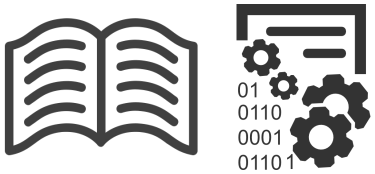
DHARMA Kickoff-Meeting,
16 September 2019, Berlin



ERC Rules on Open Access Publication and the Open Data Research Pilot

Emmanuelle Morlock (CNRS, HiSoMA Research Centre)

cnrs UNIVERSITÉ DE LYON HISTOIRE ET SOURCES DES MONDES ANTIQUES HiSoMA

erc

Making research results openly available is not an afterthought when the project is over, but, rather, it is an effective way of doing the research itself.

Outline

1. Context: OA policies in Europe
2. The Horizon 2020 OA mandate and ERC specificities
3. Key notions:
   FAIR, DMP, datapapers...

2007 - FP7
OA **pilot for publications**
ERC guidelines
recommends OA for **data**
from the beginning

2014 - Horizon 2020
OA obligation for publications + **OA data pilot**

2017 - Horizon 2020
OA by default for publication and data

2018 - EOSC
Zooming on Open Science

sept 2018
Science Europe initiative to
accelerate the transition
to full Open Access

2021 - Horizon Europe
Open Science **modus operandi**
(Open data by default)

Evolution of policies in the EU funding programs for R&I

On the road to Open Science (By Patrick Hochstenbach CC BY 4.0)

Open Science is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods.

Open research beyond academia: innovation + citizen science

Source: FOSTER e-learning platform

Benefits of Open Science

EFFICIENCY

Increase research efficiency (avoid duplication of research, more results with the same data, more participation in the work.

KNOWLEDGE

Enables the creation of new research questions.

QUALITY & INTEGRITY

wider evaluation and scrutiny by the scientific community.

COLLABORATION

International, better understanding of challenges requiring coordinate efforts at a large scale.

VISIBILITY

Share your work and make it more visible.

ECONOMIC BENEFITS

Reduce publishing costs, value creation through innovation.

What is Open Access (OA)?

H2020 => 0nline access to scientific information that is:

- **free of charge** to the end-user
- **reusable**

Scientific information:

- peer-reviewed scientific research **articles**
- research **data**

Multi-beneficiaries General Model Grant Agreement

- 29.2. Open Access to scientific publications
- 29.3. Open Access to research data

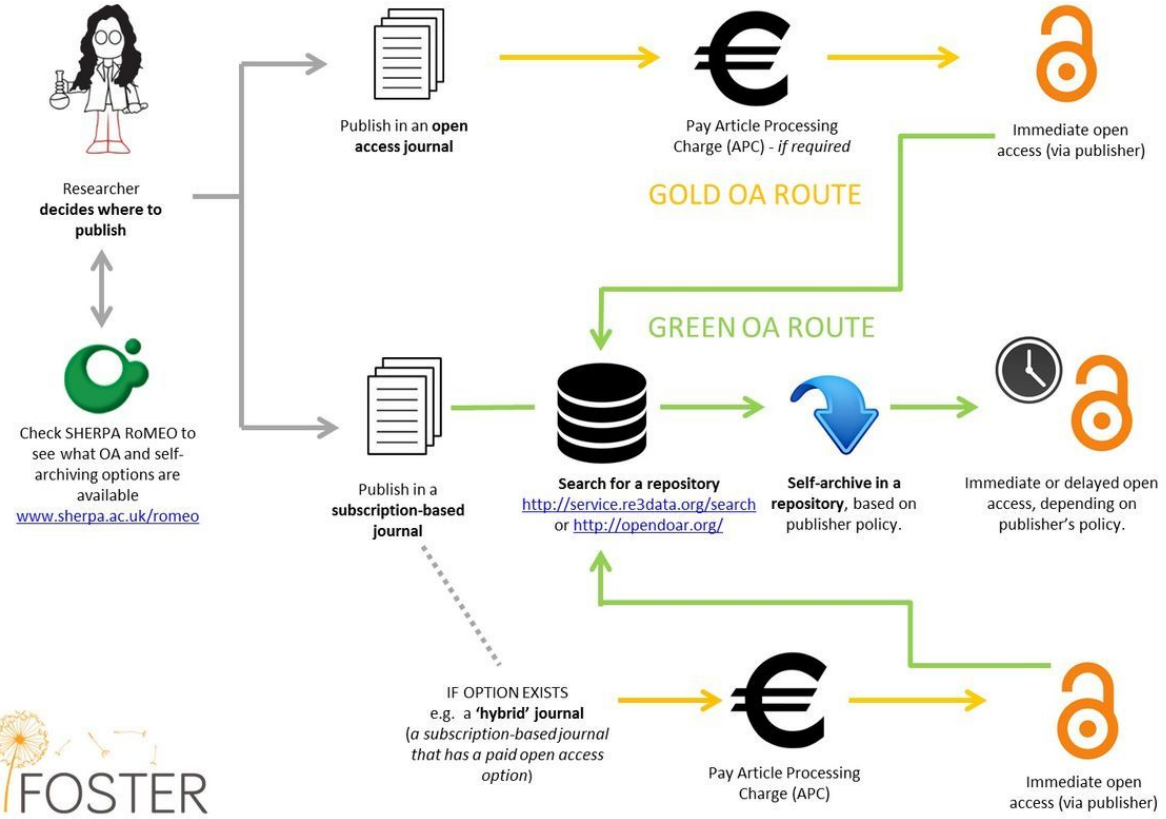
H2020 OA mandate for publications

All peer-reviewed scientific publications related to results (including monographs and book chapters) must be:

- deposited in a **repository** (even for if in a OA journal already)
- in the form of a **machine readable electronic copy**
- the published **version** or the final peer-reviewed manuscript accepted for publication
- as **soon** as possible
- with identifying **metadata**

Open access costs eligible for reimbursement when the project runs.

Publications: routes to Open Access



Maximum embargo 6 months (12 for SSH)

FOSTER

www.fosteropenscience.eu

H2020 OA mandate for data

- Limited pilot launched in 2014, in 2017
  - participating is the default option for all projects (opt out possible)
- Develop a Data Management Plan (DMP)
- Provide open data (if possible)
  - for (meta)data needed to validate results
  - other (meta)data curated or raw data specified in the DMP

As open as possible, as closed as necessary!

ERC specificities for publications

- ERC rules closely aligned to the H2020 requirements (cf. MGA 29.2)
  - expected: data underlying Publications
- Disciplinary or institutional repositories recommended
- Embargo extended to 12 months if pub. linked to SSH (upon request)
- Metadata: lighter requirements for bibliographic metadata
  - but PID mandatory (DOI)
  - Good practice: ERC grant number indicated in metadata
- Open license encouraged (CC BY)

ERC specificities for research data

Slightly different 29.3 (data): opt-in to the ORD pilot optional

- Data management must follow the FAIR principles
- Submit the first version of a DMP within 6 months
  specific template for ERC
- Deposit data mentioned in the DMP in a trusted (certified) repository
- Exceptions acceptable if it would jeopardise any objective (not just main)
- Opt-out: no justification needed
- Recommended: Information and tools used for the validation of the results

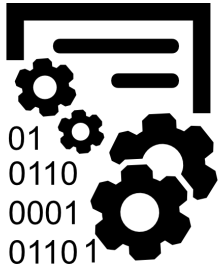
Ultimately, it is for the individual investigator to decide which data merit conservation and/or sharing.

Non compliance with OA requirements

- Grant may be reduced (art. 43)
- Other measures describes in chapter 6 of the MGA

Publications

- All publications including monographs and book chapters
- Always deposit in Open repository
- A PID is required



Data

- Degrees of data sharing
- Needs Research Data Management compliant with the FAIR principles
- Don’t forget to: think about tools/infos needed to validate results

PUBLICATIONS AND DATA

Data is increasingly becoming a legitimate citable research output.

DATA?

What is data?

Definitions vary based on the research community or discipline, legal context, data governance policies…

It's impossible to frame them as a fixed ontological category… especially in the Humanities.

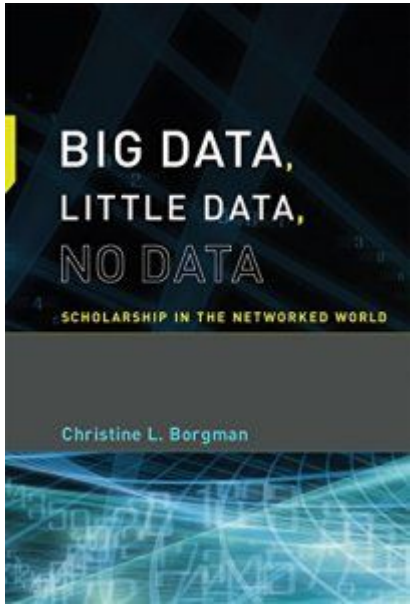
=> Everything that is produced in a research process (but not administrative of pedagogical outputs)

Publications <-> Data

Publications are arguments made by authors,

Data are evidence used to support the arguments.

C.L. Borgman



A digital, selectively constructed, machine-actionable abstraction representing some aspects of a given object of humanistic inquiry.

—Schöch, Christoph. “Big? Smart? Clean? Messy?” Data in the Humanities”. Journal of Digital Humanities. Vol. 2, n°3, 2013 2018

European Research Council

Scientific Council

Established by the European Commission

Open Research Data and Data Management Plans

Information for ERC grantees

by the ERC Scientific Council

Version 3.1
3 July 2019

FAIR?

FINDABLE

To aid automatic discovery of relevant datasets

ACCESSIBLE

Explicit limitations on the use of data, protocols for querying, downloading or copying data are made explicit for both humans and machines

INTEROPERABLE

(Meta)data should use standards, have references to other (meta)data and be machine actionable

REUSABLE

(Meta)data are sufficiently well described for humans and computers to be able to understand them and have a clear and accessible data usage licence

FINDABLE

RICH DESCRIPTIVE (META)DATA

- How will they be discovered?
- What is necessary to make explicit so they can be understood outside the team?

- F1. (Meta)data are assigned a globally unique and **persistent identifier**
- F2. Data are described with **rich metadata** (defined by R1 below)
- F3. Metadata clearly and explicitly **include the identifier** of the data they describe
- F4. (Meta)data are registered or **indexed in a searchable resource**

ACCESSIBLE

EXPLICIT ACCESS PROTOCOLS FOR BOTH HUMANS AND MACHINES

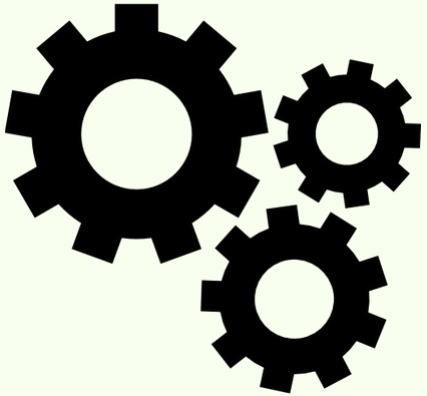
- Where the data will be found?
- Are they openly and freely accessible or are the limitations clear?

- A1. (Meta)data are retrievable by their identifier using a **standardised communications protocol**
  - A1.1 The **protocol is open**, free, and universally implementable
  - A1.2 The protocol allows for an **authentication and authorisation procedure, where necessary**

- A2. **Metadata are accessible**, even when the data are no longer available

INTEROPERABLE

INTEGRATED WITH OTHER DATA & INTEROPERATE WITH APPLICATIONS

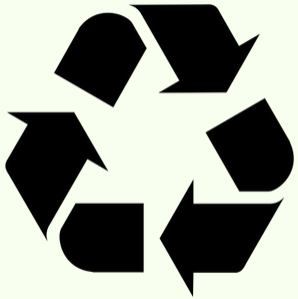


- What are the formal languages in use?
- Is the data intelinked to other (meta)data?
- Is the data machine-actionable?

- I1. (Meta)data use a **formal**, accessible, shared, and broadly applicable **language for knowledge representation**.
- I2. (Meta)data use **vocabularies** that follow FAIR principles
- I3. (Meta)data include qualified **references** to other (meta)data

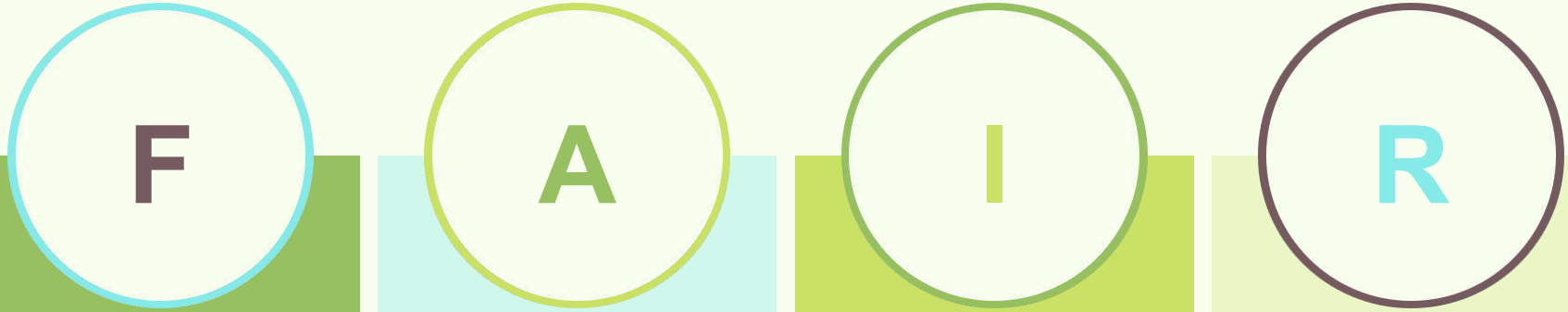
REUSABLE

OPTIMISE THE REUSE OF THE DATA



- How will they be discovered?
- Are reuse conditions clear?
- What is necessary to make explicit so they can be understood outside the team?

- R1. Meta(data) are **richly described** with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible **data usage license**
  - R1.2. (Meta)data are associated with detailed **provenance**
  - R1.3. (Meta)data meet **domain-relevant community standards**



FINDABLE

Metadata

Persistent identifier

Naming conventions

Keywords

Versioning

ACCESSIBLE

Software

Documentation

Access modalities

Data repository

INTEROPERABLE

Standards

Vocabularies

Machine actionable (Meta)data

REUSABLE

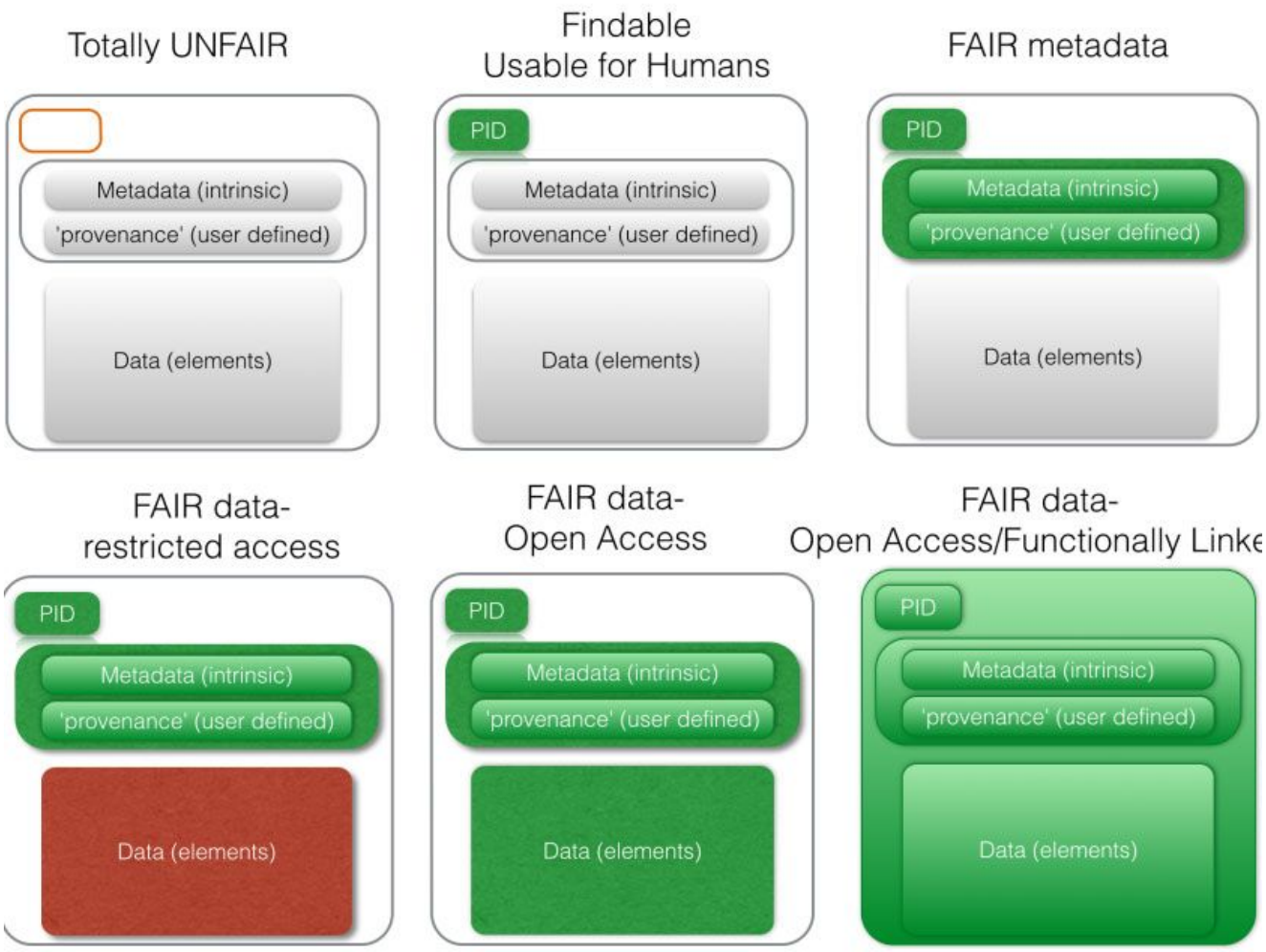
Licensing

Context

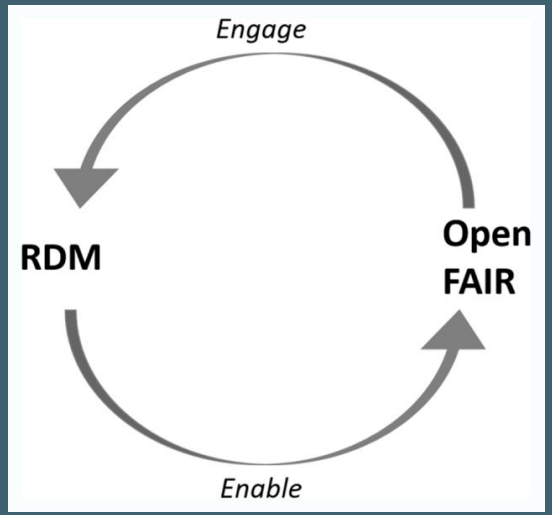
Provenance info

Methodologies

Data as increasingly FAIR Digital Objects



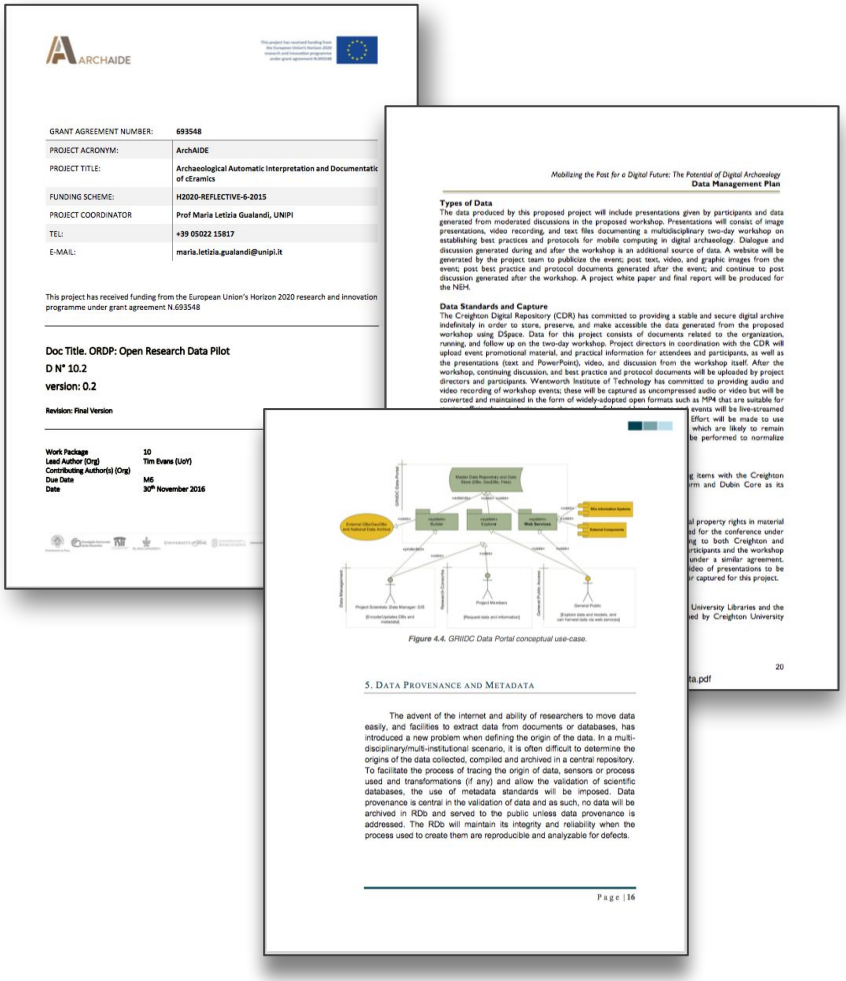
FAIR vs Open vs RDM



Source : Higman, R., Bangert, D., & Jones, S. (2019). Three camps, one destination: the intersections of research data management, FAIR and Open. *Insights*, *32*(1), 18. DOI: http://doi.org/10.1629/uksg.468

How to comply with the ORD Pilot?

1. Write a DMP
2. Find a repository matching data needs
3. Deposit (meta)data and other tools



ARCHAIDE

GRANT AGREEMENT NUMBER:	693548
PROJECT ACRONYM:	ArchAIDE
PROJECT TITLE:	Archaeological Automatic Interpretation and Documentatic of cEramics
FUNDING SCHEME:	H2020-REFLECTIVE-6-2015
PROJECT COORDINATOR	Prof Maria Letizia Gualandi, UNIPI
TEL:	+39 05022 15817
E-MAIL:	maria.letizia.gualandi@unipi.it

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N.693548

Doc Title. ORDP: Open Research Data Pilot
D N° 10.2
version: 0.2
Revision: Final Version

Work Package	10
Lead Author (Org)	Tim Evans (UoY)
Contributing Author(s) (Org)	
Due Date	M6
Date	30th November 2016

Mobilizing the Past for a Digital Future: The Potential of Digital Archaeology
Data Management Plan

Types of Data
The data produced by this proposed project will include presentations given by participants and data generated from moderated discussions in the proposed workshop. Presentations will consist of image presentations, video recording, and text files documenting a multidisciplinary two-day workshop on establishing best practices and protocols for mobile computing in digital archaeology. Dialogue and discussion generated during and after the workshop is an additional source of data. A website will be generated by the project team to publicize the event; post text, video, and graphic images from the event; post best practice and protocol documents generated after the event; and continue to post discussion generated after the workshop. A project white paper and final report will be produced for the NEH.

Data Standards and Capture
The Creighton Digital Repository (CDR) has committed to providing a stable and secure digital archive indefinitely in order to store, preserve, and make accessible the data generated from the proposed workshop using DSpace. Data for this project consists of documents related to the organization, running, and follow up on the two-day workshop. Project directors in coordination with the CDR will upload event promotional material, and practical information for attendees and participants, as well as the presentations (text and PowerPoint), video, and discussion from the workshop itself. After the workshop, continuing discussion, and best practice and protocol documents will be uploaded by project directors and participants. Wentworth Institute of Technology has committed to providing audio and video recording of workshop events; these will be captured as uncompressed audio or video but will be converted and maintained in the form of widely-adopted open formats such as MP4 that are suitable for

Effort will be made to use

20

ta.pdf

Figure 4.4. GRIIDC Data Portal conceptual use-case.

5. Data Provenance and Metadata

The advent of the internet and ability of researchers to move data easily, and facilities to extract data from documents or databases, has introduced a new problem when defining the origin of the data. In a multi-disciplinary/multi-institutional scenario, it is often difficult to determine the origins of the data collected, compiled and archived in a central repository. To facilitate the process of tracing the origin of data, sensors or process used and transformations (if any) and allow the validation of scientific databases, the use of metadata standards will be imposed. Data provenance is central in the validation of data and as such, no data will be archived in RDb and served to the public unless data provenance is addressed. The RDb will maintain its integrity and reliability when the process used to create them are reproducible and analyzable for defects.

Page | 16

Step 1: What is a DMP?

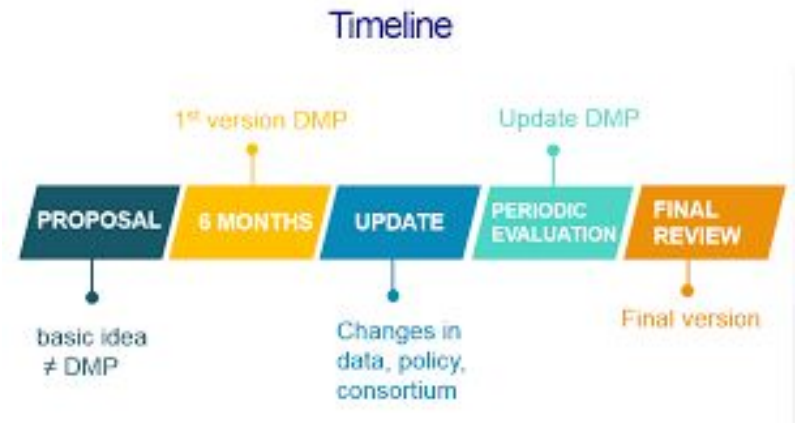
A DMP is a brief plan, formalised in a document, to define:

- how the data will be created
- how it will be documented
- who will be able to access it
- where it will be stored
- who will back it up
- whether (and how) it will be shared & preserved

Step 1: DMP Timeline

The DMP is a living document:

- initial DMP within first 6 months
- updated DMP
  - for significant changes
  - or periodical reviews
- final DMP



source: https://fr.slideshare.net/OpenAIRE_eu/openaire-webinar-open-research-data-in-h2020-149299693 (june 2019)

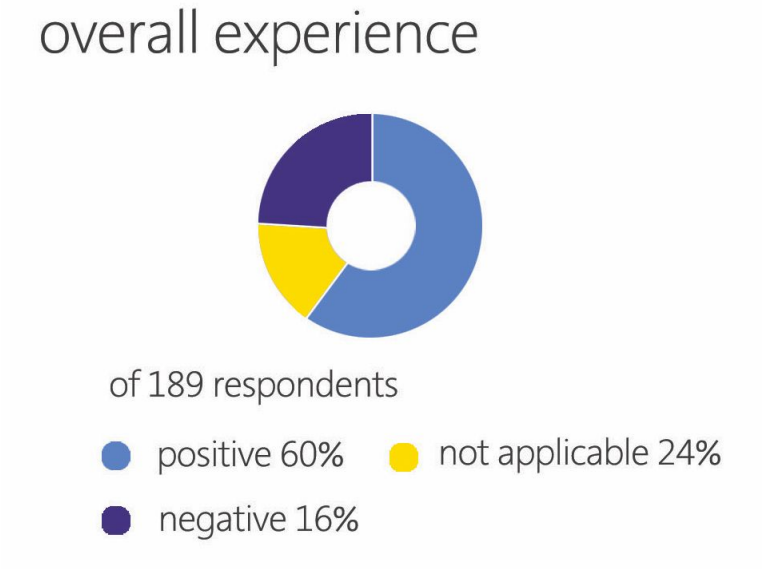
Writing a DMP (...) should be part of the research process from the outset.

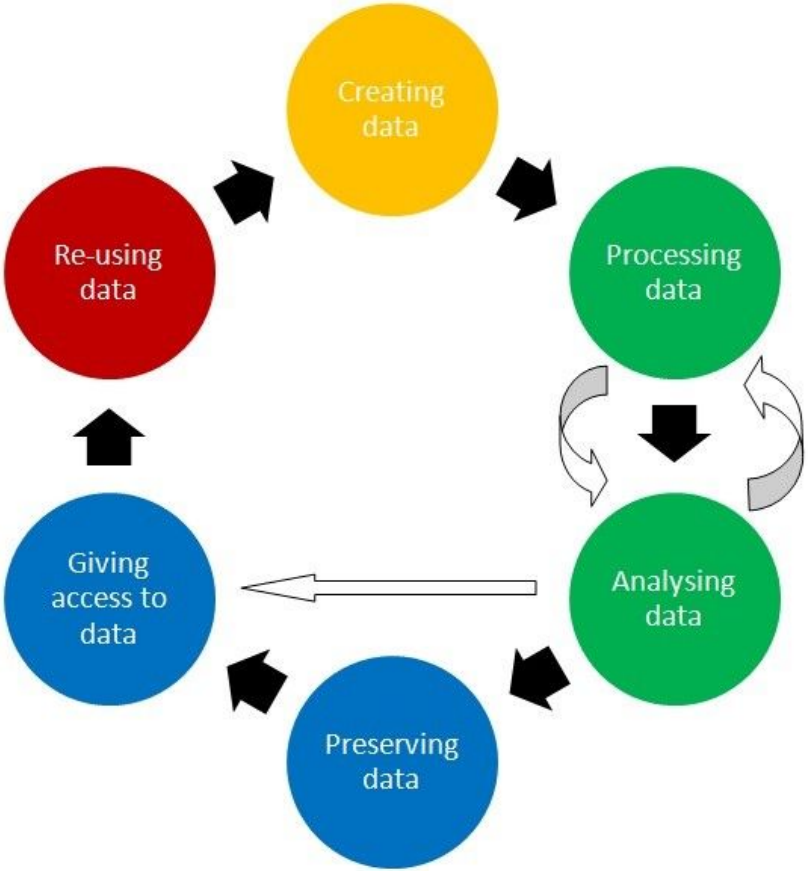
ERC, *Open Research Data and Data Management Plans* (v 3.1 3 July 2019)

DMPs can be useful...

OpenAire & FAIR data expert group survey (2017)

Report : http://doi.org/10.5281/zenodo.1120245





http://blogt.ethz.ch/innovethbib/en/2015/09/07/management-von-forschungsdaten-start-des-projekts-data-life-cycle-management/2015-09-08_data-life-cicle/

Benefits of writing a DMP

- Guides the decision making
- Centralises information about the data
- Optimizes data management
- Help building dissemination strategies

Evaluation criteria for DMPs

Performance Criteria	Performance levels		
	Detailed	Addressed but incomplete / unsatisfactory	Not addressed
1. Summary of digital outputs and technologies	Plan provides a clear description of the digital output or digital technology and the type of access envisaged e.g. 'freely available online' The summary explains how this will be achieved technically so reviewers can assess whether plans are appropriate.	There's a partial description for example an explanation of the data but not the access plans, or too little information given the scale and importance of the digital output. There is insufficient information to assess whether the overall plans are sound.	It is unclear what will be created or how. The proposal indicates that siginificant digital outputs will be created but the technical plan doesn't match this.
2a. Standards and formats	Relevant statistics are provided to explain the size, quantity and duration of the data. An explanation is given about which formats and standards will be used and why. This demonstrates an understanding about which formats are appropriate to each purpose e.g. to provide access or to archive data.	A data volume is stated, but it's difficult to interpret how many files that pertains to or how large and complex the dataset is. Formats and standards are named but it is not clear that these are understood and that choices have been made for valid reasons.	The description misses important information and doesn't give a clear sense of sounds decisions regarding the technical methodology.
2b. Hardware and software	It is clear what (if any) additional hardware and software will be required. The additional kit is clearly necessary and aligns with statements in the justification of resources. If nothing is required, the plan should clearly state this is not applicable.	The suggested resources seem unconvincing (e.g. items that should already provided by the institution or software that doesn't seem necessary for the proposed digital output).	The plan does not state that this section is not applicable.
2c.Technical methodology	There is a full description of the development process from data capture to delivery of a digital resource. Key issues such as quality control and backup are covered for all the research environments / circumstances e.g. in-house procedures and fieldwork. There is a clear timetable for technical delivery that relates to the overall project milestones in the case for support.	Certain information is missing e.g. how quality will be cheked if a number of assistants are transcribing and documenting resources. Backup is not explained for all circumstances in which data will be collected or stored. A timetable is missing, so it is unclear how the technical work aligns with the research activity.	There is no clear technical methodology or plan for how the data will be acquired, processed and used.

http://www.dcc.ac.uk/events/workshops/reviewing-data-management-plans

Step 2: Select a suitable repository

1. Domain specific repository
2. Institutional repository
3. General purpose repository such as Zenodo

Repositories are not created equal!

cf. https://doi.org/10.1371/journal.pone.0078080

Step 3: Deposit Data

WHAT?

- Data that validate the results presented in publications
- Other data: everything needed to find, asses, understand, reuse data:
  - “raw” data
  - derived data
  - code
  - Methods

Step 3: Deposit Data

PLANNED PREPARATION (with the DMP)

- Standard file formats
- Standards metadata schemas
- Documentation: what do others need to understand your data?
- (Open) Licences

Preparing data for sharing

Embedded documentation

- code, field and label description
- descriptive headers or summaries
- information recorded in the file properties (IPTC images, word doc)

Supporting documentation

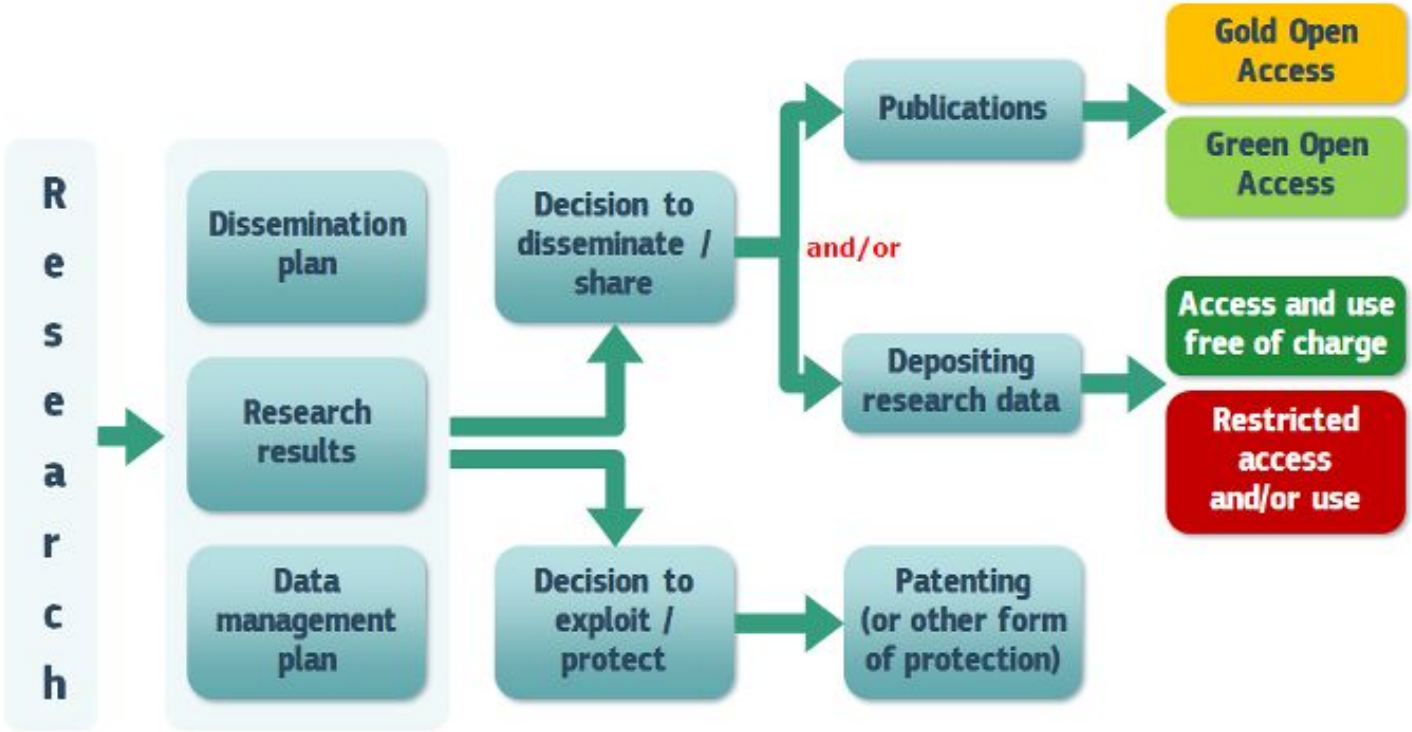
- Working papers, blog posts, field notebooks
- Questionnaires, interview guides
- Final project reports
- Catalogue metadata

Adapted from: https://www.slideshare.net/dancrane_open/data-sharing-how-what-and-why

Opening and sharing your data is not just about compliance: they are strategic decisions.

The ERC lets you decide what to share and how, but expects you to provide the necessary metadata and documentation that will make it most useful!

Decide early and record decisions in the DMP



Pragmatic advice

Prioritize

Data needed to validate results presented in publications

Data that can be re-used

Plan for sharing

Promote

Give visibility to the project with Open Data

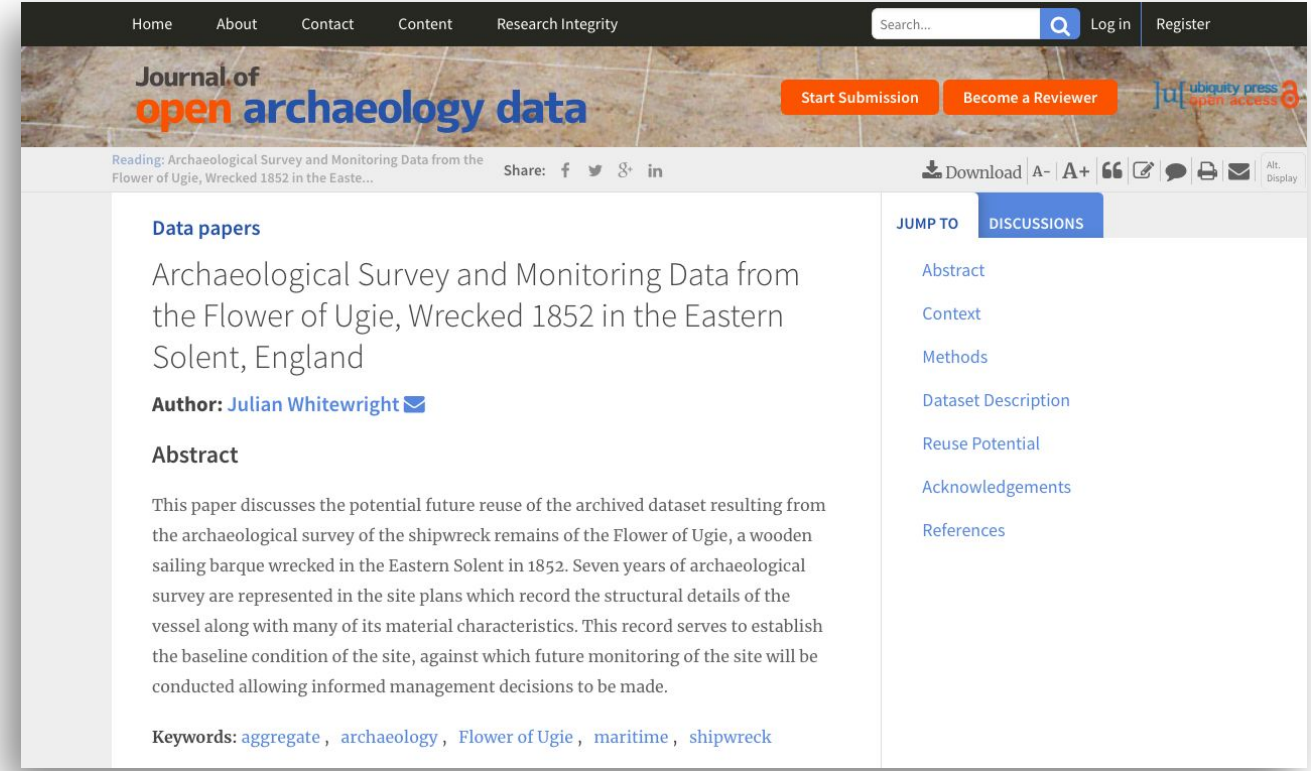
Commit to produce **data papers** for important dataset deposits (and retro-plan from that objective)

Collaborate

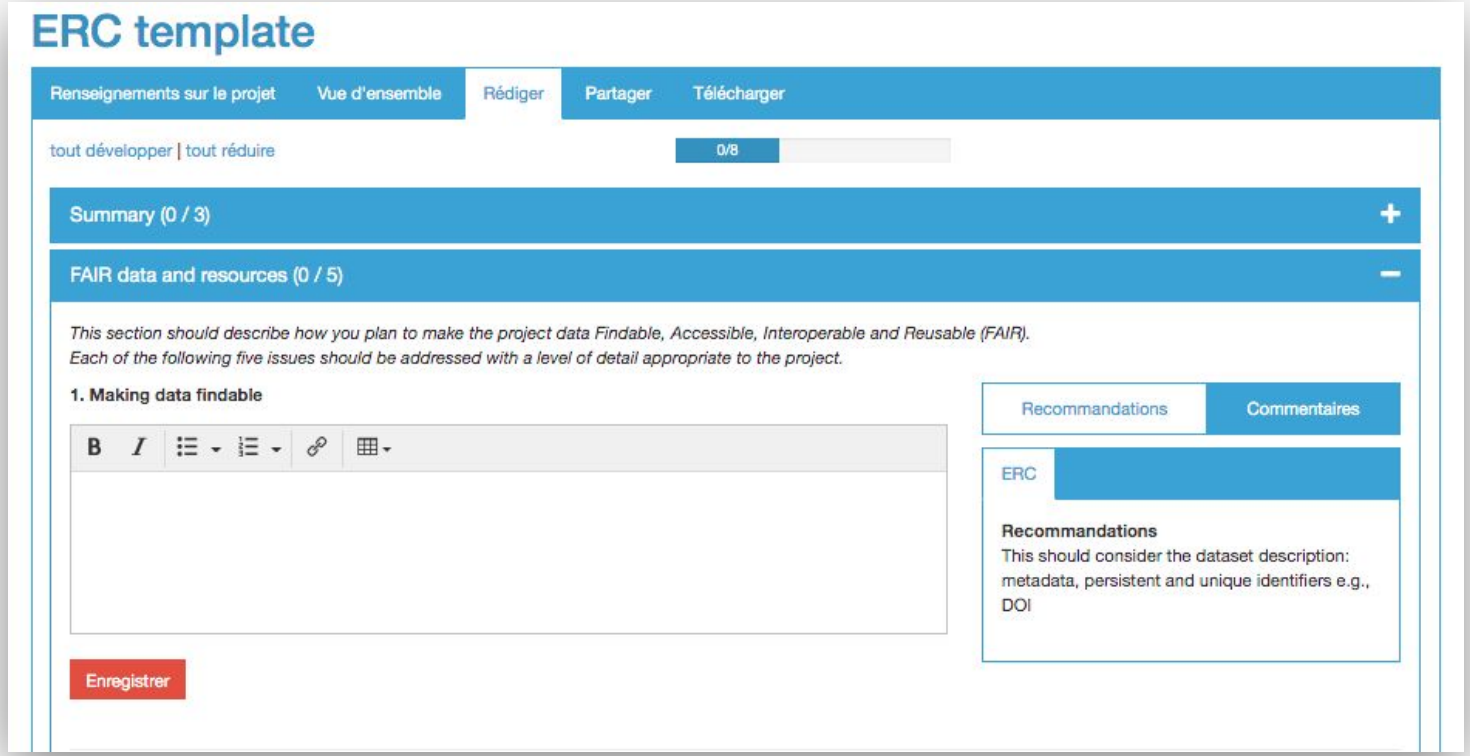
The DMP helps you think about **automatic data stewardship procedures**

Set up collective work sessions to select data for sharing/preservation define quality control steps

Datapapers Journal example



A collaborative tool for writing DMPs



Thanks!

Questions?

Credits

Presentation template by **Slidesgo**