

Open Science and Open Research Data: Requirements in H2020

Marisa Pérez Aliende
Universidad Autónoma de Madrid

uc3m | Universidad **Carlos III** de Madrid



Getafe, 14th November 2017



A close-up photograph of a red wooden sign hanging from a brass door handle. The sign is rectangular and has a slightly weathered appearance. It is suspended by a chain from a large, ornate brass handle. The background is a blue-painted door with visible wood grain and paneling. The lighting is bright, casting shadows on the door.

Welcome
WE ARE OPEN



Agenda

- 1** ■ Introduction to:
 - 1.1** ■ Open Science
 - 1.2** ■ Open Access
 - 1.3** ■ Open Data
- 2** ■ Horizon 2020, the Open Research Data Pilot
- 3** ■ How to manage and share data
 - 3.1** ■ Data management planning: Creating a DMP
 - 3.2** ■ Organizing files and data
 - 3.3** ■ Data sharing and archiving



Introduction to Open Science

Marisa Pérez Aliende
14th November 2017



Universidad
Carlos III de Madrid



What is Open Science?

“Science carried out and communicated in a manner which allows others to contribute, collaborate and add to the research effort, with all kind of data, results and protocols made freely available at different stages of the research process.”

Research Information Network

www.rin.ac.uk/our-work/data-management-and-curation/open-science-case-studies



What is Open Research?

Open Research is an interchangeable term with Open Science

“The idea that scientific knowledge of all kinds should be openly shared as early as it is practical in the discovery process.”



Why Open Research?

Open Research

embodies ideas of best research practice by opening access to results, data, protocols and other aspects of the research process. It also includes the use of open source software and open standards that offer unfettered dissemination of scientific discourse.

Reproducibility of research findings

One argument for Open Research is the findings of medical research are disseminated too slowly under the current system



Some benefits of Openness

- You can access relevant literature – not behind pay walls
- Ensures research is transparent and reproducible
- Increased visibility, usage and impact of your work
- New collaborations and research partnerships
- Ensure long-term access to your outputs
- Help increase the efficiency of research

Jones, Sarah “The benefits & practice of openness”

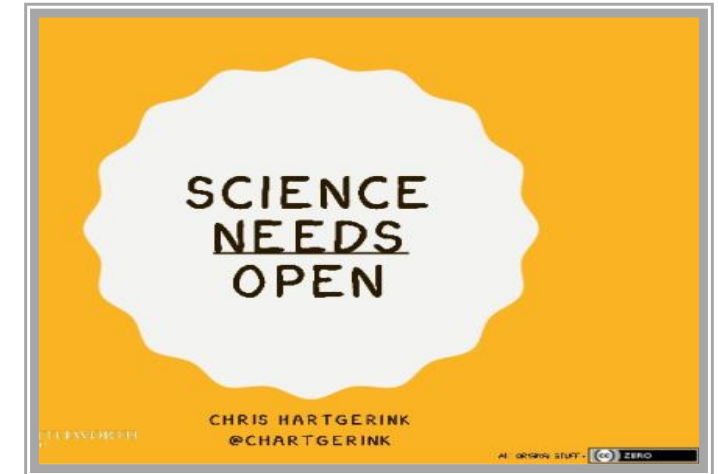
We can add:

<https://www.fosteropenscience.eu/content/benefits-openness-0>

- Increases the chances of attaining research grant
- Helps to reduce time spent finding/accessing material
- Cut down on academic fraud

(<https://www.economist.com/news/china/21586845-flawed-system-judging-research-leading-academic-fraud-looks-good-paper>)

- The research cycle acceleration
- Increase citations



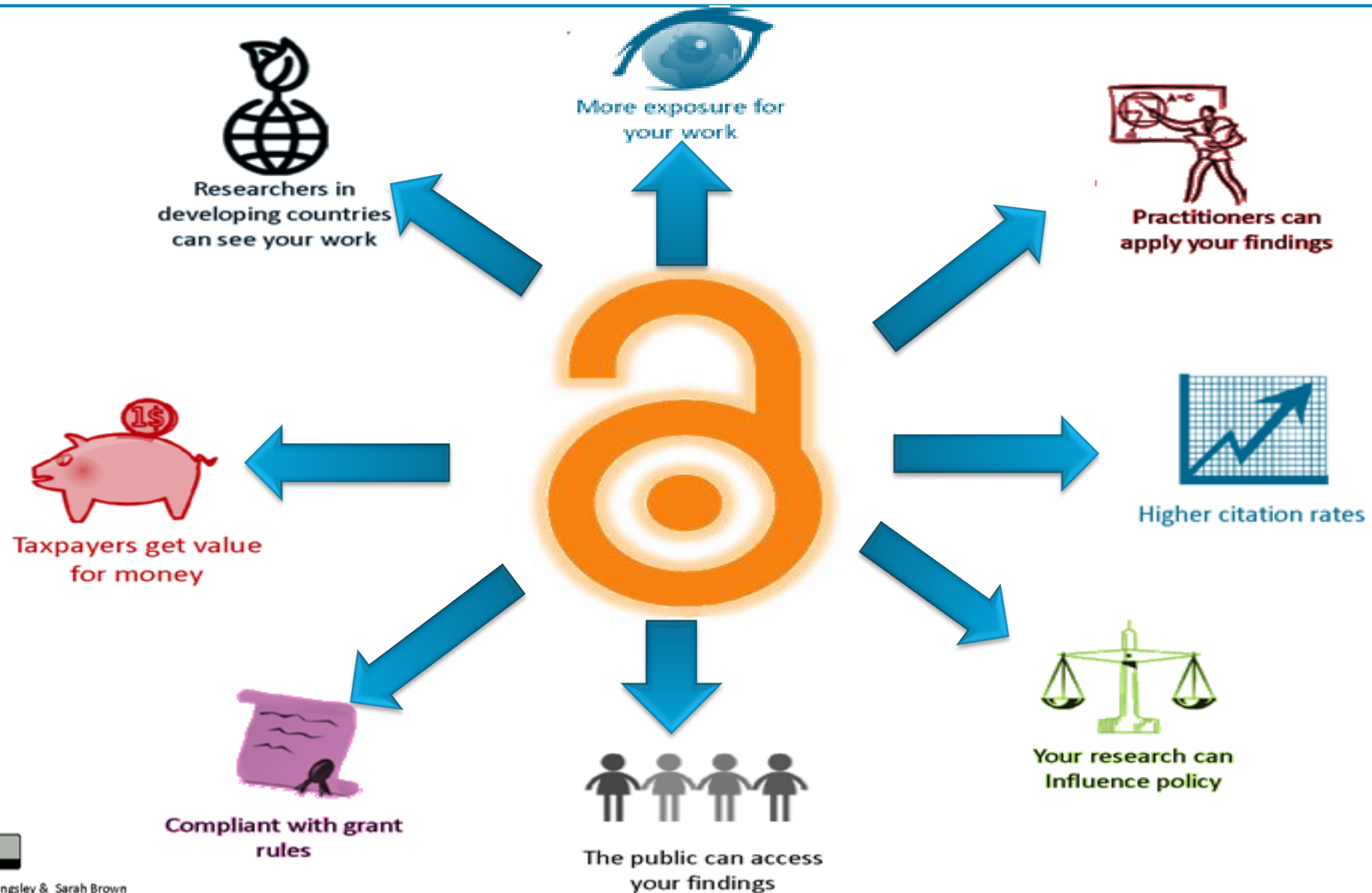
YES: Open promotes diversity of opinion
 YES: Open promotes disagreement
 YES: Open increases criticism
 YES: From open. We can learn

Hartgerink, Chris “Science needs open”





Benefits of Openness





Introduction to Open Access

Marisa Pérez Aliende
14th November 2017



Universidad
Carlos III de Madrid



Why Open Access?

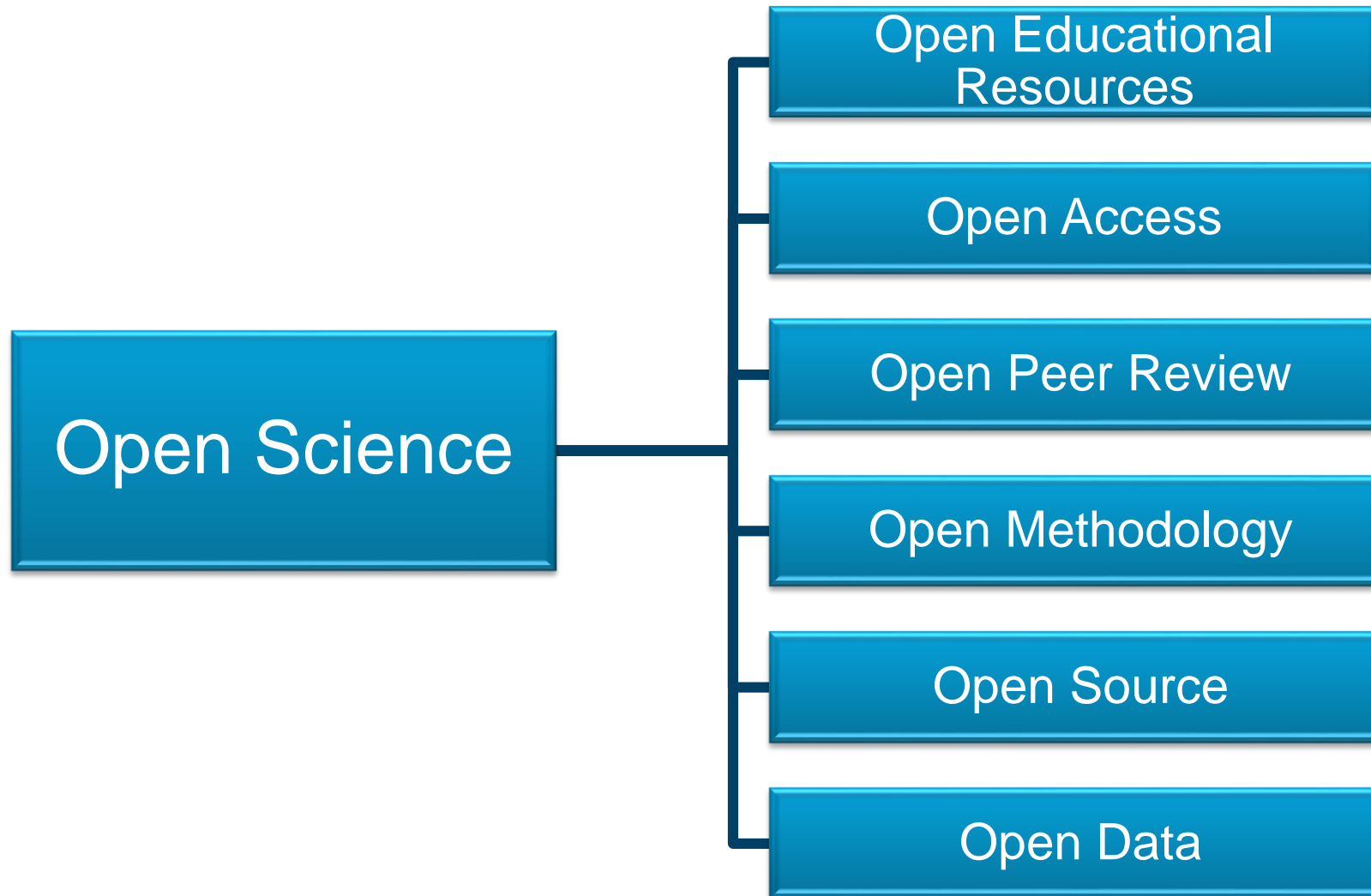
A light blue, rounded rectangular graphic with a slight 3D effect and a shadow. Inside the graphic, the words "OPEN ACCESS" are written in a stylized, bold, italicized font. "OPEN" is in blue with a black outline, and "ACCESS" is in orange with a black outline.

OPEN ACCESS

Centrum Cyfrowe "Open Access, why Open Science?"
<https://www.youtube.com/watch?v=0GbXjWLKqG0>



Not just Open Access Publishing



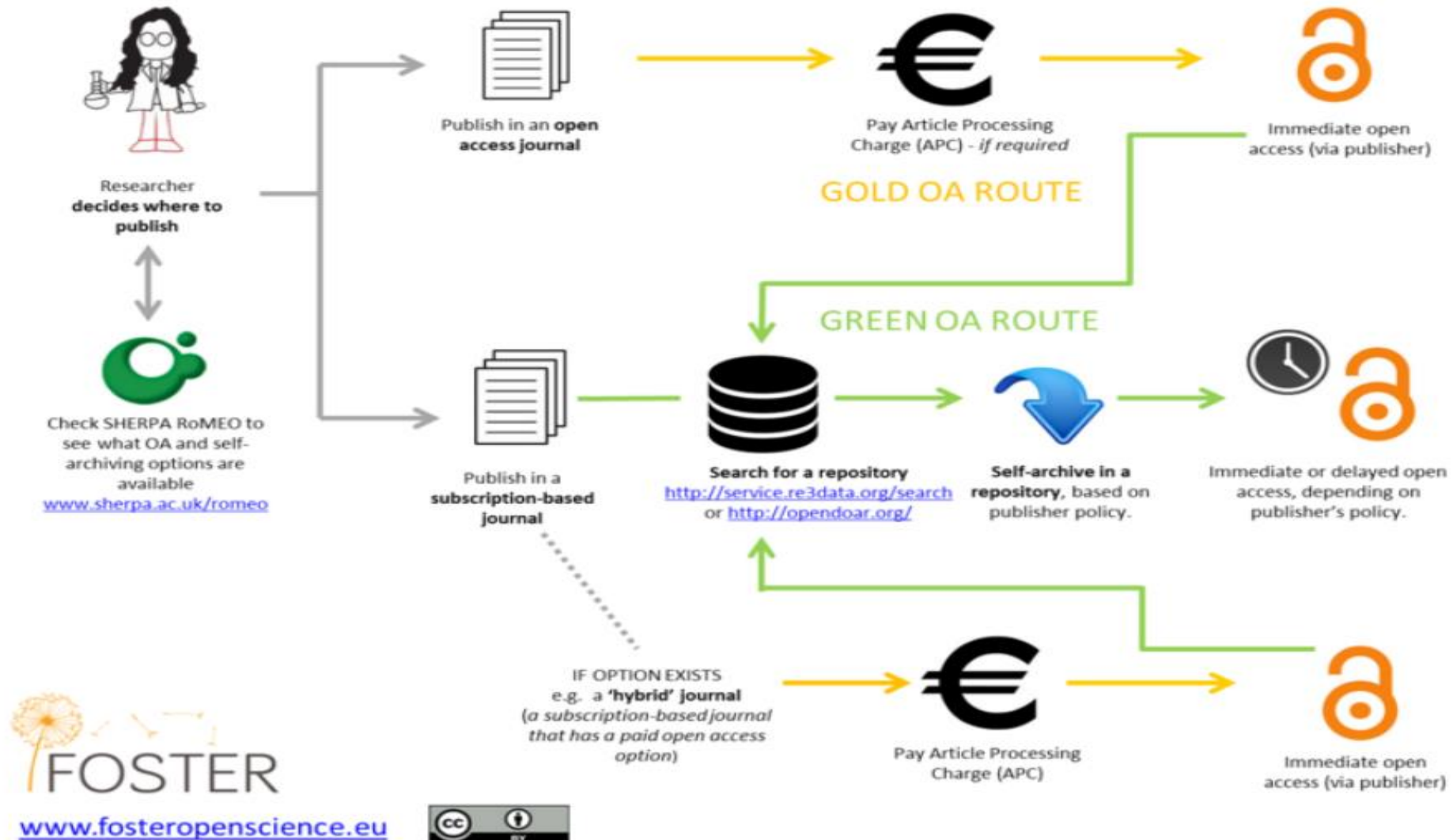
Why Open Access?



Open Access Explained

<http://www.youtube.com/watch?v=L5rVH1KGBCY>

Routes to Open Access





50 shades of open access

Type	Subtype	Who pays?	Example
Gold	"Diamond"	Institution (subsidy)	
Gold	Gold, not for profit	Author (fee)	Int. J. of the Commons
Gold	Gold, for profit	Author (fee)	PLoS
Gold	Hybrid gold, for profit ✓	Author (fee) + Library (subscription)	In Nature
Green	Last author version in repository (embargo's)	Library (subscription)	In Igitur
Green	Pre- prints	Library (subscription)	ArXiv / SSRN / Peer J preprints
Green	Working papers	Working paper archive (institutional subsidy)	In RepEc
Green	"Black" (sharing against copyright)	Publisher	Via Academica



Open Access publishing: How / When publishing a journal article...

1. Consider carefully where to publish your work: Choose an Open Access Journal

CRITERIA

Sometimes, the best place for your article is a traditional subscription based journal. You may still have options for making your article openly available:

1. Publish in a hybrid open access journal
2. Deposit the article in a repository:

- Check the Publisher's Policies

SHERPA ROMEO

- Retain your rights

NEGOCIATE!!!

- Deposit the article in a repository



Open Access publishing: How / When publishing a journal article...

Criteria to consider when evaluating the quality of open journals:

- Any help?:
 - [Think, Check Submit](#) : A check list to make sure you choose trusted journals for your research
 - Directory of Open Access Journals ([DOAJ](#))
 - Check if the publisher is a member of:
 - Open Access Scholarly Publishers' Association ([OASPA](#))
 - Committee on Publication Ethics ([COPE](#))
 - [Principles of transparency and best practice in scholarly publishing](#):
 - Peer review process
 - Governing body
 - Editorial team/contact information
 - Author fees
 - Copyright
 - Process for identification of and dealing with allegations of research misconduct
 - Ownership and management
 - Web site
 - Name of journal
 - Conflict of interests ...
- **Ask your librarian**

When publishing OA books, check OA publishers:

- *Open Book Publishers (social sciences & humanities)*
- *INTECH open science open minds (science, technology & medicine)*





Open Access publishing: How / When publishing a journal article...

Predatory publishers

There is no one standard definition of what constitutes a predatory publisher but generally they are those publishers who charge a fee for the publication of material without providing the publication services an author would expect such as peer review and editing. Missing out on these important steps can undermine the final product and [perpetuates bad research](#) in general and exploits the Open Access publishing model.

University of Cambridge

<https://osc.cam.ac.uk/about-scholarly-communication/author-tools/considerations-when-choosing-journal/predatory-publishers>



How to spot a PREDATORY PUBLISHER

Claire Sewell
Office of Scholarly Communication
ces43@cam.ac.uk

Sewell, Claire "How to spot a predatory publisher"
<https://www.youtube.com/watch?v=He9GJybTtUw&feature>



Open Access publishing: How / When publishing a journal article...

Fake Journals and Scam Conferences

“Predatory publishers present themselves as legitimate journals, claiming to provide peer review and editorial services. In truth, they will publish nearly anything that is sent to them (and paid for) with little to no quality control, as evidenced by the numerous [examples](#) of deliberately false papers submitted—and successfully published—by journalists and academics hoping to prove a point. This lack of quality control allows poorly conducted and outright false research to be published, sometimes leading to greater consequences down the road as these papers are cited or used as the basis for further studies.”

Zimmerman, Erin “How to keep your research out of fake journals and scam conferences”

<http://www.cdnsiencepub.com/blog/how-to-keep-your-research-out-of-fake-journals-and-scam-conferences.aspx>

**The publisher
send e-mail
poorly written**

**Lower fees than
author-pays
journal**

**Not indexed in
databases
(PubMed, WoS)**

**False editorials
boards**

**Lack of quality
control**



Open Access publishing: How / When publishing a journal article...

1. Consider carefully where to publish your work: Choose an Open Access Journal

CRITERIA

Sometimes, the best place for your article is a traditional subscription based journal. You may still have options for making your article openly available:

1. Publish in a hybrid open access journal
2. Deposit the article in a repository:

- Check the Publisher's Policies

SHERPA ROMEO

- Retain your rights

NEGOCIATE!!!

- Deposit the article in a repository



Open Access publishing: How / When publishing a journal article...

SHERPA/RoMEO ... opening access to research

Home • Search • Journals • Publishers • FAQ • Suggest • About

English | Español | Magyar | Nederlands | Português

Publisher copyright policies & self-archiving

Search

Journal titles or ISSNs Publisher names

Exact title starts with contains ISSN

[Advanced Search](#) [Search](#) [Reset](#)

Use this site to find a summary of permissions that are normally given as part of each publisher's copyright transfer agreement.

Special RoMEO Pages

- RoMEO Statistics
- Application Programmers' Interface (API)
- Publisher Categories in RoMEO
- Definitions and Terms

Additions and Updates [RSS1 Feed](#)

- Open Journals - Open Journals - 31-Oct-2017
- Columbia University, Teachers College - Columbia University, Teachers College - 30-Oct-2017
- University of Lincoln, Doctoral School - University of Lincoln, Doctoral School - 30-Oct-2017

Other SHERPA Services

- SHERPA/FACT - Funders & Authors Compliance Tool
- SHERPA/JULIET - Research funders' open access policies

Jisc

This work is licensed under [CC BY-NC-ND](#). [About using our content](#) [Contact us](#)

Journal:	Advances in Surgery (ISSN: 0065-3411)
RoMEO:	This is a RoMEO green journal
Paid OA:	A paid open access option is available for this journal.
Author's Pre-print:	✓ author can archive pre-print (ie pre-refereeing)
Author's Post-print:	✓ author can archive post-print (ie final draft post-refereeing)
Publisher's Version/PDF:	✗ author cannot archive publisher's version/PDF
General Conditions:	<ul style="list-style-type: none"> Authors pre-print on any website, including arXiv and RePEC Author's post-print on author's personal website immediately Author's post-print on open access repository after an embargo period of between 12 months Permitted deposit due to Funding Body, Institutional and Governmental policy or mandate, may be required to comply with embargo periods of 12 months Author's post-print may be used to update arXiv and RePEC Publisher's version/PDF cannot be used Must link to publisher version with DOI Author's post-print must be released with a Creative Commons Attribution Non-Commercial No Derivatives License
Mandated OA:	Compliance data is available for 57 funders
Paid Open Access:	Open Access
Copyright:	Unleashing the power of academic sharing - Sharing Policy - Sharing and Hosting Policy FAQ - Green open access - Journal Embargo Period List (pdf) - Journal Embargo List for UK Authors - Attaching a User License (pdf)
Updated:	01-Jul-2016 - Suggest an update for this record
Link to this page:	http://www.sherpa.ac.uk/romeo/issn/0065-3411/
Published by:	Elsevier : 12 months [Commercial Publisher] - Green Policies in RoMEO
For:	Mosby [Imprint]
Guidance:	Please see the list of Publisher Categories in RoMEO for guidance on interpreting the priority of multiple publishers.
<p>These summaries are for the journal's <i>default</i> policies, and changes or exceptions can often be negotiated by authors. <i>All information is correct to the best of our knowledge but should not be relied upon for legal advice.</i></p>	



Open Access publishing: How / When publishing a journal article...

Understanding your rights:

- Pre-print
- Post-print
- Publisher version

RoMEO colours archiving policies:

ROMEIO colour	Archiving policy
green	can archive pre-print <i>and</i> post-print or publisher's version/PDF
blue	can archive post-print (ie final draft post-refereeing) or publisher's version/PDF
yellow	can archive pre-print (ie pre-refereeing)
white	archiving not formally supported





Open Access publishing: How / When publishing a journal article...

Retain your rights



Publishers are becoming more friendly towards author rights, due to funder mandates for Open Access.

Publishers' friendly agreements

- [Nature](#)
- Public Library of Science ([PLoS](#))

Author's Addendum
(attach to a journal publisher's copyright agreement)

- Scholar's Copyright Addendum Engine ([SCAE](#))
- SPARC Author Addendum



*When publishing a **book** there are also options available when negotiating the author agreement. If there is not OA publishing options, negotiate to deposit in an institutional or subject repository, with an embargo period.*



Open Access publishing: How / When publishing a journal article...

License your work!

Creative Commons licenses grant the copyright of a work to its rightful author, indicating to any third party what they can or cannot do with it, without giving up all control.

LICENSES	TERMS
	Attribution BY Others can copy, distribute, display, perform and remix your work if they credit your name as requested by you
	No Derivative Works ND Others can only copy, distribute, display or perform verbatim copies of your work
	Share Alike SA Others can distribute your work only under a license identical to the one you have chosen for your work
	Non-Commercial NC Others can copy, distribute, display, perform or remix your work but for non-commercial purposes only.

	Can someone use it commercially?	Can someone create new versions of it?
Attribution		
Share Alike		Yup, AND they must license the new work under a Share Alike license.
No Derivatives		
Non-Commercial		Yup, AND the new work must be non-commercial, but it can be under any non-commercial license.
Non-Commercial Share Alike		Yup, AND they must license the new work under a Non-Commercial Share Alike license.
Non-Commercial No Derivatives		



Open Access publishing: How / When publishing a journal article...

Deposit your article in a repository

Once you've ensured that you have the right to make your article available through a repository, you can deposit it in your institution's repository, a subject repository, or both!

- Speak to your University Library and deposit in your IR:

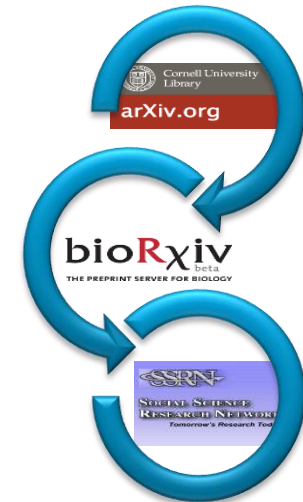
Universidad Carlos III de Madrid
Biblioteca -Archivo

- Identify a repository in your field:

- The Directory of Open Access Repositories

- Registry of Open Access Repositories

ROAR
Registry of
Open Access
Repositories



BOOKS: these Institutional or subjects repositories are valid for books.

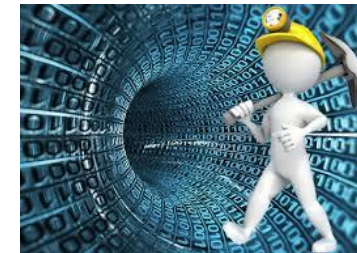
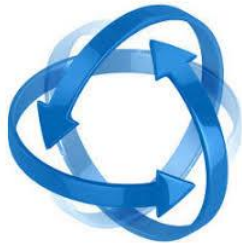
But, in this case, the publisher can use **The Open Access Publishing in European Networks (OAPEN)**, a central repository for hosting and disseminating OA books in humanities and social sciences.

Open Access to Publications

- Free, immediate online access to the results of research



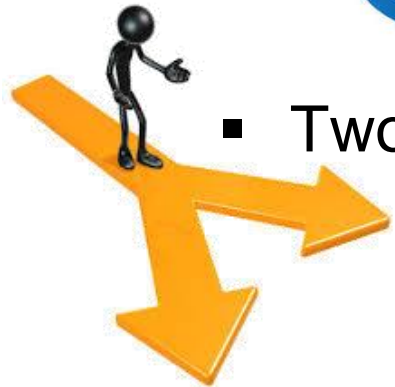
- Free to reuse, e.g. to build tools to mine the content



- Two routes to make sure anyone can access your papers:

- Gold route: paying APC to ensure publishers makes copy open
 - Green route: self-archiving Open Access copy in repository

- Find out what your publisher allows on [SHERPA ROMEO](https://www.sherpa.ac.uk/romeo/)





Introduction to Open Data

Marisa Pérez Aliende
14th November 2017



Universidad
Carlos III de Madrid



Open Data

WHAT IS OPEN?

OPEN KNOWLEDGE' IS ANY CONTENT, INFORMATION OR DATA THAT PEOPLE ARE FREE TO USE, RE-USE AND REDISTRIBUTE — WITHOUT ANY LEGAL, TECHNOLOGICAL OR SOCIAL RESTRICTION

“ Open knowledge is what open data becomes when it's useful, usable and used.”



<https://okfn.org/opendata/>

“Open data and content can be freely used, modified and shared by anyone for any purpose”

<http://opendefinition.org>



What is Research Data?

“Research data, unlike other types of information, is collected, observed, or created, for purposes of analysis to produce original research results”

University of Edinburg

“Recorded factual material commonly accepted in the scientific community as necessary to document and support research findings.”

National Medical Research Council. Singapore



What is Research Data?

“Data are facts, observations or experiences on which an argument, theory or test is based. Data may be numerical, descriptive or visual. Data may be raw or analysed, experimental or observational. Data includes: laboratory notebooks; field notebooks; primary research data (including research data in hardcopy or in computer readable form); questionnaires; audiotapes; videotapes; models; photographs; films; test responses. Research collections may include slides; artefacts; specimens; samples. Provenance information about the data might also be included: the how, when, where it was collected and with what (for example, instrument). The software code used to generate, annotate or analyse the data may also be included.”



What is Research Data?

Tim Berners-Lee proposal for five star open data:

<http://5stardata.info>

★	make your stuff available on the Web (whatever format) under an open licence	Example
★ ★	make it available as structured data (e.g. Excel instead of a scan of a table)	Example
★ ★ ★	use non-proprietary formats (e.g. CSV instead of Excel)	Example
★ ★ ★ ★	use URIs to denote things, so that people can point at your stuff	Example
★ ★ ★ ★ ★	link your data to other data to provide context	Example

Research Data?

- All researchers work with data



but what you call data will depend on your discipline.



- As a humanities scholar you might talk about your primary resources or texts.



- In Social Science, you may think in terms of survey results, interviews and statistics.



- In Science, different terms for the output of your experiments and observations





Research Data can be

Qualitative
or
Quantitative

Print, digital,
and physical
formats

Existing data

Collecting or
creating new
data

Research data needs to be cared, so the results
of your research can be validated and built upon



Classification of Research Data

- **Observational:** data captured in real-time, usually irreplaceable. For example, sensor data, survey data, sample data, neuro-images.
- **Experimental:** data from lab equipment, often reproducible, but can be expensive. For example, gene sequences, chromatograms, toroid magnetic field data.
- **Simulation:** data generated from test models where model and metadata are more important than output data. For example, climate models, economic models.
- **Derived or Compiled:** data is reproducible but expensive. For example, text and data mining, compiled database, 3D models.
- **Reference or Canonical:** a (static or organic) conglomeration or collection of smaller (peerreviewed) datasets, most probably published and curated. For example, gene sequence databanks, chemical structures, or spatial data portals.



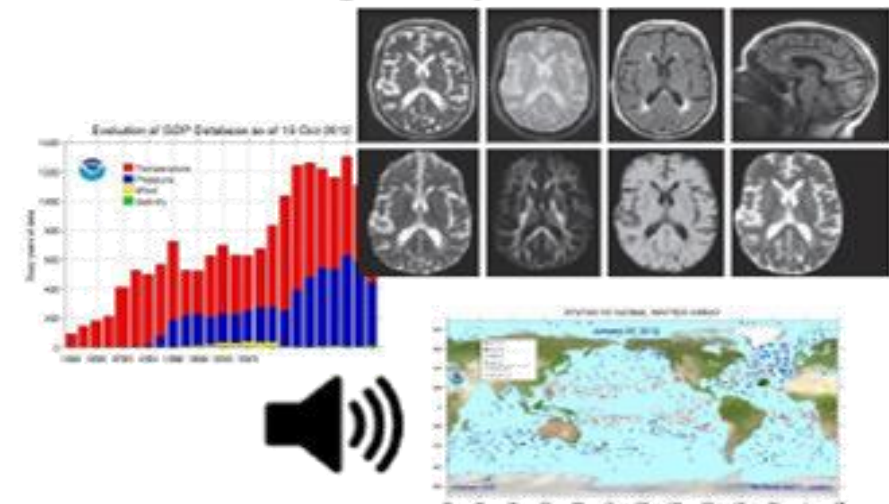
Research Data Formats

- Text - flat text files, Word, Portable Document Format (PDF), Rich Text Format (RTF), Extensible Mark-up Language (XML).
- Numerical - Statistical Package for the Social Sciences (SPSS), Stata, Excel.
- Multimedia - jpeg, tiff, dicom, mpeg, quicktime.
- Models - 3D, statistical. • Software - Java, C.
- Discipline specific - Flexible Image Transport System (FITS) in astronomy, Crystallographic Information File (CIF) in chemistry.
- Instrument specific - Olympus Confocal Microscope Data Format, Carl Zeiss Digital Microscopic Image Format (ZVI).



Research Data (traditional & electronic) may include all the following

- Documents (text, Word), spreadsheets
- Laboratory notebooks, field notebooks, diaries
- Questionnaires, transcripts, codebooks
- Audiotapes, videotapes
- Photographs, films
- Test responses
- Slides, artefacts, specimens, samples
- Collection of digital objects acquired and generated during the process of research
- Data files
- Database contents (video, audio, text, images)
- Models, algorithms, scripts
- Contents of an application (input, output, logfiles for analysis software, simulation software, schemas)
- Methodologies and workflows
- Standard operating procedures and protocols





Research Data

The following research records may also be important to manage during and beyond the life of a project:

- Correspondence (electronic mail and paper-based correspondence)
- Project files
- Grant applications
- Ethics applications
- Technical reports
- Research reports
- Master lists
- Signed consent forms

University of Edinburgh

<https://www.ed.ac.uk/information-services/research-support/research-data-service>



Open Data in Science

- Science is based on building on, reusing and openly criticising the published body of scientific knowledge.
- For science to effectively function it is crucial that science data be made open.
- By open data in science we mean that it is freely available on the public internet permitting any user to download, copy, analyse, re-process, pass them to software or use them for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself.





Panton Principles: *Principles for Open Data in Science*

1. When publishing data make an explicit and robust statement of your wishes.
2. Use a recognized waiver or license that is appropriate for data.
3. If you want your data to be effectively used and added to by others it should be open as defined by the [Open Knowledge/Data Definition](#) – in particular non-commercial and other restrictive clauses should not be used.
4. Explicit dedication of data underlying published science into the public domain via PDDL or CCZero is strongly recommended and ensures compliance with both the Science Commons [Protocol for Implementing Open Access Data](#) and the [Open Knowledge/Data Definition](#).

Panton Principles, Principles for open data in science. Murray-Rust, Peter; Neylon, Cameron; Pollock, Rufus; Wilbanks, John; (19 Feb 2010). Retrieved [23/10/2017] from

<https://pantonprinciples.org/>



Long Tail Research Data

BIG DATA

It is the data of big science that are shared among the great teams. Scientists often have sophisticated infrastructures that help them manage those volumes of data. Well organized, frequently used and cited.

SMALL SCIENCE / LITTLE SCIENCE

The long tail of research data:

It subsumes large portions of data that are highly heterogeneous, managed predominantly locally within each researcher's environment and frequently not properly transferred to and managed within well-curated repositories.

Proll, Meixner y Rauber "Precise data identification services for Long Tail Research Data". 2017

Big Data	Long-Tail Data
Homogeneous	Heterogeneous
Large	Small
Common standards	Unique standards
Integrated	Not-integrated
Central curation	Individual curation
Disciplinary repositories	Institutional, general or no repositories

The background is a deep blue gradient. A glowing horizon line curves across the middle. Above the horizon, a small globe icon is positioned where the letter 'O' in 'HORIZON' would be. Radiant light beams emanate from behind the globe icon, creating a lens flare effect. The text 'HORIZON 2020' is written in a white, sans-serif font, with the globe icon acting as the letter 'O'.

HORIZON 2020

The Open Research Data Pilot



What is Horizon 2020?

Horizon 2020 is the biggest EU Research and Innovation programme ever with nearly €80 billion of funding available over 7 years (2014 to 2020) – in addition to the private investment that this money will attract. It promises more breakthroughs, discoveries and world-firsts by taking great ideas from the lab to the market.

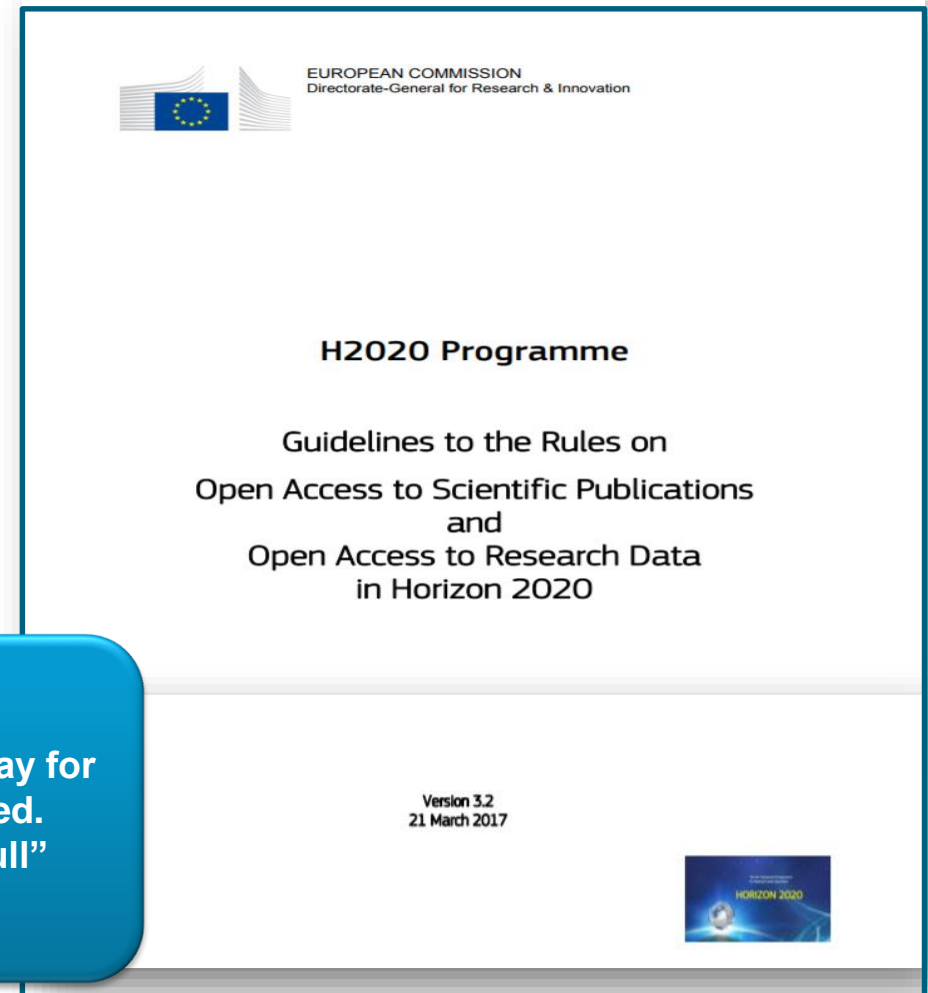
<https://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020>

Why Open Access to publications and Open Data in H2020?

Broader access to scientific outputs helps to:

- **build on previous research results** (improved quality of results)
- **encourage collaboration and avoid duplication of effort** (greater efficiency)
- **speed up innovation** (faster progress to market means faster growth)
- **involve citizens and society** (improved transparency of the scientific process)

“The European Commission considers that there should be no need to pay for information funded from the public purse each time it is accessed or used. Moreover, it should benefit European businesses and the public to the full”





Why Open Access to publications and Open Data in H2020?

Version 4.1 - 26 October 2017

ARTICLE 29 — DISSEMINATION OF RESULTS — OPEN ACCESS — VISIBILITY OF EU FUNDING

29.1 Obligation to disseminate results

Unless it goes against their legitimate interests, each beneficiary must — as soon as possible — ‘**disseminate**’ its results by disclosing them to the public by appropriate means (other than those resulting from protecting or exploiting the results), including in scientific publications (in any medium).

This does not change the **obligation to protect results** in Article 27, the confidentiality obligations in Article 36, the security obligations in Article 37 or the obligations to protect personal data in Article 39, all of which still apply.

29.2 Open access to scientific publications

Each beneficiary must ensure open access (free of charge, online access for any user) to all peer-reviewed scientific publications relating to its results.

In particular, it must:

- (a) as soon as possible and at the latest on publication, deposit a machine-readable electronic copy of the published version or final peer-reviewed manuscript accepted for publication in a repository for scientific publications;

Moreover, the beneficiary must aim to deposit at the same time the research data needed to validate the results presented in the deposited scientific publications.

- (b) ensure open access to the deposited publication — via the repository — at the latest:
 - (i) on publication, if an electronic version is available for free via the publisher, or
 - (ii) within six months of publication (twelve months for publications in the social sciences and humanities) in any other case.
- (c) ensure open access — via the repository — to the bibliographic metadata that identify the deposited publication.

The bibliographic metadata must be in a standard format and must include all of the following:



Obligation to:

- Protect results (art. 27)
- Confidentiality (art. 36)
- Security (art.37)
- Protect personal data (art. 39)



Model amendment to publishing agreements

This model is not mandatory but reflects the obligations for the beneficiary under the H2020 grant agreements

OPEN ACCESS PUBLISHING AGREEMENT

> Instructions and footnotes in blue should be deleted.
 > For options [in square brackets]: choose the applicable option. Options not chosen should be deleted.
 > For fields in [grey in square brackets]: enter the appropriate data.

ADDENDUM

(To be filled out by the beneficiary/author and the publisher. This model is not mandatory but reflects the obligations for the beneficiary under the H2020 grant agreements. It can be supplemented by further provisions agreed between the parties, provided they are compatible with the Grant Agreement. The Commission/Agency takes no responsibility for the use of this model.)

This 'Addendum' is **between** the following parties:

on the one part,

1. the publisher

[full official name (short name)], established in [official address in full], represented by [....].

and

on the other part,

1. 'the corresponding author':

[full name], [official address in full], represented by [....]

and the following **other authors**

2. [full name], [official address in full], represented by [....]

3. [full name], [official address in full], represented by [....]

[same for each author].

With this Addendum, the **parties agree to complement and amend** the attached Publication Agreement concerning the publication [insert name of publication] in the Journal [insert name of journal] with the following open access clause:

Open access

The author(s) retain(s) the right to:

FP7 template Open Access Publishing Agreement 5.0.0-1 20.03.2017

a) deposit a machine-readable electronic copy of the published version or the final manuscript (after peer review) in an institutional, centralised and/or subject-based repository;

b) provide open access (i.e. free-of-charge access to the electronic copy to anyone) through this repository:

(i) immediately, if the publication itself is published 'open access' (i.e. if an electronic version is also available free of charge to the reader via the publisher) or

(ii) within 12/6/3 months after publication.

In case of conflicting provisions, this Addendum takes precedence over the Publication Agreement.

All other provisions of the Publishing Agreement remain unchanged.

This Addendum enters into force on the day of the last signature.

[OPTION if addendum signed after publication: It takes effect on (insert publication date).]

SIGNATURES

For the authors:

name date signature

name date signature

name date signature

For the publisher:

date signature stamp

Done in two originals, in English

History of changes		
Version	Publication date	Change
1.0	20.03.2017	1. Initial version

1. Choose 12 months for publications in the social sciences and humanities and 6 months for publications in other domains

Choose 12 months for publications in the social sciences and humanities and 6 months for publications in other domains



Horizon2020: Annotated Model Grant Agreement (AGA)

29.3 Open access to research data

[OPTION 1 for actions participating in the open Research Data Pilot: Regarding the digital research data generated in the action ('data'), the beneficiaries must:

- (a) deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:
 - (i) the data, including associated metadata **needed to validate the results** presented in scientific publications as soon as possible;
 - (ii) **other data, including associated metadata**, as specified and within the deadlines laid down in the **'data management plan'** (see Annex I);
- (b) provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).

This does not change the obligation to protect results in Article 27, the confidentiality obligations in Article 36, the security obligations in Article 37 or the obligations to protect personal data in Article 39, all of which still apply.

As an exception, the beneficiaries do not have to ensure open access to specific parts of their research data if the achievement of the action's main objective, as described in Annex I, would be jeopardised by making those specific parts of the research data openly accessible. In this case, the data management plan must contain the reasons for not giving access.]

Doesn't apply to all data (researchers to define as appropriate)
 Don't have to share data if inappropriate – exemptions apply
 Data underlying publications should be shared as soon as possible. Other data to be shared in timeframe specified in the DMP

Jones, Sarah "Horizon 2020 Open Research Data Pilot: What is required"
<https://www.fosteropenscience.eu/sites/default/files/pdf/2289.pdf>



H2020 Open Research Data (ORD) Pilot

Open access to research data refers to the right to access and re-use research data.

'Research data' refers to information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images.

The focus is on research data that is available in digital form.

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

The ORD pilot aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects and takes into account the need to balance openness and protection of scientific information, commercialisation and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions.

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



H2020 areas participating in the Pilot

Areas participating in the pilot
from 2014 – 2016.

Projects in other areas can
participate on a voluntary basis.

Future and Emerging Technologies

Research infrastructures – part e-Infrastructures

Leadership in enabling and industrial technologies – Information and
Communication Technologies

Societal Challenge: 'Secure, Clean and Efficient Energy' – part Smart
cities and communities

Societal Challenge: 'Climate Action, Environment, Resource
Efficiency and Raw materials' – except raw materials

Societal Challenge: 'Europe in a changing world – inclusive, innovative
and reflective Societies'

Science with and for Society

From January 2017 participating in the ORD pilot will be the default option (is extended to cover all thematic areas of Horizon 2020), but retaining opt-out possibilities.



ORD reasons for opting-out ?

The beneficiaries may **opt-out** of the pilot at any stage
(at the proposal **or** during the lifetime of a project, partially (selected datasets) **or** entirely)

Participation in the Open Research Data Pilot is **not** part of the project evaluation

- Participation is incompatible with:
 - The obligation to protect results that are expected to be commercially or industrially exploited.
 - The need for confidentiality in connection with security issues.
 - Rules on protecting personal data.
- Participation would mean that the project's main aim might not be achieved.
- The project will not generate / collect any research data.
- There are other legitimate reasons .

These issues should be described in the project Data Management Plan

Not all data can be open

The Commission's approach
"As open as possible, as closed as necessary"



H2020: FAIR data principles

Beneficiaries make their research data FAIR:

Findable

- (meta)data are assigned a globally unique and persistent identifier
- data are described with rich metadata (defined by R1 below)
- metadata clearly and explicitly include the identifier of the data it describes
- (meta)data are registered or indexed in a searchable resource

Accessible

- (meta)data are retrievable by their identifier using a standardized communications protocol
 - the protocol is open, free, and universally implementable
 - the protocol allows for an authentication and authorization procedure, where necessary
- metadata are accessible, even when the data are no longer available

Interoperable

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

Reusable

- meta(data) are richly described with a plurality of accurate and relevant attributes
 - (meta)data are released with a clear and accessible data usage license
 - meta(data) are associated with detailed provenance
 - (meta)data meet domain-relevant community standards



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



<https://www.force11.org/group/fairgroup/fairprinciples>



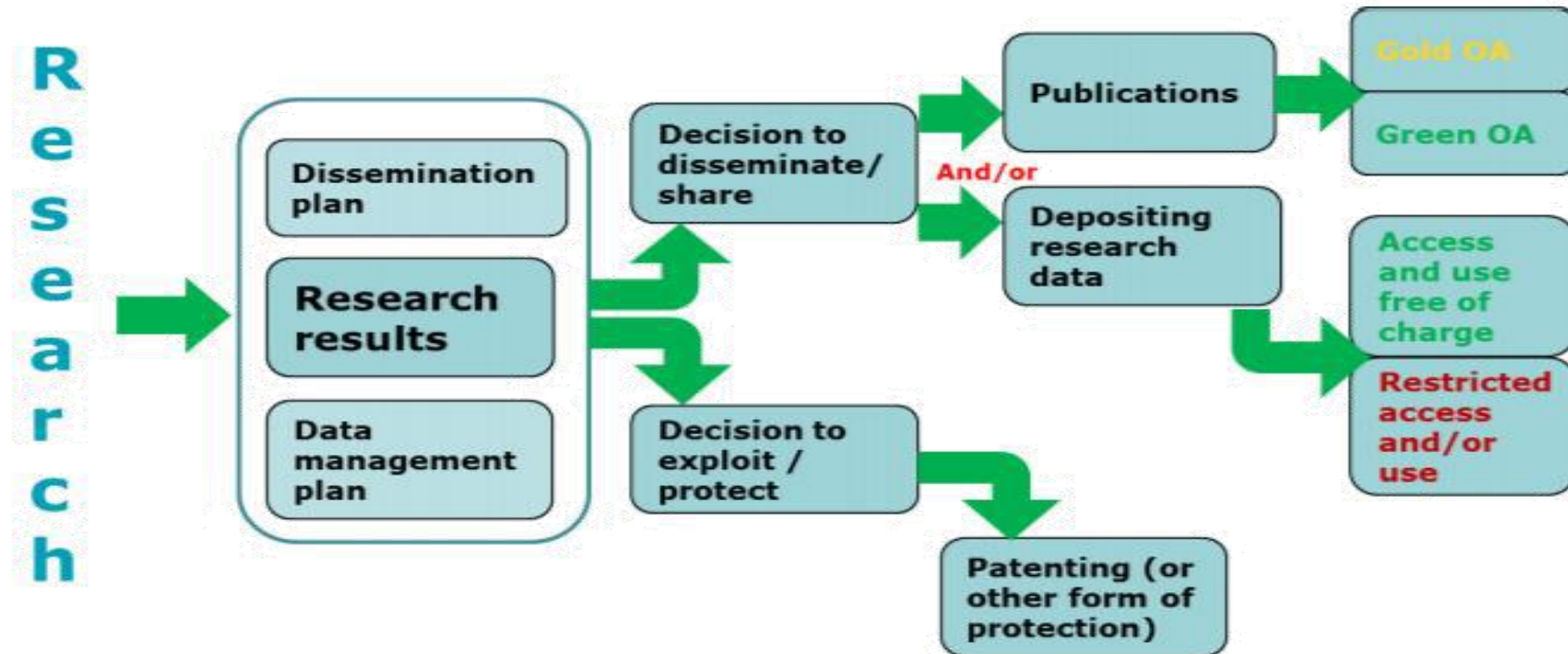
Key requirements of the ORD Pilot

Beneficiaries participating in the ORD pilot will:

- Deposit data in a research data repository
- Take measures to enable third parties to access, mine, exploit, reproduce and disseminate (free or charge for any user) this research data
- Provide information via the chosen repository about tools and instruments necessary for validating the results (where possible provide the tools and instruments themselves)



Open Access to Research Data



Graph: Open access to scientific publication and research data in the wider context of dissemination and exploitation

“H2020 Programme: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in H2020”. 21 marzo2017 v3.2,
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf



Other research funders' open access policies

- Sherpa-Juliet



- ROARMAP: Registry of Open Access Repository Mandates and Policies



- Johns Hopkins University. Data Management Resources. Funder Data Related Mandates



- Digital Curation Centre (DCC)



<http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies>

<http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies>



HORIZON 2020

Data management planning:

Creating a DMP



What is data management?

“Research data management (or RDM) is a term that describes the organization, storage, preservation, and sharing of data collected and used in a research project. It involves the everyday management of research data during the lifetime of a research project (for example, using consistent file naming conventions). It also involves decisions about how data will be preserved and shared after the project is completed (for example, depositing the data in a repository for long-term archiving and access).”

University of Pittsburgh



Data Documentation Initiative – DDI (2008)



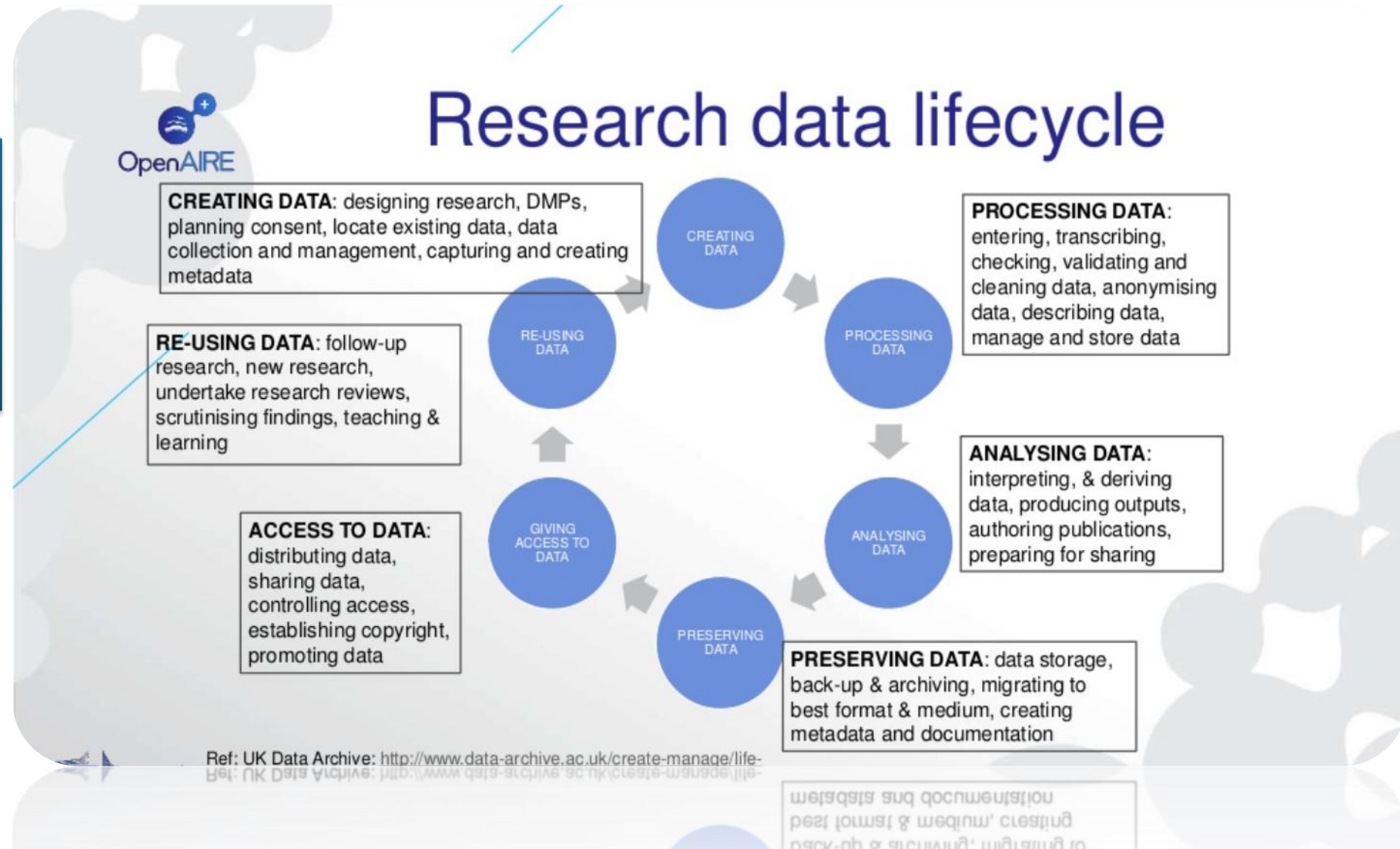
Why manage data?

- **Save time:** Planning ahead for your data management needs will save you time and resources
- **Increase your research impact:** Publications and data openly shared are cited more often
- **Preserve data:** Depositing data in a repository safeguards your investment. Many data sets are unique and can only be collected once
- **Increase your research efficiency:** With proper data management, you'll be able to quickly understand and locate the data you've collected
- **Meet grant/funder requirements**
- **Promote new discoveries:** By sharing data with other researchers



Research data lifecycle

The data life cycle is a part of every scientific research project, and demonstrates the various process involved in data management.



58





What is a Data Management Plan?

Data Management Plans (DMPs) are a key element of good data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. As part of making research data findable, accessible, interoperable and re-usable (FAIR), a DMP should include information on:

- The handling of research data during and after the end of the project
- What data will be collected, processed and/or generated
- Which methodology and standards will be applied
- Whether data will be shared/made open access and
- How data will be curated and preserved (including after the end of the project)



A DMP is required for all projects participating in the extended ORD pilot, unless they opt out of the ORD pilot. However, projects that opt out are still encouraged to submit a DMP on a voluntary basis.



DMPs definitions

“A data management plan is a formal document you develop at the start of your research project which outlines all aspects of your data (i.e., what you will do with your data during and after your research project).”

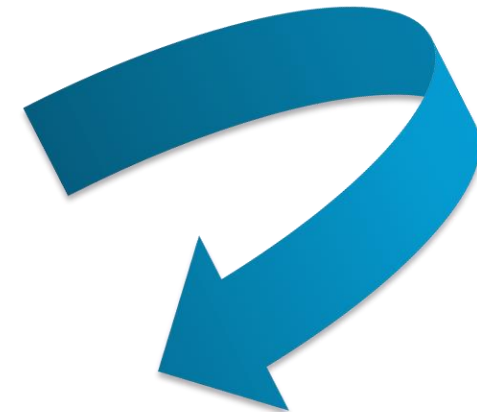
NTU Libraries

“A data management plan is a document that describes how you will collect, organise, manage, store, secure, back up, preserve, and share your data.”

The University of Queensland

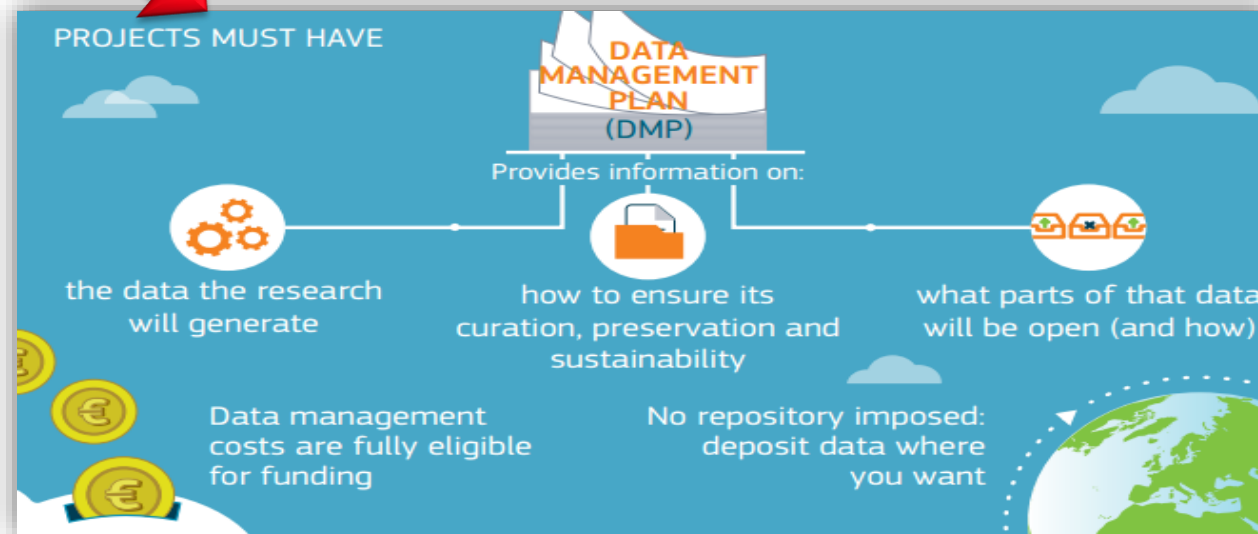
“The DMP should describe whether and how data generated through the course of the proposed research will be shared and preserved (including timeframe), or explain why data sharing and/or preservation are not possible or scientifically appropriate.”

NASA



NASA's Terrestrial Ecology Program now requires that each proposal include a **Data Management Plan (DMP)** of up to **two pages**.

DMPs definitions





Why DMPs

Most funders require researchers to submit a data management plan with their grant applications.

Writing a DMP provides the impetus for researchers to consider and propose to their funders how they will manage and share their data before project work begins.

This benefits researchers as it establishes the framework and resources to support their research data, which translates to better time management and lower costs during project work.

The University of Edinburgh. Research Data Service



Benefits of writing a DMP

If you commit to the data management plan that you submit to your funder, you also stand to gain the following short- and long-term benefits:

- Comply with research funder mandates
- Find and understand your data when needed
- Preventing duplication of effort by enabling others to use your data.
- Document your data to ensure access and continuity when you or other colleagues depart or new research staff start
- If required, allows for the validation of your published results
- Ensure your research is visible and has impact
- Get credit when others cite your work
- Sharing your data can lead to collaboration and advances in research
- Enhancing data security and minimising the risk of data loss

The University of Edinburgh. Research Data Service

Data management plans must be continuously maintained and kept up-to-date throughout the course of the research



DMPs: information to include

Describe the data that your research will generate or collect	Types of data, format for files in which data will be stored, estimated size
Describe how you will annotate and/or describe the data, including the metadata standards and tools (if any) that will be employed	Applicable standards for metadata content and format, including tools/software used to capture and edit the metadata
How will the data be organized, stored and protected during the research project?	Storage methods and backup procedures for the data. Description of the facilities and equipment will be required What measures will be taken for protection of privacy and confidentiality, for sensitive data. Consider security, intellectual property and other right.
How will the data be shared with others, during and/or after the project?	Deposit your data in repositories that facilitate access, with descriptive information to enable finding, DOIs, links to publications. Provide access and restrictions, who may access, under what conditions, a timeline for providing access.
Where and how will the data be archived/preserved for long-term access?	Plans for preserving data in accessible form. Timeline: how long data are to be preserved. Documenting the resources needed to fulfil preservation goals.



DMPs: information to include

What research data you will be creating or collecting.

Who will be responsible for each aspect of the management plan you are developing.

What policies (funding, institutional, and legal) will apply to your data.

How the data will be organised (folder structures, file naming conventions, file versioning).

How the data will be documented during the collection and analysis phase of your research.

What data management practices (backups, storage, access control, archiving) you will use to store & secure your data.

What facilities and equipment will be required (hard-disk space, backup server, repository).


Who will have ownership and access rights to your data.

How the data will be preserved and made available in the long term once your research is completed.



How to write a DMP

Checklist that presents the main questions or themes that researchers may want to cover when writing a DMP

 Checklist for a Data Management Plan, v4.0	
<p>Please cite as: DCC. (2013). <i>Checklist for a Data Management Plan. v.4.0</i>. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/data-management-plans</p>	
DCC Checklist	DCC Guidance and questions to consider
Administrative Data	
ID	A pertinent ID as determined by the funder and/or institution.
Funder	State research funder if relevant
Grant Reference Number	Enter grant reference number if applicable [POST-AWARD DMPs ONLY]
Project Name	If applying for funding, state the name exactly as in the grant proposal.
Project Description	<p>Questions to consider:</p> <ul style="list-style-type: none"> - What is the nature of your research project? - What research questions are you addressing? - For what purpose are the data being collected or created? <p>Guidance:</p> <p>Briefly summarise the type of study (or studies) to help others understand the purposes for which the data are being collected or created.</p>
PI / Researcher	Name of Principal Investigator(s) or main researcher(s) on the project.
PI / Researcher ID	E.g ORCID http://orcid.org/
Project Data Contact	Name (if different to above), telephone and email contact details
Date of First Version	Date the first version of the DMP was completed
Date of Last Update	Date the DMP was last changed
Related Policies	<p>Questions to consider:</p> <ul style="list-style-type: none"> - Are there any existing procedures that you will base your approach on? - Does your department/group have data management guidelines? - Does your institution have a data protection or security policy that you will follow? - Does your institution have a Research Data Management (RDM) policy? - Does your funder have a Research Data Management policy? - Are there any formal standards that you will adopt? <p>Guidance:</p> <p>List any other relevant funder, institutional, departmental or group policies on data management, data sharing and data security. Some of the information you give in the remainder of the DMP will be determined by the content of other policies. If so, point/link to them here.</p>

1. Administrative Data
2. Data collection
3. Documentation & metadata
4. Ethics and legal compliance
5. Storage and backup
6. Selection and preservation
7. Data sharing
8. Responsibilities & resources



H2020 and DMPs

In the Pilot, the first version of the DMP must be delivered within the first 6 months of the project, once the project has its funding approved and has started. So it is not required at the proposal stage

The Commission provides a DMP template, the use of which is recommended but voluntary.

The DMP is not part of the proposal evaluation.

The DMP is intended to be a living document, needs to be updated over the course of the project whenever significant changes arise, such as:

- new data
- changes in consortium policies
- or changes in consortium composition and external factors.

The DMP should be updated as a least during mid-term and the end of the project.



H2020 Template

ANNEX 1 Horizon 2020 FAIR Data Management Plan (DMP) template

The template is a set of questions that it should answered with a level of detail appropriate to the project.

1. Data Summary
2. FAIR data
 - 2.1 Making data findable, including provisions for metadata
 - 2.2 Making data openly accessible
 - 2.3 Making data interoperable
 - 2.4 Increase data re-use (through clarifying licenses)
3. Allocation of resources
4. Data security
5. Ethical aspects
6. Other issues



H2020 Template

SUMMARY TABLE 1

FAIR Data Management at a glance: issues to cover in your Horizon 2020 DMP

This table provides a summary of the Data Management Plan (DMP) issues to be addressed, as outlined in Annex I. You should refer to the annex and the main text of the guidelines for further guidance.

DMP component	Issues to be addressed
1. Data summary	<ul style="list-style-type: none"> State the purpose of the data collection/generation Explain the relation to the objectives of the project Specify the types and formats of data generated/collected Specify if existing data is being re-used (if any) Specify the origin of the data State the expected size of the data (if known) Outline the data utility: to whom will it be useful
2. FAIR Data	
2.1. Making data findable, including provisions for metadata	<ul style="list-style-type: none"> Outline the discoverability of data (metadata provision) Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers? Outline naming conventions used Outline the approach towards search keyword Outline the approach for clear versioning 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

2.2 Making data openly accessible	<ul style="list-style-type: none"> Specify which data will be made openly available? If some data is kept closed provide rationale for doing so Specify how the data will be made available Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)? Specify where the data and associated metadata, documentation and code are deposited Specify how access will be provided in case there are any restrictions
2.3. Making data interoperable	<ul style="list-style-type: none"> Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability. Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?
2.4. Increase data re-use (through clarifying licences)	<ul style="list-style-type: none"> Specify how the data will be licenced to permit the widest reuse possible Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why Describe data quality assurance processes Specify the length of time for which the data will remain re-usable
3. Allocation of resources	<ul style="list-style-type: none"> Estimate the costs for making your data FAIR. Describe how you intend to cover these costs Clearly identify responsibilities for data management in your project Describe costs and potential value of long term preservation

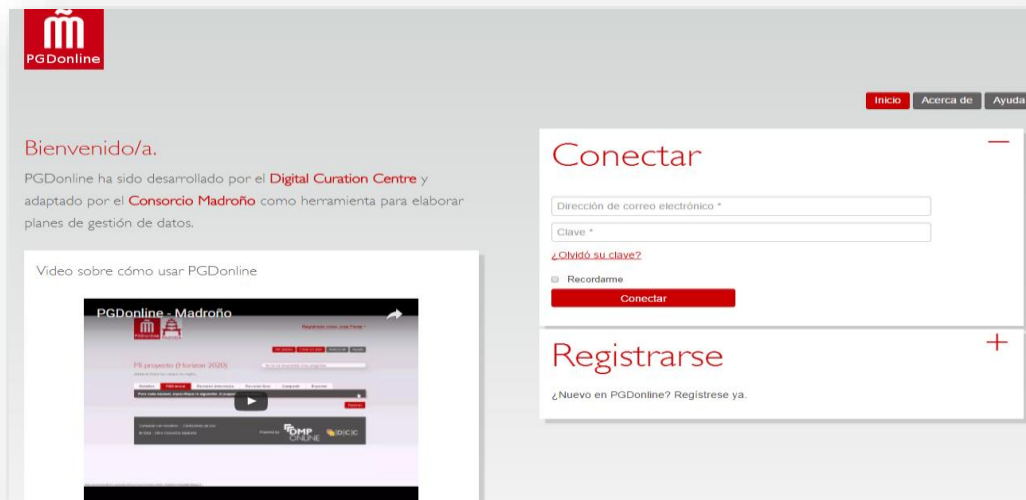
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Using a data management planning tool

There are tools to support researchers in writing their data management plan:

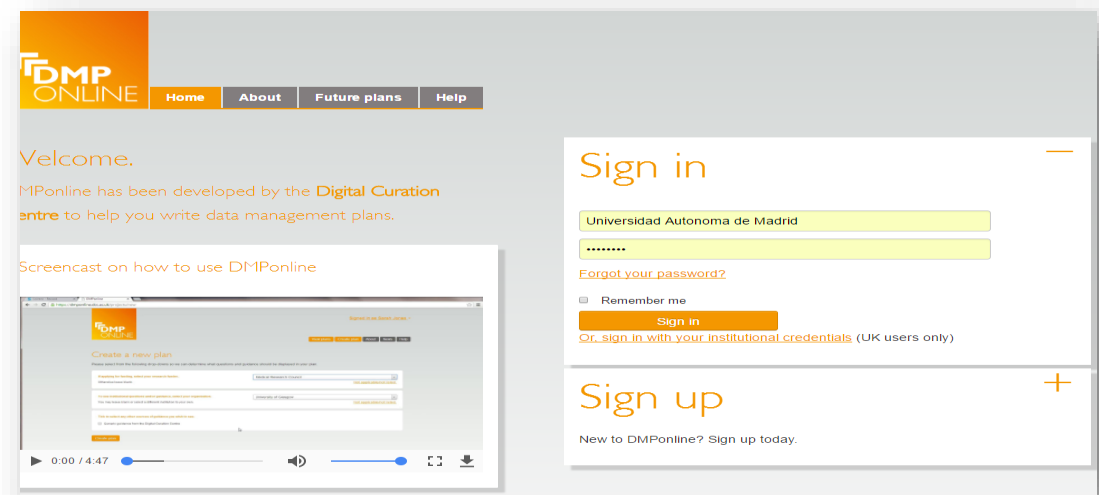
- **DMP-online:** is the DCC data management planning tool. A flexible web-based tool to assist users to create personalised plans according to their context or research funder. Includes a template for Horizon 2020.
- **DMP-Tool:** is the University of California Curation Center online tool.

PGDOnline



<http://pgd.consorciomadrone.es>

DMPOnline



<https://dmponline.dcc.ac.uk>



DMPonline

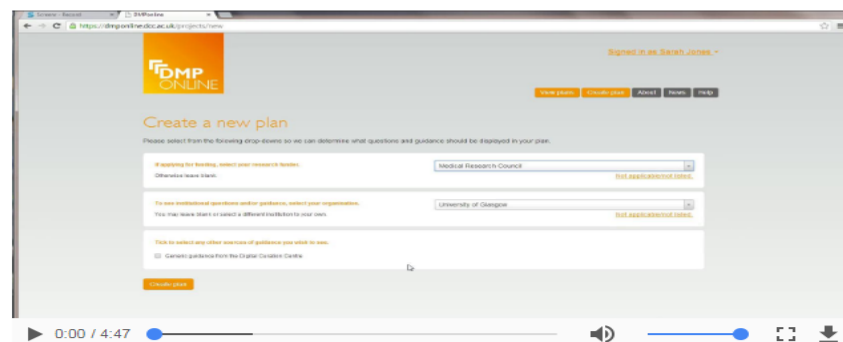


[Home](#) [About](#) [Future plans](#) [Help](#) [Change language](#)

Welcome.

DMPonline helps you to create, review, and share data management plans that meet institutional and funder requirements. It has been jointly developed by the Digital Curation Centre (DCC) and the University of California Curation Center (UC3).

Screencast on how to use DMPonline



[Contact us](#) | [Terms of use](#)

Sign in

Email address *

Password *

[Forgot your password?](#)

☐ Remember me

Sign in

[Or, sign in with your institutional credentials](#) (UK users only)

Create account

New to DMPonline? Create an account today.



DMPonline

← → ↻ Es seguro | https://dmponline.dcc.ac.uk/projects/new

Signed in as Marisa Perez ▾

DMP ONLINE View plans Create plan About Future plans Help

Create a new plan

Please select from the following drop-downs so we can determine what questions and guidance should be displayed in your plan.

If you aren't responding to specific requirements from a funder or an institution, [select here to write a generic DMP](#) based on the most common themes.

If applying for funding, select your research funder.
Otherwise leave blank.

Funder

- Biotechnology and Biological Sciences Research Council (BBSRC)
- Cancer Research UK (CRUK)
- Economic and Social Research Council (ESRC)**
- Engineering and Physical Sciences Research Council (EPSRC)
- European Commission (Horizon 2020)
- Medical Research Council (MRC)
- National Science Foundation (USA)
- Natural Environment Research Council (NERC)

Confirm plan details


You have selected the Default DMP, which is based on the DCC Checklist. This offers a generic set of DMP questions and guidance. For more details see: [DMP checklist 2013](#).

Cancel Create plan

Contact us | Terms of use | DMPonline previous version
© 2004 - 2017 Digital Curation Centre (DCC)



DMPonline



[View plans](#)
[Create plan](#)
[About](#)
[Future plans](#)
[Help](#)

Create a new plan

Please select from the following drop-downs so we can determine what questions and guidance should be displayed in your plan.

If you aren't responding to specific requirements from a funder or an institution, [select here to write a generic DMP](#) based on the most common themes.

If applying for funding, select your research funder.

Otherwise leave blank.

European Commission (Horizon 2020)

Not applicable/not listed.

To see institutional questions and/or guidance, select your organisation.

You may leave blank or select a different organisation to your own.

University of Edinburgh

De Montfort University

University of Derby

University of Dundee

University of Durham

ELIXIR

Edge Hill University

Edinburgh Napier University

University of Edinburgh

Tick to select any other sources of guidance you wish to see.

☒ DCC guidance
 ☐ Roslin Institute

Create plan

Confirm plan details

Where your funder or institution doesn't have specific requirements (or if you left these options blank), you will see the DCC Checklist. This offers a generic set of DMP questions and guidance. For more details see: [DMP checklist 2013](#).

Funder: European Commission (Horizon 2020)

Institution: University of Edinburgh

Template: Horizon 2020 DMP

Other guidance:

DCC guidance

Cancel

Yes, create plan



DMPonline

The image displays three overlapping screenshots of the DMPonline web application, illustrating the steps to create a Horizon 2020 Data Management Plan (DMP).

Screenshot 1 (Left): Shows the 'Plan details' tab. A message at the top states 'Plan was successfully created.' Below, the 'My plan (Horizon 2020 DMP)' form is visible. Fields include Plan name (My plan (Horizon 2020 DMP)), ID, Grant number, Principal Investigator/Researcher (Marisa Perez), Principal Investigator/Researcher ID, Plan data contact, and Description. A 'Create plan' button is present.

Screenshot 2 (Middle): Shows the 'Initial DMP' tab. It displays a progress bar at the top right indicating 'Signed in as Marisa Perez' and '0/9' questions answered. Below the progress bar, a list of sections is shown, each with a question count and answer status:

- 1. Data summary (1 question, 0 answered)
- 2. FAIR data (4 questions, 0 answered)
- 3. Allocation of resources (1 question, 0 answered)
- 4. Data security (1 question, 0 answered)
- 5. Ethical aspects (1 question, 0 answered)
- 6. Other (1 question, 0 answered)

Screenshot 3 (Right): Shows the 'Detailed DMP' tab for the '2. FAIR data' section. It indicates '0/9 questions answered' and 'approx. 15% of available space used'. The content area provides guidance on FAIR data, stating: 'In general terms, your research data should be 'FAIR' that is findable, accessible, interoperable and re-usable. These principles precede implementation choices and do not necessarily suggest any specific technology, standard or implementation-solution.'

2.1 Making data findable, including provisions for metadata:

- Outline the discoverability of data (metadata provision)
- Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?
- Outline naming conventions used
- Outline the approach towards search keyword
- Outline the approach for clear versioning
- Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

A 'Guidance' tab is active, showing 'EC Guidance' which states: 'The Research Data Alliance provides a [Metadata Standards Directory](#) that can be searched for discipline-specific standards and associated tools.'



DMPonline

My plan (Horizon 2020 DMP)

Plan details	Initial DMP	Detailed DMP	Final review DMP	Share	Export
<p>From here you can download your plan in various formats. This may be useful if you need to submit your plan as part of a grant application. Select what format you wish to use and click to 'Export'.</p>					
Initial DMP					+
Detailed DMP					-
<p>Format</p> <div> <div> pdf csv html json pdf text xml docx </div> <div>Export</div> </div>					
Final review DMP					+

[Contact us](#) | [Terms of use](#) | [DMPonline previous version](#)

© 2004 - 2017 Digital Curation Centre (DCC)





H2020 DMPs in Zenodo

- Candela Bravo, Alexandre Almeida, & Christoforos Kachris. (2017). D8.3 Data Management Plan (Intermediate version). Zenodo. <http://doi.org/10.5281/zenodo.936394>

Grant: VINEYARD - Versatile Integrated Accelerator-based Heterogeneous Data Centres (687628)

- Jones, B., Amsaghrou, R., & Shiers, J. (2016). D1.1_Data Management Plan. Zenodo. <http://doi.org/10.5281/zenodo.48171>
- Gerosa, Matteo, Bressan, Serena, & Pistore, Marco. (2017). H2020 692819 SIMPATICO - D1.3: Data management plan v1 (Version Version 1). Zenodo. <http://doi.org/10.5281/zenodo.1040799>
- Monica Caballero, & M^o Eugenia Fuenmayor, . (2015). AutoPost-D6.3: Data management plan. Zenodo. <http://doi.org/10.5281/zenodo.56107>



Other funders DMP

- Donaldson, M. , Schwamm, H. and Campbell, F. (2017) EPSRC Data Management Plan Assessment Rubric v2.0. Documentation. University of Glasgow, Zenodo.

<http://dx.doi.org/10.5281/zenodo.247087>

Template: EPSRC

- Calli Jenkerson, Steven Foga. USGS CDR/ECV DMP.

<https://dmptool.org/plans/5119.pdf>

Template: NSF-AGS: Atmospheric and Geospace Sciences

- Jennifer McWhorter. **Coastal Data Information Program (CDIP).**

<https://dmptool.org/plans/15485.pdf>

Template: NOAA Data Sharing

More examples:



Organizing files and data

Marisa Pérez Aliende
14th November 2017



Universidad
Carlos III de Madrid

File organization

Good file management in research helps to:

- **Identify**
- **Locate**
- **Use data effectively files**

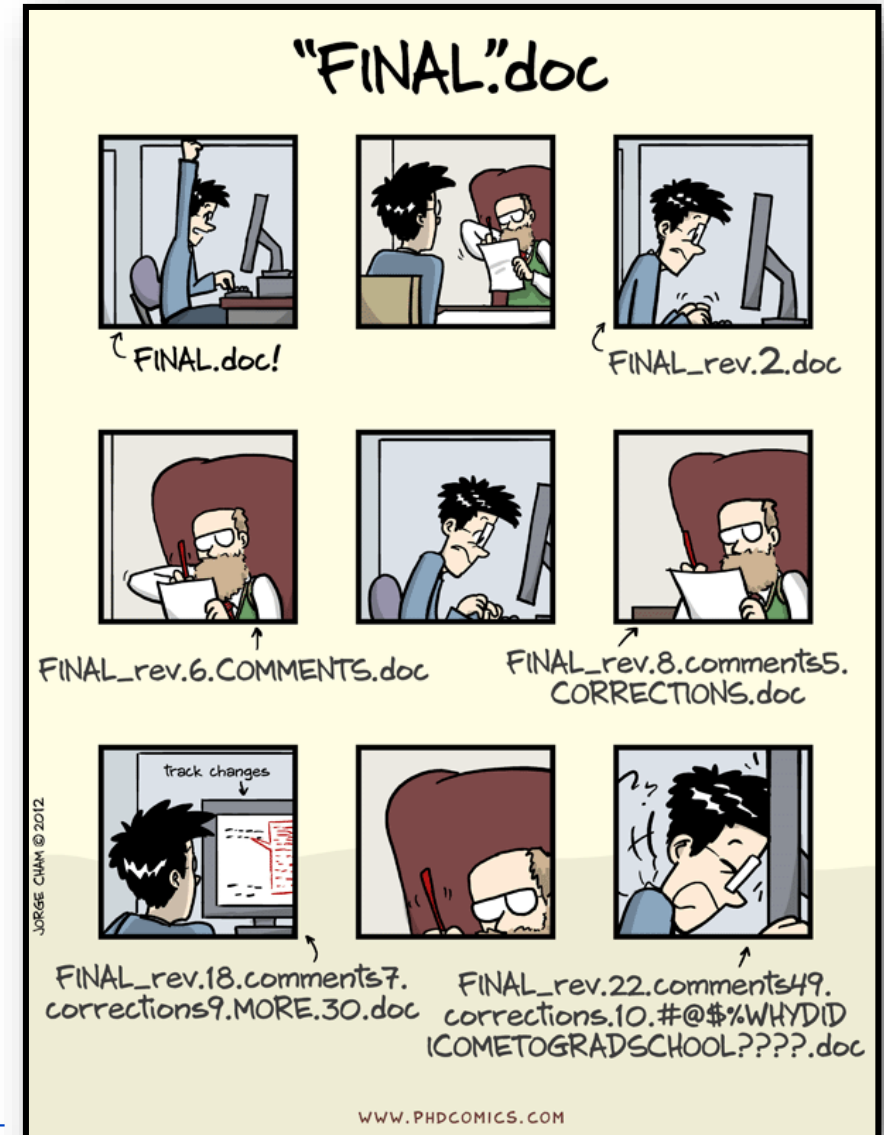
Why file organization is important?

Once a research gets underway, there may be multiple files in various formats, multiple versions, methodologies, etc., all relating to that research



Can someone else understand or use those data files?

<http://phdcomics.com/comics.php?f=1531>

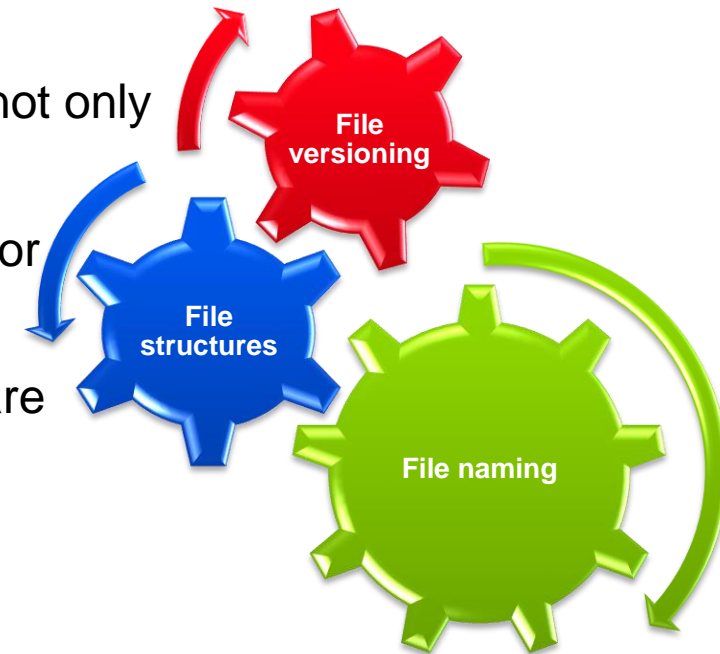




File organization: Best practices

Research data file and folder need to be labelled and organized in a systematic way so that they are both identifiable and accessible for current and future users.

- Data files are distinguishable from each other without their container folder.
- They are easier to locate and browse, and they can be retrieved not only by the creator but by other users as well
- Sorted in a logical sequence and are not accidentally overwritten or deleted
- Good data file naming prevents confusion when multiple people are working and shared files



But ... what do we mean by file organization?



File naming conventions

Help you stay organized by making it easy to identify the file(s) that contain the information that you are looking for just from its title and by grouping files that contain similar information close together:

- Maintain order
- Include same information

Best Practice	Example	
Limit the file name to 32 characters (preferably less!)	32CharactersLooksExactlyLikeThis.csv	
When using sequential numbering, use leading zeros to allow for multi-digit versions For a sequence of 1-10: 01-10 For a sequence of 1-100: 001-010-100	NO YES	ProjID_1.csv ProjID_12.csv ProjID_01.csv ProjID_12.csv
Don't use special characters & , * % # ; * () ! @ \$ ^ ~ ' { } [] ? < > -	NO	name&date@location.doc
Use only one period and use it before the file extension	NO NO YES	name.date.doc name_date..doc name_date.doc
Avoid using generic data file names that may conflict when moved from one location to another	NO YES	MyData.csv ProjID_date.csv



Some examples

sam_monarch_wing_20160115_CM_001.tif
 [instrument]_[item]_[date]_[collector]_[ascension#].ext

FileOrgSlides_20170118.pptx
 [class][material]_[date].ext

SevilletaLTER_NM_2001_NPP.csv
 [project name]_[state]_[year]_[dataset].ext
SevilletaLTER_NM_2001_NPP_20170117.csv
 [project name]_[state]_[year]_[dataset]_[analysisID].ext

Malinowski "Data Management: File organization". 2017

https://libraries.mit.edu/data-management/files/2014/05/FileOrgSlides_20170118sm.pdf

Information to consider including in file names

1. Project or experiment name or acronym
2. Location/spatial coordinates
3. Researcher name/initials
4. Date or date range of experiment
5. Type of data
6. Conditions
7. Version number of file
8. File extension

Include a README.txt file in the directory that explains the naming convention along with any abbreviations or codes used.

Check for established file naming conventions in your discipline.

20130503_DOEProject_DesignDocument_Smith_v2-01.docx
 20130709_DOEProject_MasterData_Jones_v1-00.xlsx
 20130825_DOEProject_Ex1Test1_Data_Gonzalez_v3-03.xlsx
 20130825_DOEProject_Ex1Test1_Documentation_Gonzalez_v3-03.xlsx
 20131002_DOEProject_Ex1Test2_Data_Gonzalez_v1-01.xlsx
 20141023_DOEProject_ProjectMeetingNotes_Kramer_v1-00.docx



File versioning

Save new versions
Establish a consistent convention

- Use ordinal numbers (1,2,3,etc) for major version changes and a decimal for minor changes:
data_1
data_1.1
data_1.2
- Use dates to distinguish between successive versions.
- Avoid imprecise “final” labels: “final”, “review”, “raw”...
- Put older versions in a separate folder
- Use version control software:
Subversion
Box
TortoiseSVN ...



File structure

Hierarchical:

Items organized in folders and subfolders:

Avoid overlapping categories

Not big folders

Don't let the structure get too deep

**Document your
system**

consistently

Example file structure systems/directory hierarchy conventions:

```

/[Project]/[Grant Number]/[Event]/[Date]
/[Project]/[Sub-project]/[Run of an experiment]/[Person]/[Date]
/[Research area]/[Project]/[Data vs. documentation]/[Date]
/[Project]/[Type of file]/[Person]/[YYYYMMDD]
/[Instrument]/[Date]/[Sample]
  
```

For the butterfly project:

```

/butterfly/images/mcneill/20140117
/butterfly/tabular/mcneill/20140117
/butterfly/projectDocs/
/butterfly/literature/subject/
  
```



McNeill, D. y Bailey, H. "Research Data Management: File organization. 2014.
<https://libraries.mit.edu/data-management/files/2014/05/file-organization-july2014.pdf>



File formats

The file formats you use have a direct impact on your ability to open those files at a later date and on the ability of other people to access those data.

non-proprietary (open) file format when possible

Some preferred file formats

- Containers: TAR, GZIP, ZIP
- Databases: XML, CSV
- Geospatial: SHP, DBF, GeoTIFF, NetCDF
- Moving images: MOV, MPEG, AVI, MXF
- Sounds: WAVE, AIFF, MP3, MXF
- Statistics: ASCII, DTA, POR, SAS, SAV
- Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Tabular data: CSV
- Text: XML, PDF/A, HTML, ASCII, UTF-8
- Web archive: WARC

Stanford University,
best practices for file formats

Type	Recommended	Avoid for data sharing
Tabular data	CSV, TSV, SPSS portable	Excel
Text	Plain text, HTML, RTF PDF/A only if layout matters	Word
Media	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	Quicktime H264
Images	TIFF, JPEG2000, PNG	GIF, JPG
Structured data	XML, RDF	RDBMS

Proprietary Format	Alternative/Preferred Format
Excel (.xls, .xlsx)	Comma Separated Values (.csv) ASCII
Word (.doc, .docx)	plain text (.txt), XML, PDF/A, HTML, ODF or if formatting is needed, PDF/A (.pdf)
PowerPoint (.ppt, .pptx)	PDF/A (.pdf), ODP, JPEG 2000, PDF, PNG
Photoshop (.psd)	TIFF (.tif, .tiff),
Quicktime (.mov)	MPEG-4 (.mp4), MOV, AVI, MXF
Sounds	WAVE, AIFF
Containers	TAR, GZIP, ZIP
Databases	XML, CSV



Documentation

Documentation is capturing the work that you performed in ways that others could read it, understand what you did and be able to reproduce your work if needed. Your documentation should include both your day to day activities (what you did) as well as "big picture" information (why you did it).

Ex.: laboratory notebooks, experimental protocols, questionnaires...

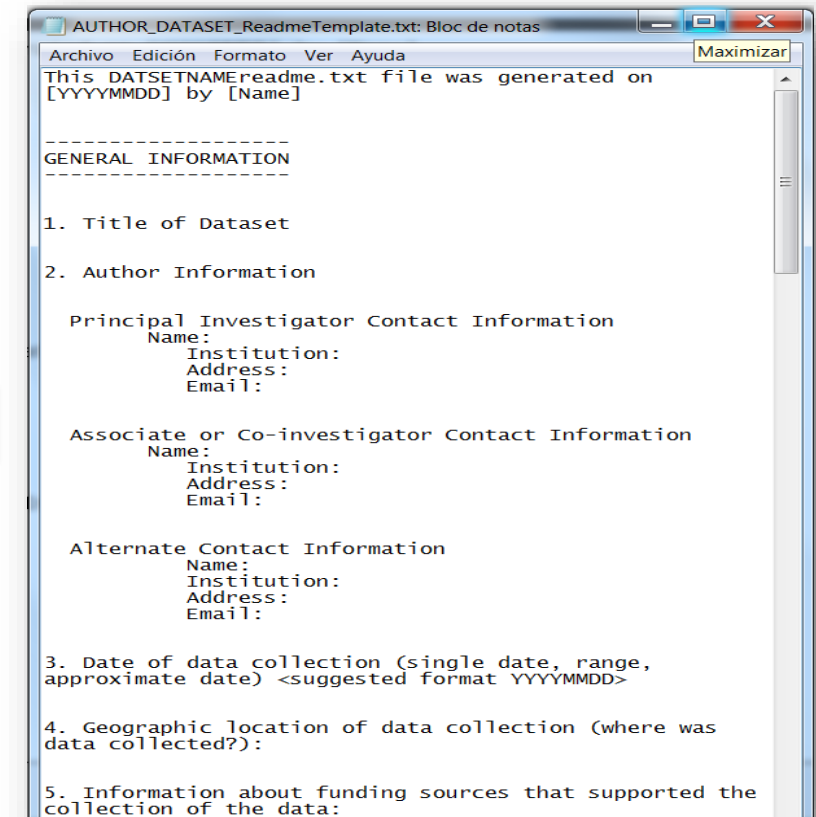
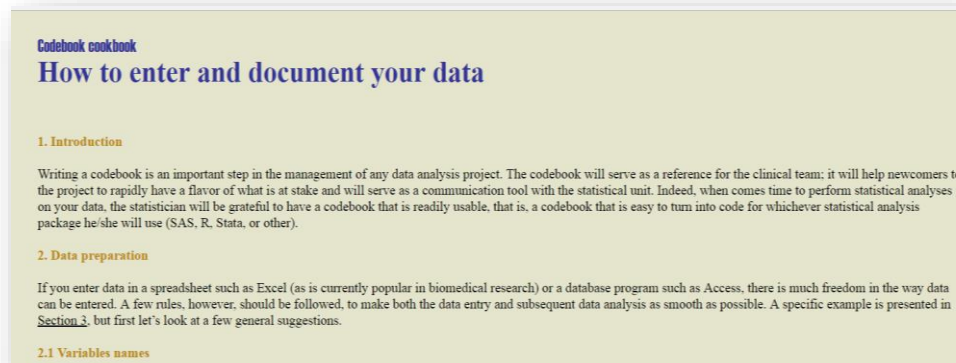
Create one readme file for each data file, whenever possible.

- Readme files (readme.txt)

A readme file provides information about a data file and is intended to help ensure that the data can be correctly interpreted.

- Codebooks and data dictionaries

Template readme file



University of Cornell. Guidelines for writing "readme" style metadata



Data sharing and archiving

Marisa Pérez Aliende
14th November 2017



Universidad
Carlos III de Madrid

Sharing data

Challenges of sharing:

- It takes time and effort to make data shareable
- Perceived risk from loss of control
- Protection of sensitive data
- Unclear or problematic ownership of the data
- A lack of incentives for sharing data



[MANTRA Sharing, Preservation and Licensing](#)

Benefits of sharing:

- Reinforces open scientist research
- Verification and replication of original results
- Promotes new research
- Encourage collaboration
- Reduce costs by avoiding duplicate data collection efforts
- Protects against fraudulent data
- Enhances the visibility and overall impact of research projects
- Preserves data for future
- Helps the broader community and individuals researchers do better research



Ethical issues when sharing: Anonymisation

Before sharing your data, you should consider the privacy, intellectual property, copyright, or licensing issues to be addressed with regard to the sharing.

Anonymisation is a valuable tool that allows data to be shared, whilst **preserving privacy**. The process of anonymising data requires that identifiers are changed in some way such as being removed, substituted, distorted, generalised or aggregated.

Data protection

Anonymisation: managing data protection risk code of practice

ico.

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income	Expenses/month
22	F	S017	£20,000	£1,100
25	M	S018	£22,000	£1,300
30	M	S016	£32,000	£1,800
35	F	S017	£31,500	£2,000
40	F	S015	£68,000	£3,500
50	M	S014	£28,000	£1,200

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income (low if <25,000; medium if between 25,000 to 45,000; high if >45,000)	Expenses/month (low if <1,800; medium if between 1,800 to 2,400; high if >2,400)
20-24	F	S017-19	low	low
25-29	M	S017-19	low	low
30-34	M	S014-16	medium	medium
35-39	F	S017-19	medium	medium
40-44	F	S014-16	high	high
50-54	M	S014-16	medium	low

Interview and page number	Original	Changed to
Int1		
p1	Age 27	Age range 20-30
p1	Spain	European country
p3	Manchester	Northern metropolitan city or English provincial city
p2	20th June	June
p2	Amy (real name)	Moir (pseudonym)

Restrict the upper or lower ranges of a continuous variable to hide outliers if the values for certain individuals are unusual or atypical within the wider group researched. In such circumstances the unusually large or small values might be collapsed into a single code, even if the other responses are kept as actual quantities, or one might code all responses.

example: Annual salary could be 'top-coded' to avoid identifying highly paid individuals. A top code of £100,000 or more could be applied, even if lower incomes are not coded into groups.



Licensing data

A license will define what others may or may not do with your data.

The most widely-recognized:

- Creative Commons
- Open Data Commons
- Public Domain



very simple, factual datasets

6 permutations

Data to be used automatically

Version 4 or later

designed specifically for databases

OpenDataCommons.org

Legal tools for Open Data

• Attribution License (ODC-By)

allows licensees to copy, distribute and use the database, to produce works from it and to modify, transform and build upon it for any purpose.

This {DATA(BASE)-NAME} is made available under the Open Data Commons Attribution License: <http://opendatacommons.org/licenses/by/{version}>.

• Open Database License (ODC-OdbL)

is the same as ODC-By but it adds a copyleft condition that applies to new databases derived from the database, this condition would be satisfied by future versions of the same licence or a compatible one as judged by the licensor. And DRM can only be applied to the database or a new database derived from it if an alternative copy without the restrictions is made equally available.

This {DATA(BASE)-NAME} is made available under the Open Database License: <http://opendatacommons.org/licenses/odbl/1.0/>. Any rights in individual contents of the database are licensed under the Database Contents License: <http://opendatacommons.org/licenses/dbcl/1.0>



Public domain licenses



The most permissive

CC0 1.0 Universal (CC0 1.0) Public Domain Dedication



Is for dedicating works to the public domain, as a waiver of a person's right to work

Public Domain Mark (PDM 1.0) Public Domain Dedication



Is to allow public domain works to be more easily discovered and recognised as such, but it should not be used for waiving rights. Is intended for use with works free of known copyright restrictions.

Open Data Commons Public Domain Dedication and License (PDDL)

accomplishes much the same thing in much the same way as CC0, but is worded specifically in database terms.

Other Open Source Initiative ([OSI](#)) approved licenses:

- [GNU. General Public License](#)
- [MIT license](#)
- [Mozilla Public License 2.0](#)

Licensing assistance

Open Definition Licenses

License	Domain	By	SA	Comments
Creative Commons CCZero (CC0)	Content, Data	N	N	Dedicate to the Public Domain (all rights waived)
Open Data Commons Public Domain Dedication and Licence (PDDL)	Data	N	N	Dedicate to the Public Domain (all rights waived)
Creative Commons Attribution 4.0 (CC-BY-4.0)	Content, Data	Y	N	
Open Data Commons Attribution License (ODC-BY)	Data	Y	N	Attribution for data(bases)
Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA-4.0)	Content, Data	Y	Y	
Open Data Commons Open Database License (ODbL)	Data	Y	Y	Attribution-ShareAlike for data(bases)

Licensing assistance:

- European Data Portal. [Licensing Assistant](https://www.europeandataportal.eu/en/content/show-license) (<https://www.europeandataportal.eu/en/content/show-license>)
- EUDAT licensing wizard helpt to pick licence for data and software <https://www.eudat.eu/services/userdoc/license-selector>
- OER IPR [Support Project](http://www.web2rights.com/OERIPRSupport/) (<http://www.web2rights.com/OERIPRSupport/>)

H2020 guidelines point to CC BY 4.0 o CC0.

MÁS PERMISIVA

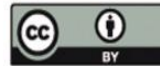


Domínio Público

(CC0):
Europeana, Figshare,
Open Goldberg V.

- Compartir
- Copiar
- Remezclar
- Ganar dinero

- Menciono al autor
(en algunas jurisdicciones)



Reconocimiento

(by):
PLOS, Saylor.org.

- Compartir
- Remezclar
- Ganar dinero

- Menciono al autor



Reconocimiento – CompartirIgual

(by-sa):
Wikipedia, Wikimedia,
Arduino, P2PU

- Compartir
- Remezclar
- Ganar dinero

- Menciono al autor
- Mantengo la misma licencia (by-sa)



Reconocimiento – SinObrDerivada

(by-nd):
Drupal, Behance,
GNU, Free Software Foundation.

- Compartir
- Ganar dinero

- Menciono al autor
- No hago remezclas



Reconocimiento – NoComercial

(by-nc):
Brooklyn Museum,
Wired.com Photography

- Compartir
- Remezclar

- Menciono al autor
- No gano dinero



Reconocimiento – NoComercial – CompartirIgual

(by-nc-sa):
MIT Open CourseWare

- Compartir
- Remezclar

- Menciono al autor
- No gano dinero
- Mantengo la misma licencia (by-nc-sa)



Reconocimiento – NoComercial – SinObrDerivada

(by-nc-nd):
Videos TED Talks,
Propublica

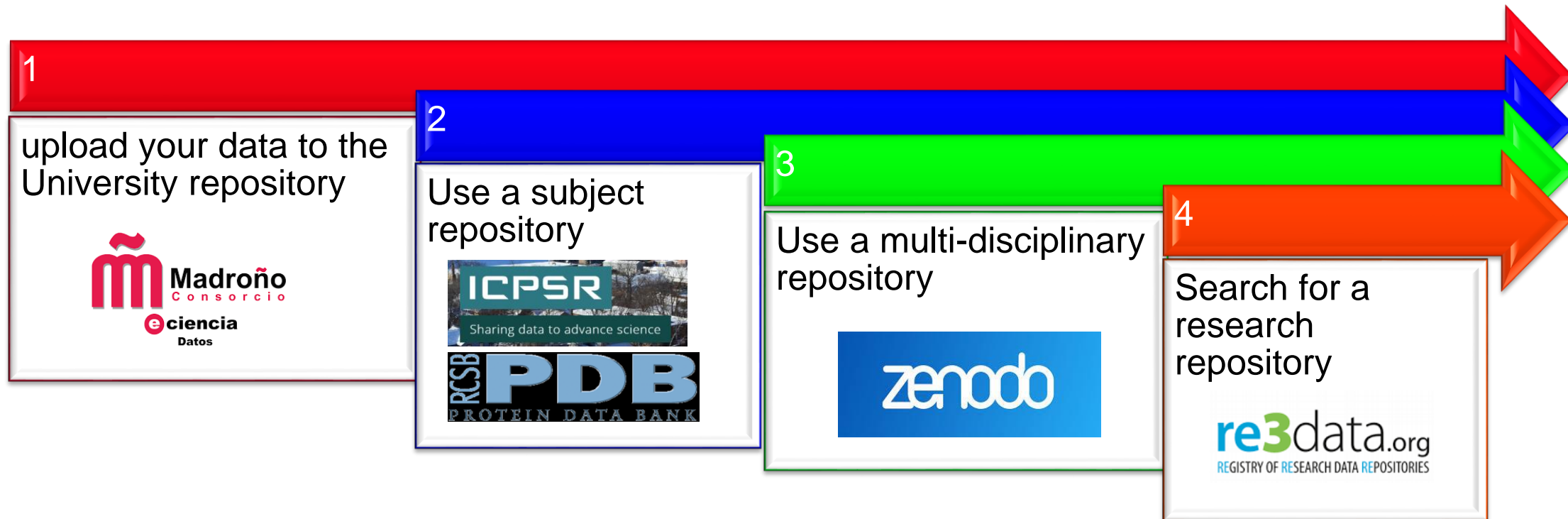
- Compartir

- Menciono al autor
- No hago remezclas
- No gano dinero
- Mantengo la misma licencia (by-nc-nd)

Archiving

H2020 ORD pilot participants are asked to deposit data in a research data repository:

A digital archive collecting and displaying datasets and their metadata





Citing research data

Data citation:

“The practice of providing a reference to data in the same way as researches routinely provide a bibliographic reference to printed resources”



Broadly-applicable data citation standards have not yet been established; use standards adopted by relevant academic journal, data repository, or professional organization



<https://www.force11.org/datacitation>

The Data Citation Principles cover purpose, function and attributes of citations.

These principles recognize the dual necessity of creating citation practices that are both human understandable and machine-actionable.



Citing research data

The challenges associated with data publication vary across disciplines.
Data Cite recommends the following format:

Creator (Publication Year). Title. Publisher. Identifier



It may also be desirable to include information about two optional properties:

Creator (Publication Year). Title. Version. Publisher. Resource Type. Identifier

Where:

Publisher: The organization that provides access to the dataset, ex.: Dryad, Zenodo

Resource Type: Examples, data base, dataset

Data Cite in collaboration with Crossref, mEDRA e ISTICT:

The DOI Citation Formatter
(<http://citation.crosscite.org/>)

From DOI build a complete cite in multiple citation styles

DOI Citation Formatter

Paste your DOI:

10.1016/j.physletb.2015.12.039

For example 10.1145/2783446.2783605

Select Formatting Style:

chicago-author-date

Begin typing (e.g. Chicago or IEEE.) or use the drop down menu.

Select Language and Country:

en-US

Begin typing (e.g. en-GB for English, Great Britain) or use the drop down menu.

Format

Khachatryan, V., A.M. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, et al. 2016. "Search for a Higgs Boson Decaying into $\gamma \gamma \rightarrow \ell\ell$ with Low Dilepton Mass in Pp Collisions at $\sqrt{s} = 8$ TeV." Physics Letters B. Elsevier BV. doi:10.1016/j.physletb.2015.12.039.

Copy to clipboard

Do you want to integrate this service? Check the [Documentation](#)

DOI Registration Agencies



Crossref

mEDRA





When depositing your data...

Main criteria:

Certification as Trustworthy Digital Repository with an explicit ambition to keep the data available in long term

- Are the repository's terms and conditions acceptable? Matches your and your funder needs
- Where is your data going to be stored?
- Are you allowed to provide information about your ORCID ID?
- Will your dataset be given a permanent DOI?
- Does provide guidance on how to cite deposited data?
- Is the repository used by the people in your discipline?
- Does the repository allow you to describe your data sufficiently, so it is easy to find?



DATA OR DIDN'T HAPPEN

Britney, Kristin "Rethinking Research Data"
<https://youtu.be/dXKbkpilQME>

Thank you

Marisa Pérez Aliende
Universidad Autónoma de Madrid
mp.aliende@uam.es