?

§4 Security:

- Address data recovery as well as secure storage and transfer of sensitive data

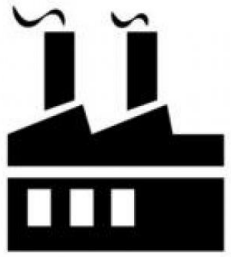
§5 Ethical aspects:

- To be covered by ethics review, ethics section of DoA and ethics deliverables. Otherwise: describe.

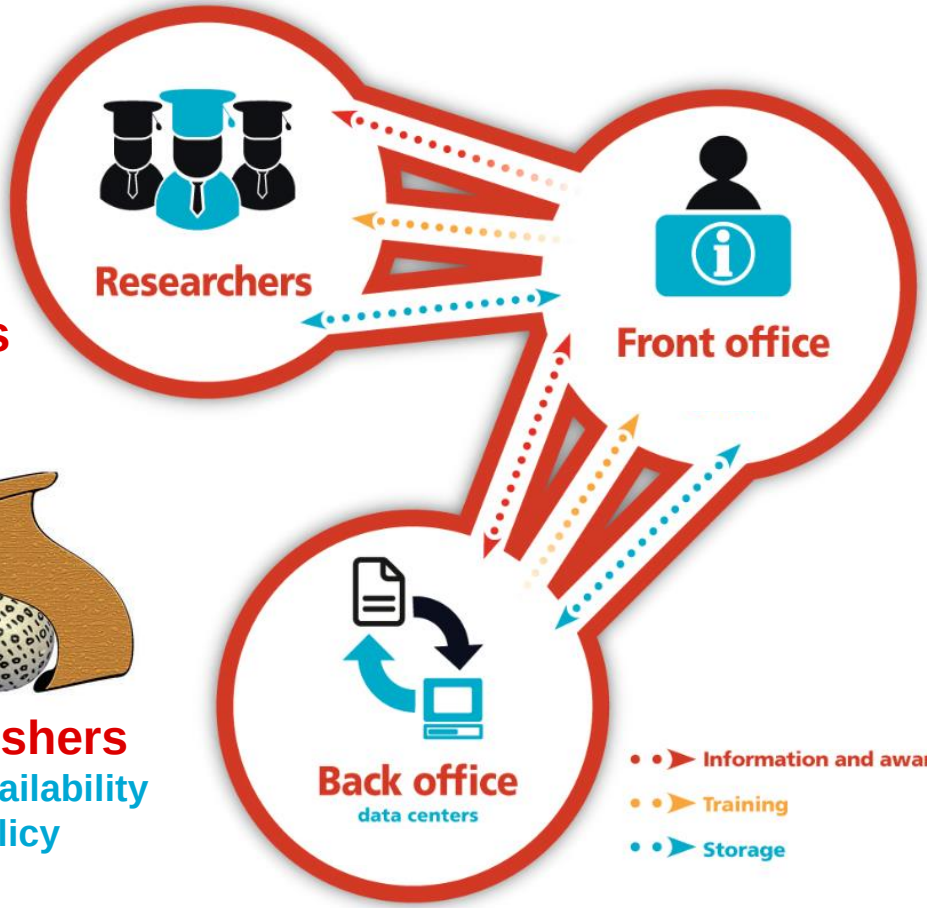
Overwhelmed?

A DMP is also a communication instrument!

Planning trick 2: include stakeholders



Commercial partners



Institution

RDM policy

Facilities

France

Andre Dazy

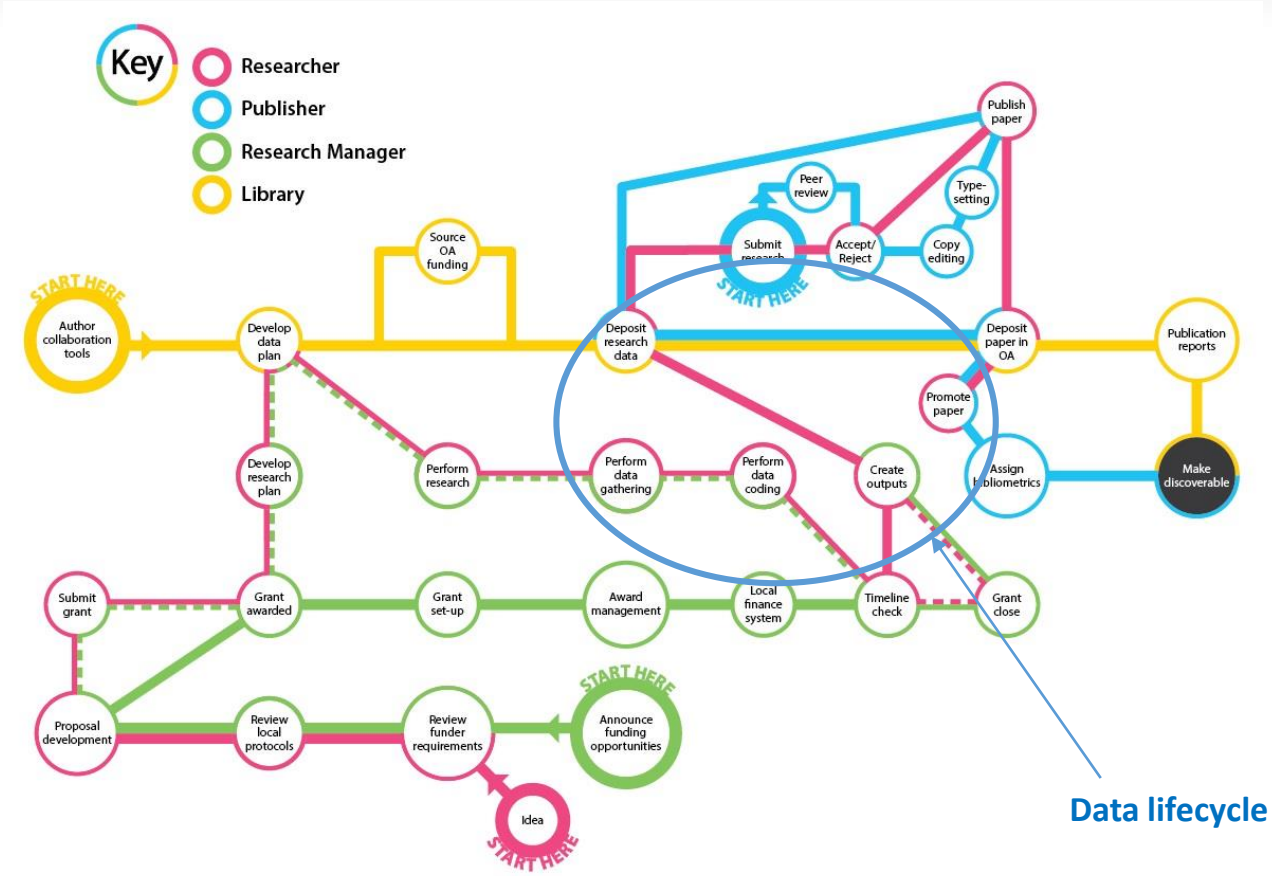
OpenAIRE Helpdesk

Publishers

Data Availability policy

Research funders

The *data* lifecycle as part of *research* lifecycle



"Open Access Tube Map" (CC-BY) - Awre, Chris L.; Stainthorp, Paul; and Stone, Graham (2016) "Supporting Open Access Processes Through Library Collaboration", *Collaborative Librarianship*: Vol. 8 : Iss. 2 , Article 8.

Responsibilities in RDM

- The principal investigator – ultimately responsible for the data and for data management
- Researchers, research assistants and/or data managers – involved in day-to-day data management
- The institution's management – draft and enforce data policies; raise data awareness
- The institution's research office consisting of library, IT and legal services – provide external data, tools, secure storage and access; expertise on rights management and ethics, data citation, metadata, access and licenses, funder requirements; raise data awareness
- Research funders – encourage good data practices; invest in data infrastructure; raise data awareness
- Project partners in academic and other research institutions as well as commercial partners
- Academic publishers – impose requirements on the availability of data underlying submitted and/or published papers; provide identifiers to cite papers and link to related data
- Research data repositories – preserve data long term; provide persistent identifiers and data discovery service

https://www.openaire.eu/briefpaper-rdm-infonoads

H2020 open data pilot: to remember

Some important notions:

- Data should be made open when possible, restricted when necessary
- Metadata and other standards
- Arrangements with the identified repository
- Documentation about the software needed to access the data
- Licences to permit the widest reuse possible
- Potential value of long-term preservation

”Costs related to open access to research data in Horizon 2020 are eligible for reimbursement during the duration of the project under the conditions defined in …”:

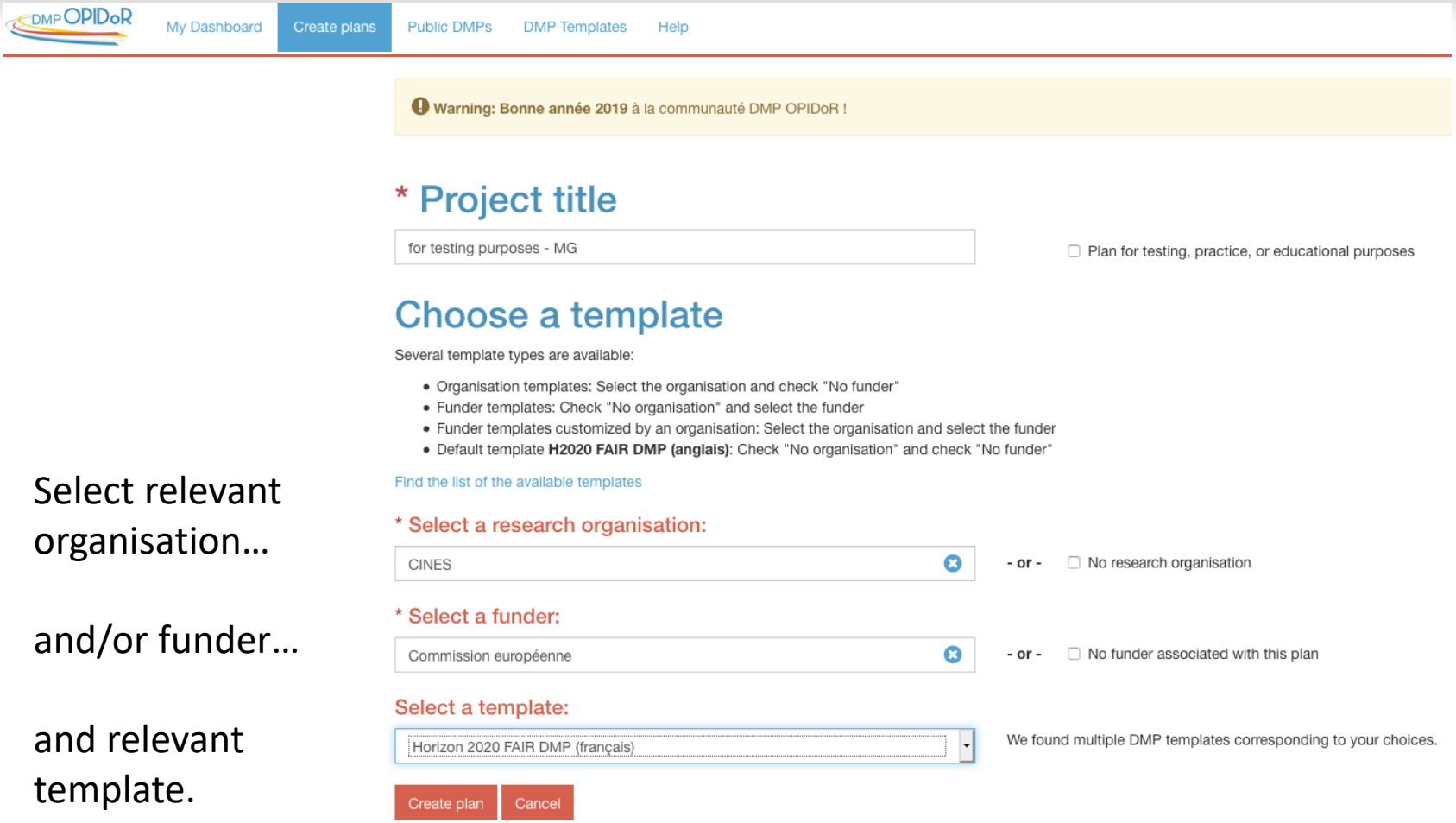
  - Only if budgeted in the proposal and granted;
  - Only during the project

Recommended: Estimating RDM Costs tool
https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs

https://pixabay.com/en/tools-construct-craft-repair-864983/

Opidor: https://dmp.opidor.fr/

Login with your institutional account.



Select relevant organisation…

and/or funder…

and relevant template.

NB: template "Horizon 2020 DMP" is a previous version – not for new projects.

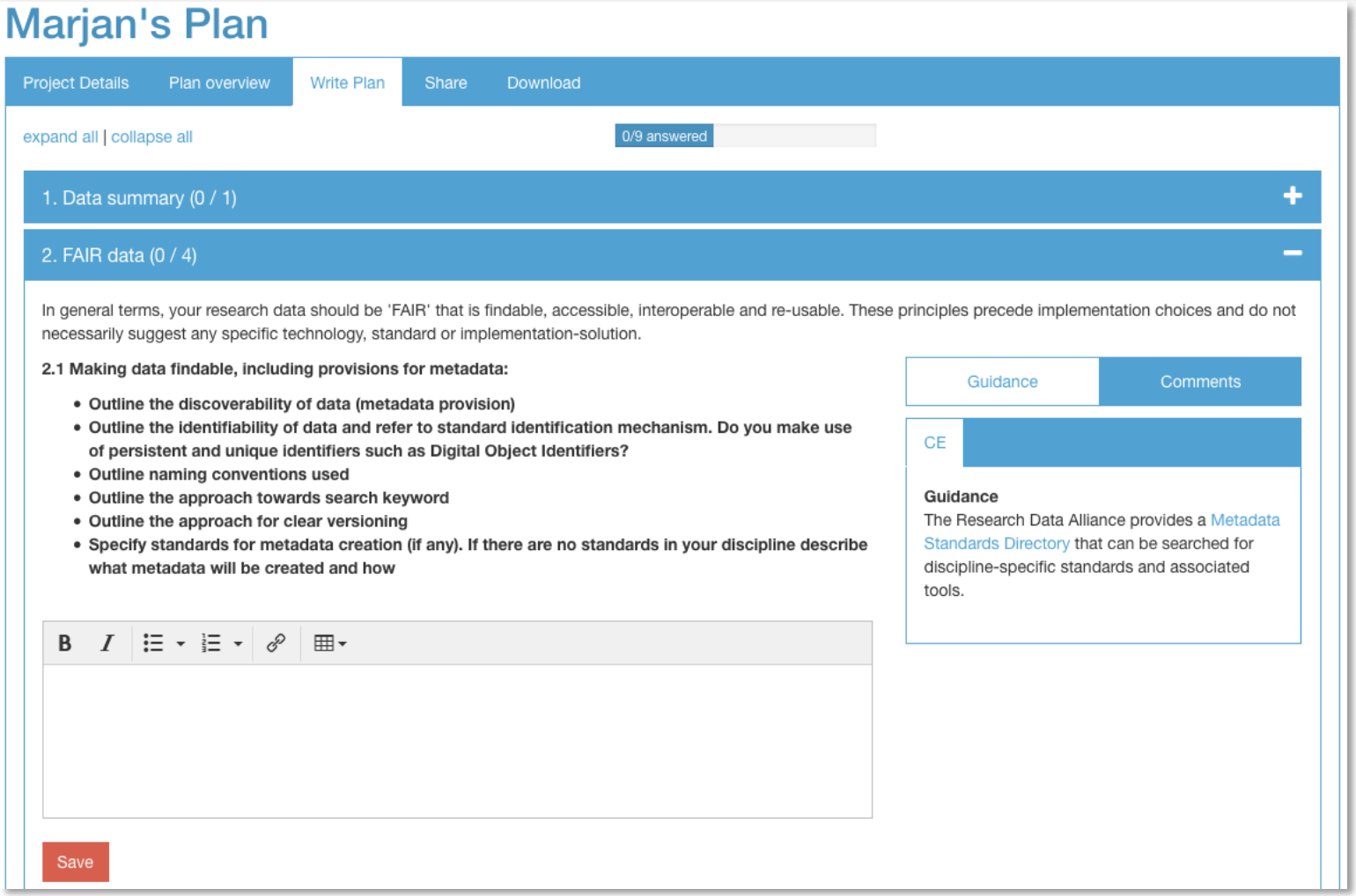
Horizon2020 template in English

https://dmp.opidor.fr/public_templates and select:

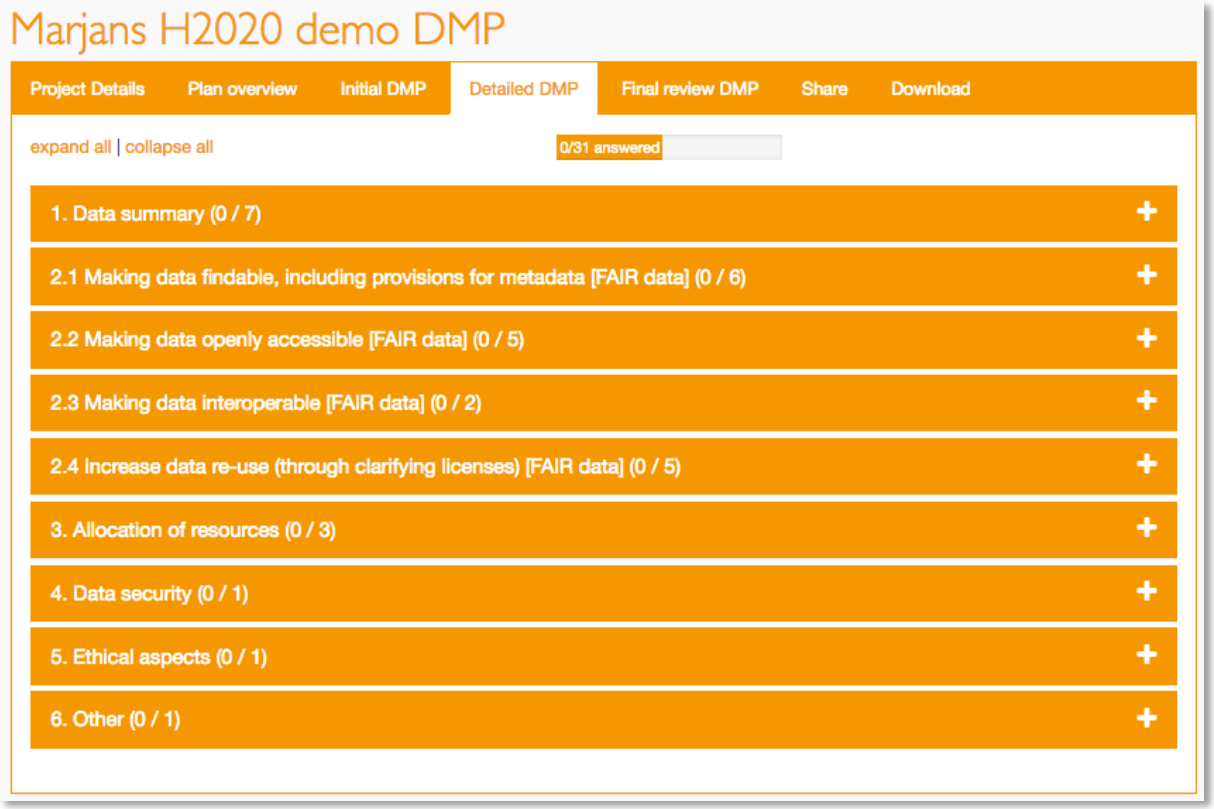
Horizon 2020 FAIR DMP (anglais)	Commission européenne	Funder	The Commission is running a flexible pilot under Horizon 2020 called the Open Research Data Pilot (ORD pilot). Projects participating in the pilot must submit a first version of the DMP (as a deliverable) within the first 6 months of the project. The DMP needs to be updated over the course of the project whenever significant changes arise.	31-10-2018	DOCX PDF	

Click + to open this template and create a DMP.

Sample question and guidance



Alternative (original) DMP tool: https://dmponline.dcc.ac.uk/



Both tools…

- contain the EC's Horizon2020 DMP template & guidance;
- allow you to collaborate with others on your DMP (under construction);
- allow you to export your DMP.

Another kind of tool: FAIR metrics

DANS developing FAIR badges for existing data

Findable (defined by metadata (PID included) and documentation)
1. No PID nor metadata/documentation 2. PID without or with insufficient metadata 3. Sufficient/limited metadata without PID 4. PID with sufficient metadata 5. Extensive metadata and rich additional documentation available
Accessible (defined by presence of user license)
1. Metadata nor data are accessible 2. Metadata are accessible but data is not accessible (no clear terms of reuse in license) 3. User restrictions apply (i.e. privacy, commercial interests, embargo period) 4. Public access (after registration) 5. Open access unrestricted
Interoperable (defined by data format)
1. Proprietary (privately owned), non-open format data 2. Proprietary format, accepted by Certified Trustworthy Data Repository 3. Non-proprietary, open format = 'preferred format' 4. As well as in the preferred format, data is standardised using a standard vocabulary format (for the research field to which the data pertain) 5. Data additionally linked to other data to provide context

FAIR badge scheme

- FAIR as proxy for data “quality” or “fitness for (re-)use”
- We want to create a badge system using the FAIR principles to assess datasets in a repository
- Score each FAIR dimension on a 5-point scale
- Consider Reusability as the resultant of the other three:
  - (F+A+I)/3=R
- Pilot study was useful. DANS is now redesigning.

2 User Reviews
1 Archivist Assessment
24 Downloads

Assessing the FAIRness of existing datasets

The tool runs a series of questions (maximum of 5 per principle) which follow routing options to display the star rating scored per principle.

- In pairs: explore FAIRdat and assess datasets from various disciplines
- How do you like it? (5 minutes)

FAIRdat prototype: https://www.surveymonkey.com/r/fairdat

Handout with datasets: https://goo.gl/749dmf

https://pixabay.com/en/bale-straw-bales-agriculture-457709/

RDM is all in a day's work

Data can be anything, and always needs to be properly managed.

Decisions made early affect what you can do later.

Planning is more important than the plan itself, but

- Plan for the desired end result
- Involve the other stakeholders
- Make an explicit DMP
- Keep it up-to-date

Did you know?

- When you integrate Open Science in your European research proposal, this makes your proposal more competitive.
  - Grigorov, Ivo; Elbæk, Mikael; Rettberg, Najla; Davidson, Joy: "Winning Horizon 2020 with Open Science". https://doi.org/10.5281/zenodo.12247

- There is evidence that grant proposals are receiving praise for including a DMP outline – even though in H2020 a DMP is not required at the proposal stage, and not a competitive point.

So: you better start early on a concrete and convincing DMP ;-)
Good luck!

Thanks to Ivo Grigorov (Technical University of Denmark, FOSTER project) for sharing these quotes.

Oh, and **AMID PROFS = FAIR OS DMP**

Acknowledgements:
https://eoscpilot.eu/
https://eudat.eu/
https://www.eosc-hub.eu/
https://www.openaire.eu/
https://www.project-freya.eu/en
and many colleagues involved in training and project communication

marjan.grootveld@dans.knaw.nl
https://dans.knaw.nl/en/projects/