IBEC

Institute for Bioengineering of Catalonia

Open Science Workshops

OSW2 · Data management

April 2022
Fidel Bellmunt

Data management workshop

Program

1. Introduction, the Open Science framework
2. Definitions and the Research Data Lifecycle
3. Research Data Management policies, funders current requirements
4. The Data Management Plan
5. The FAIR principles
6. IBEC's Research Data Management Policy, procedures and tools
7. Choosing a data repository
8. Licenses and copyright

1. Introduction, the Open Science framework

Definitions:

Open Science is science done right. (popular)

Open science refers to a new approach to the scientific process based on cooperative work and new ways of disseminating knowledge, improving accessibility to and re-usability of research outputs by using digital technologies and new collaborative tools. (EC, 2018)

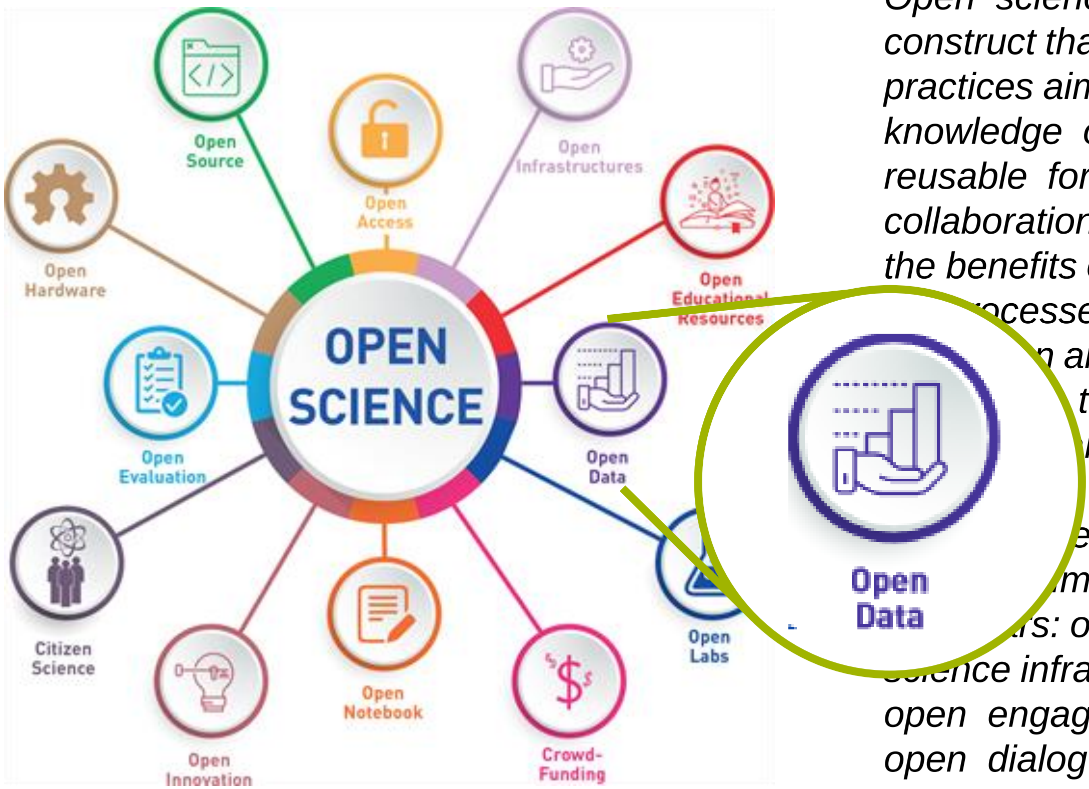
Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks. (Vicente-Saez y Martínez, 2018)

Interesting resources about OS:

➔ Podcast **Open Science Stories** by Heidi Seibold: https://anchor.fm/opensciencestories

➔ **Passport for Open Science** – A Practical Guide for PhD Students: https://www.ouvrirlascience.fr/passport-for-open-science-a-practical-guide-for-phd-students/

Data management workshop - 1. Introduction



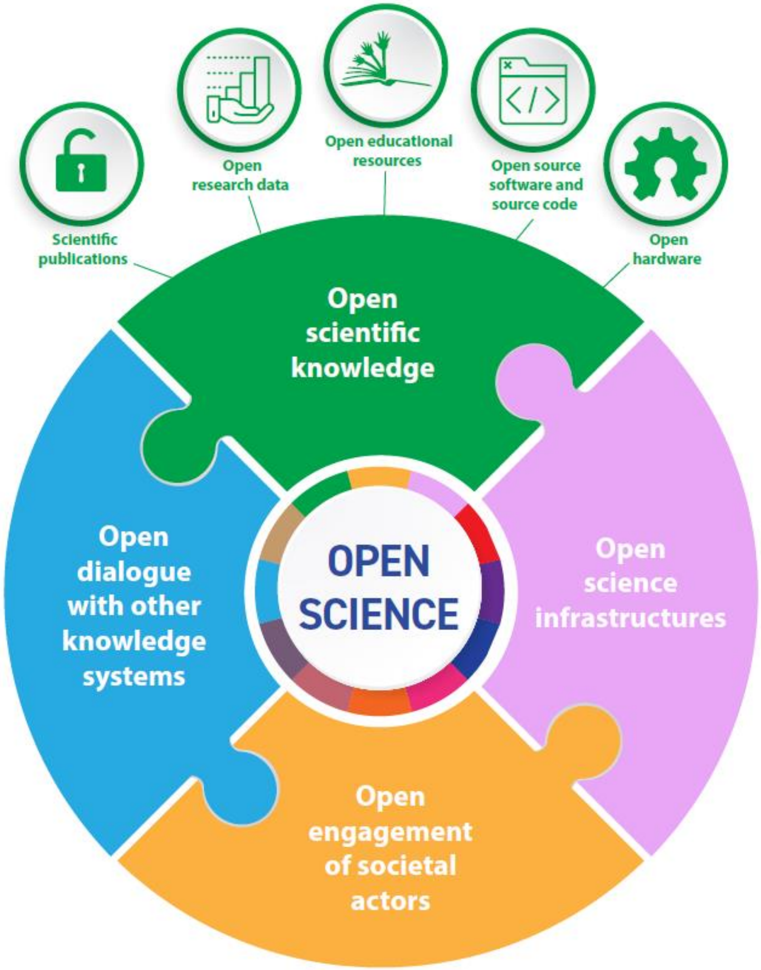
Open science is defined as an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open ocesses of scientific knowledge creation, n and communication to societal actors traditional scientific community. It ll scientific disciplines and aspects practices, including basic and ences, natural and social sciences manities, and it builds on the following rs: open scientific knowledge, open cience infrastructures, science communication, open engagement of societal actors and open dialogue with other knowledge systems.

(UNESCO Recommendation on Open Science, 2021): https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

IBEC
Institute for Bioengineering of Catalonia

Data management workshop

2. Definitions and the Research Data Life Cycle


Scientific publications
Open research data
Open educational resources
Open source software and source code
Open hardware
Open scientific knowledge
Open dialogue with other knowledge systems
OPEN SCIENCE
Open science infrastructures
Open engagement of societal actors

b. Open research data that include, among others, digital and analogue data, both raw and processed, and the accompanying metadata, as well as numerical scores, textual records, images and sounds, protocols, analysis code and workflows that can be openly used, reused, retained and redistributed by anyone, subject to acknowledgement. Open research data are available in a timely and user-friendly, human- and machine-readable and actionable format, in accordance with principles of good data governance and stewardship, notably the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, supported by regular curation and maintenance.

(UNESCO Recommendation, 2021, p. 9)

Data management workshop - 2. Definitions

Research data is any information that has been collected, observed, generated or created to validate original research findings. Research data may be arranged or formatted in a such a way as to make it suitable for communication, interpretation and processing. Data comes in many formats, both digital and physical.

The **research data lifecycle** describes the different stages research data go through before, during, and after a research project. Various data management activities take place within each stage of the data lifecycle, and the choices made in one stage influence the next one.

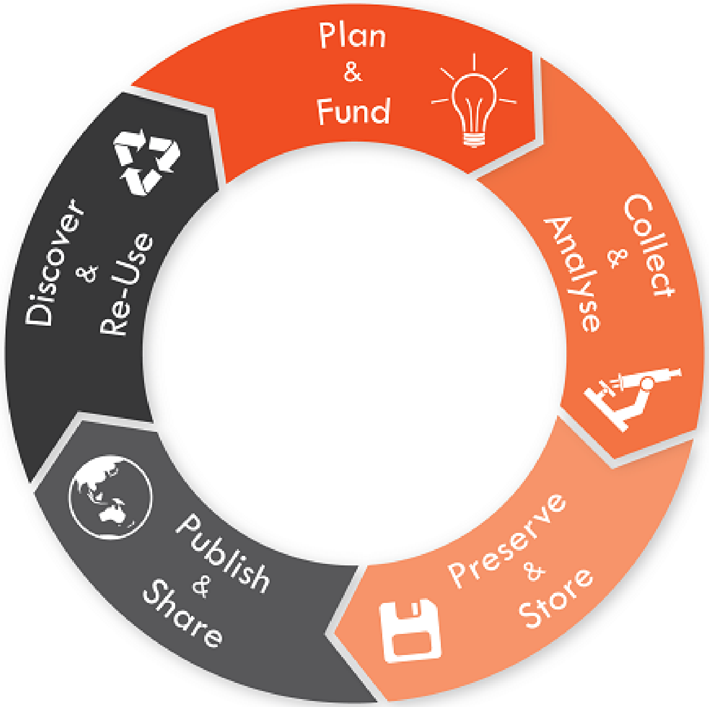
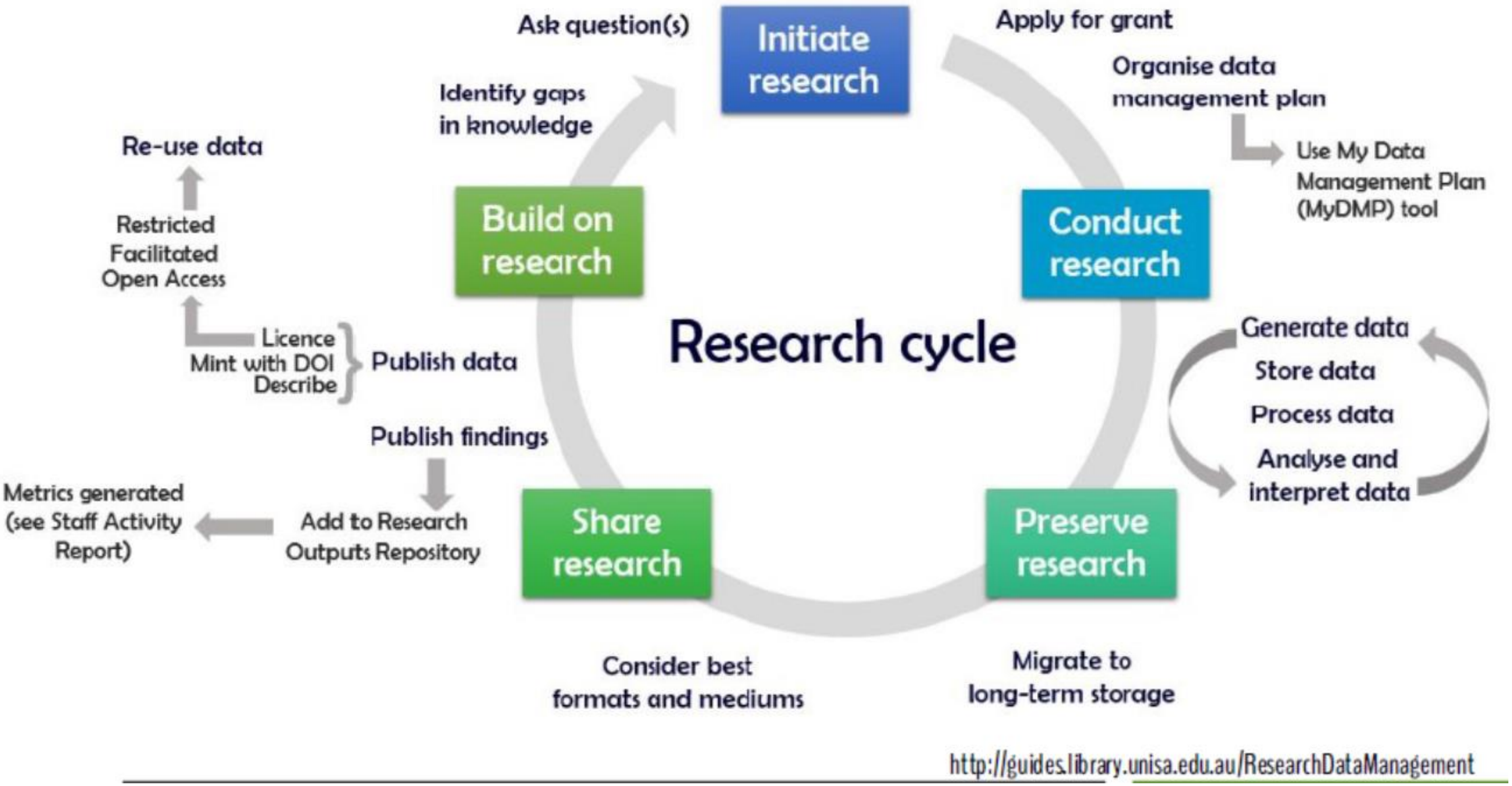


Image from OSF: https://osf.io/fr8pd/wiki/home/

Data management workshop - 2. Research Data Lifecycle

Actions to carry on during the research data lifecycle:


Research cycle
Initiate research
Conduct research
Preserve research
Share research
Build on research
Ask question(s)
Identify gaps in knowledge
Apply for grant
Organise data management plan
Use My Data Management Plan (MyDMP) tool
Generate data
Store data
Process data
Analyse and interpret data
Migrate to long-term storage
Consider best formats and mediums
Publish findings
Add to Research Outputs Repository
Metrics generated (see Staff Activity Report)
Publish data
Licence
Mint with DOI
Describe
Restricted
Facilitated
Open Access
Re-use data
http://guides.library.unisa.edu.au/ResearchDataManagement

Infographic by: Ignasi Labastida, Universitat de Barcelona

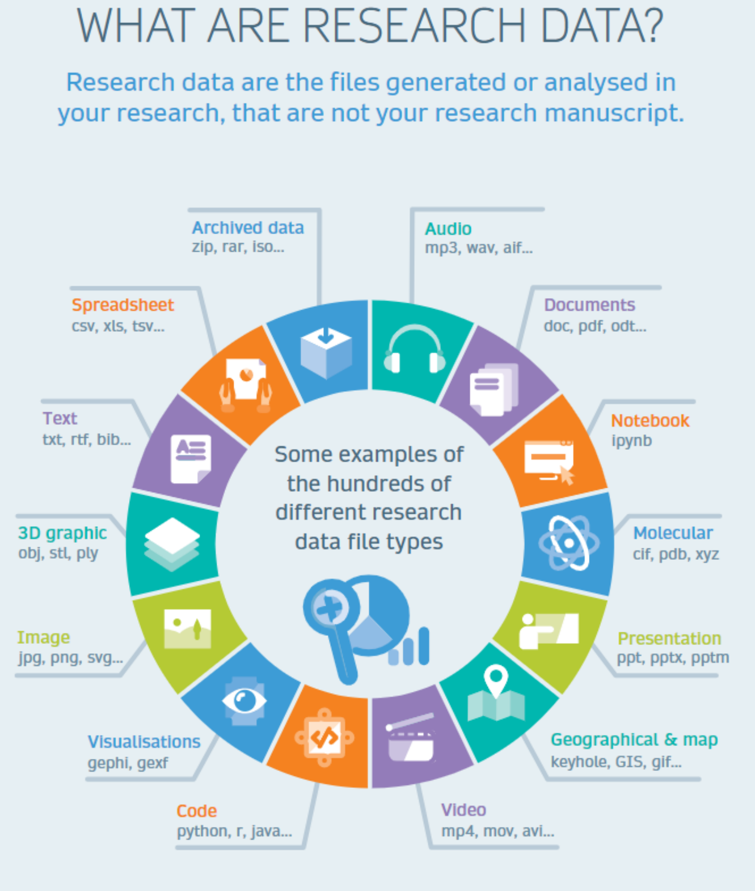
Data management workshop - 2. Research Data Management

Research data management (RDM) refers to the organization, storage and preservation of data created during a research project. Through its lifecycle, it covers initial planning, day-to-day processes and long-term archiving and sharing. It’s required to *make excellent science* and to comply with *funders requirements*:

- ✓ Ensure research **integrity and replication**
- ✓ Ensure research data and records are accurate, complete, authentic and **reliable**
- ✓ Increase your research **efficiency**
- ✓ **Save** time and resources in the long run
- ✓ Enhance data **security** and minimize the risk of **data loss**
- ✓ Prevent **duplication** by enabling others to use your data
- ✓ Comply with practices conducted in industry and commerce
- ✓ Protect your institution from reputational, financial & legal risk
- ✓ Fulfill **publisher** requirements
- ✓ Fulfill **funding** body grant requirements

Data management workshop - 2. Research Data Management

Research data types



Infographic by Science Nature: https://researchdata.springernature.com/documents/web_a92645_what-are-research-data-revision

- documents (text, Word), spreadsheets;
- questionnaires, transcripts, codebooks;
- audiotapes, videotapes;
- photographs, films;
- test responses;
- slides, artefacts, specimens, samples;
- digital objects acquired and generated during research;
- data files;
- database contents;
- models, algorithms, scripts;
- contents of an application (input, output, logfiles for analysis software, simulation software, schemas);
- methodologies and workflows.

What should be done to manage research data correctly?

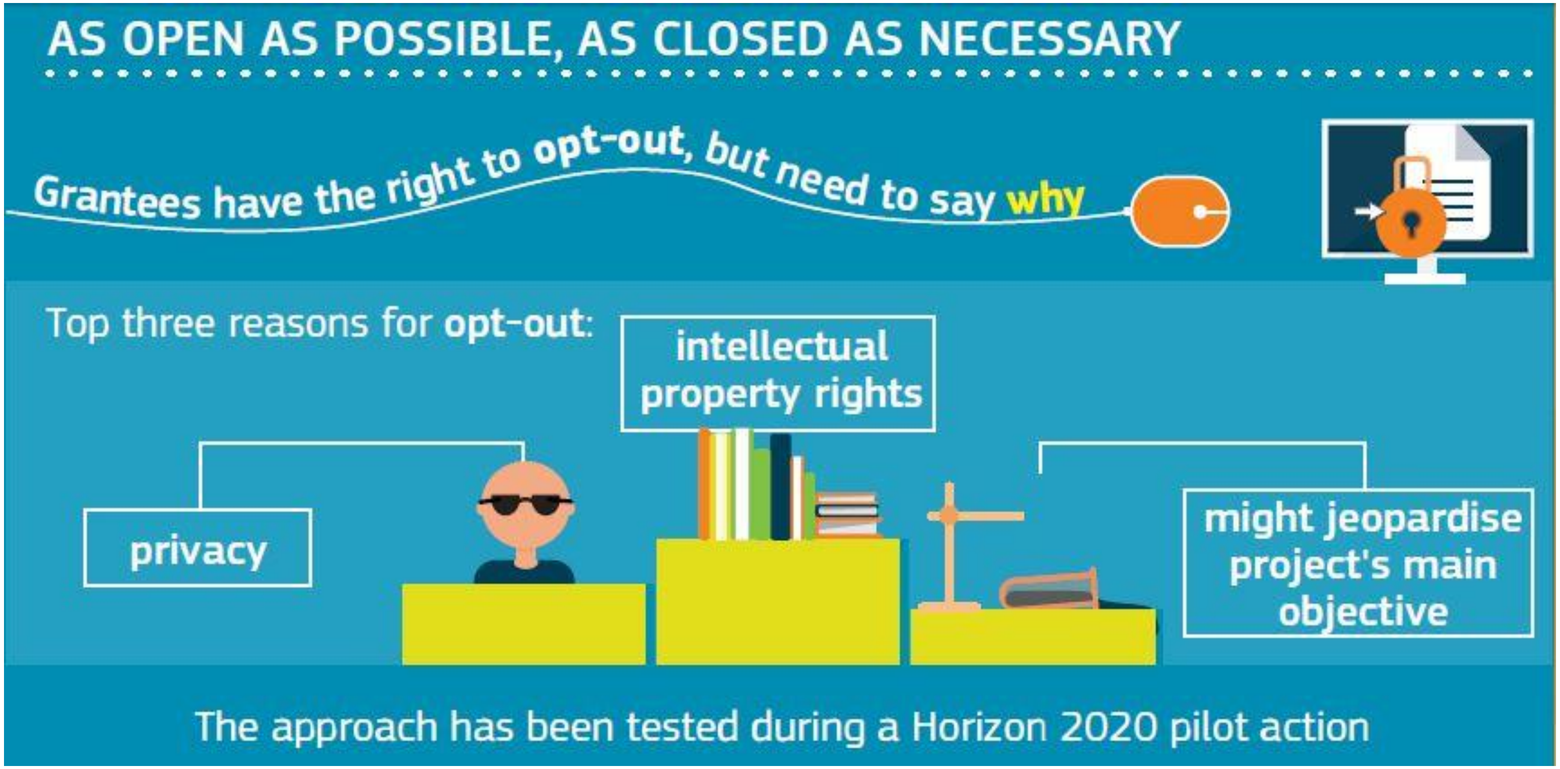
- Planning and describing data-related work *before* it takes place
- Documenting your data (and processing/workflows) so that others can find and understand it
- Choosing open (or at least standardised) file formats where possible
- Storing data safely during a project
- Depositing data in a trusted archive at the end of the research
- Creating metadata records for datasets and licensing them appropriately
- Linking publications to the datasets (and increasingly the code and protocols)

Need of:

- ➔ IT infrastructures,
- ➔ Methodologies,
- ➔ Tools,
- ➔ Standards,
- ➔ Protocols.

3. Research Data Management policies, funders current requirements

Many funders, especially the EC, ask for open access to research data under the principle “**as open as possible, as closed as necessary**“, and in either case, a Data Management Plan.


AS OPEN AS POSSIBLE, AS CLOSED AS NECESSARY
Grantees have the right to opt-out, but need to say why
Top three reasons for opt-out:
intellectual property rights
privacy
might jeopardise project's main objective
The approach has been tested during a Horizon 2020 pilot action

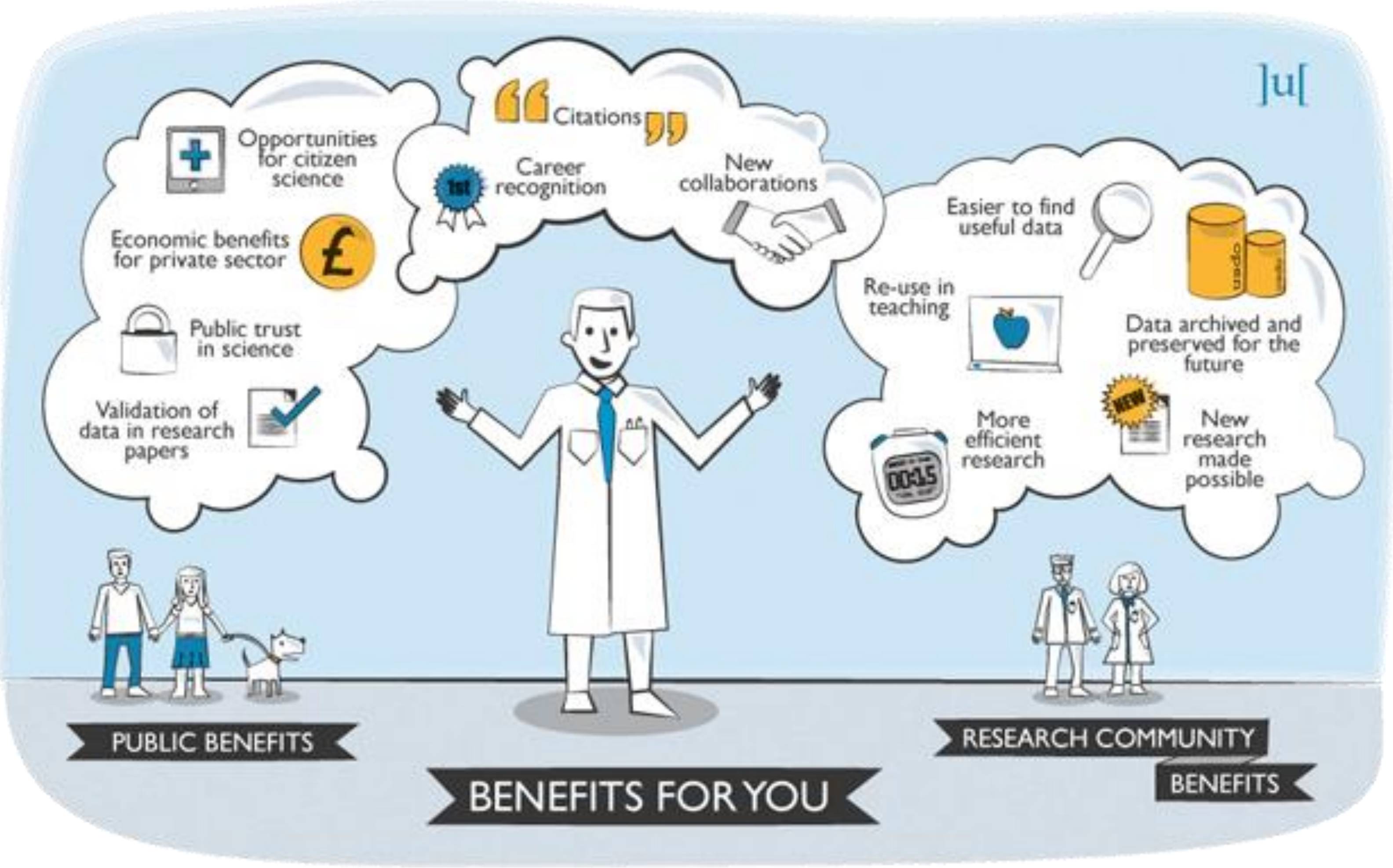
Data management benefits

There are numerous reasons why research data management has become a policy requirement of major funders and research institutions: fundamentally and most importantly it improves academic quality, transparency and the robustness of the scholarly record. But also:

- DURABILITY: Simply put, fewer important datasets will be lost
- SPEED: The research process becomes faster
- IMPACT and LONGEVITY: Data linked to publications receive more citations, over longer periods
- EFFICIENCY: Data collection can be funded once, and used many times for a variety of purposes
- ACCESSIBILITY: Interested third parties can (where appropriate) access and build upon publicly-funded research outputs with minimal barriers to access

Data management workshop - 3. Research Data Management policies, funders current requirements

Benefits on opening data


]u[
Opportunities for citizen science
Economic benefits for private sector
£
Public trust in science
Validation of data in research papers
Citations
Career recognition
1st
New collaborations
Easier to find useful data
Re-use in teaching
open
Data archived and preserved for the future
More efficient research
00:15
NEW
New research made possible
PUBLIC BENEFITS
BENEFITS FOR YOU
RESEARCH COMMUNITY
BENEFITS

(c) poster: Hole, Brian (2012) Poster: The Journal of Open Archaeology Data. Figshare. https://doi.org/10.6084/m9.figshare.96890.v1

Data management workshop - 3. Research Data Management policies, funders current requirements

Funders - European Comission

	HORIZON 2020	HORIZON EUROPE
Measures to ensure reproducibility of research outputs - under the latest available version of the Creative Commons Attribution International Public Licence (CC BY) or a licence with equivalent rights - information about the research outputs, tools and instruments needed to validate the conclusions of scientific publications or to validate/re-use research data		Mandatory
Open access to research data under the principle "as open as possible, as closed as necessary"	Partially: Only for projects which are part of open Research Data Pilot)	Mandatory (but exploitation, protection of IPR, security and privacy rules have a higher priority)
Research output management (Data Management Plan, DMP)	Partially: Only for projects which are part of open Research Data Pilot	Mandatory
Responsible management of research data and metadata of all research outputs (publications, data, software, algorithms, protocols, models, workflows …) in line with the FAIR principles	Partially (not in Grant Agreement but in related documentation for DMP under the Open Research Data Pilot)	Mandatory
Digital or physical access to the results needed to validate the conclusions of scientific publications		Additional mandatory practice imposed in the conditions of the call
In cases of public emergency, immediate open access to all research outputs, if requested by the granting authority		Additional mandatory practice imposed in the conditions of the call

Funders

”la Caixa” Foundation

As regards the research data:

Beneficiaries of a ”la Caixa” Foundation grant must draw up a **data management plan** […]

Thus:

- The first version of the plan should be submitted within a period not exceeding six months from initiation of the project. An updated version of the plan is required when the project undergoes intermediate review, and the definitive version, together with the final report, must be presented on conclusion of the project.
- Among other aspects, the plan should refer to the data that will be generated or used, how and when they will be shared, where they will be available, how they will be preserved and, in the case of sensitive data, the nature of any processing that will be undertaken to meet applicable current legislation.
- Beneficiaries should make public, at least, the data which supports the published results together, if necessary, with any material required (software, setups, etc.) for their understanding or analysis. Such publication of data may be done in any reliable repository or archive.
- In the event of there being other project results that could be disseminated (software, setups, etc.), the beneficiaries must also make them public in an appropriate repository within a period not exceeding six months from conclusion of the project.

https://fundacionlacaixa.org/en/caixaresearch-management-policy-open-access-research-results

Funders - Spanish context

Proyectos Generación de Conocimiento: Research data should be deposited in institutional, national, or international repositories within a period of no longer of 2 years counting from the end date of the project.

Catalan Government: Pacte Nacional per la Societat del Coneixement (2020) that sets the *Catalan open science strategy*:

- Open access to scientific publications.
- The publication of FAIR (findable, accessible, interoperable and reusable) scientific data.
- The creation of new infrastructures to integrate the resources of the Catalan research system into the European ecosystem of the European Open Science Cloud (EOSC).
- Responsible research and innovation policies. Increase the value of scientific culture as an essential tool to form a responsible and critical society and strengthen ad hoc training in this respect.

4. The Data Management Plan

A data management plan or DMP is a formal document that outlines how data are to be handled both during a research project, and after the project is completed. The goal of a data management plan is to consider the many aspects of data management, metadata generation, data preservation, and analysis before the project begins; this may lead to data being well-managed in the present and prepared for preservation in the future.

Wikipedia: https://en.wikipedia.org/wiki/Data_management_plan

What is data management planning?

The act of planning how you will manage all aspects of your data before you begin collecting or creating it. It may include:

- Where and how you will store and back-up your data;
- How and when it will be shared with collaborators;
- The steps required to anonymise sensitive data;
- When, how, and where your data will be preserved and shared.

Writing DMPs can help to:

- Make informed decisions to anticipate & avoid problems.
- Avoid duplication, data loss and security breaches.
- Develop procedures early on for consistency.
- Ensure data are accurate, complete, reliable and secure.
- Save time and effort to make your lives easier.
- Plan to share data early on and increase impact.

Elements of data management

Dataset

A set of files containing both research data - usually numeric or encoded - and documentation sufficient to make the data re-usable.

Documentation

Any digital files such as a codebook, technical or methodology report or user guide, which explain the research data's production, provenance, processing or interpretation.

Metadata

Information about a data item in the repository, including descriptive metadata such as title and other fields used in a citation, and administrative metadata such as date of submission. Usually conforms to a standard to allow computer-to-computer interoperability.

Digital repository

Differs from other digital collections in that: content is deposited in a repository, whether by the content creator, owner or third party; the repository architecture manages content as well as metadata; the repository offers a minimum set of basic services e.g. put, get, search, access control; the repository must be sustainable and trusted, well-supported and well-managed.

DMP Contents at Horizon 2020 template

1. Data summary (description, origin, total size, formats, usefulness…)
2. FAIR data:
   - Making data findable, including provisions for metadata
   - Making data openly accessible
   - Making data interoperable
   - Increase data re-use (through clarifying licenses)
3. Allocation of resources
4. Data security
5. Ethical aspects
6. Other relevant aspects

DMP Tools

DMPTool (MIT, US): https://dmptool.org/

DMP Online: (DCC, UK): https://dmponline.dcc.ac.uk/

Eina DMP (CSUC, Cat) https://dmp.csuc.cat/

ARGOS (OpenAire, EU): https://argos.openaire.eu/home

DMP Guides

Data Management at Horizon 2020 Online Guide

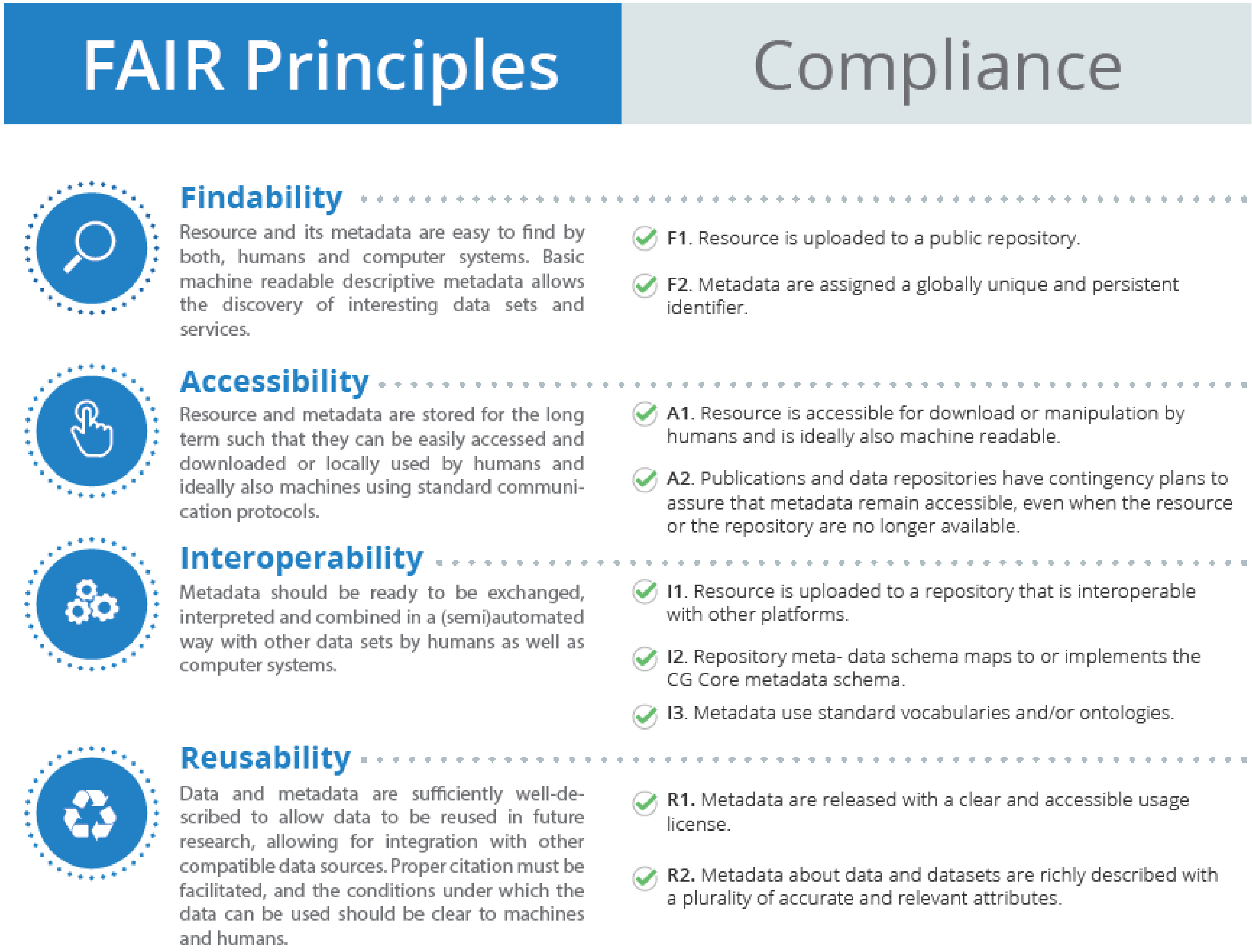
https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

5. The FAIR principles

FAIR Data Principles (Findable, Accessible, Interoperable, Re-usable) support knowledge discovery and innovation as well as data and knowledge integration and promote sharing and reuse of data. The principles help data and metadata to be ‘machine readable’, supporting new discoveries through the harvest and analysis of multiple datasets.

If you cannot make your data open, at least make it FAIR!

Data management workshop - 5. The FAIR principles



FAIR Principles Compliance

Findability

Resource and its metadata are easy to find by both, humans and computer systems. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.

✓ **F1**. Resource is uploaded to a public repository.

✓ **F2**. Metadata are assigned a globally unique and persistent identifier.

Accessibility

Resource and metadata are stored for the long term such that they can be easily accessed and downloaded or locally used by humans and ideally also machines using standard communication protocols.

✓ **A1**. Resource is accessible for download or manipulation by humans and is ideally also machine readable.

✓ **A2**. Publications and data repositories have contingency plans to assure that metadata remain accessible, even when the resource or the repository are no longer available.

Interoperability

Metadata should be ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets by humans as well as computer systems.

✓ **I1**. Resource is uploaded to a repository that is interoperable with other platforms.

✓ **I2**. Repository meta- data schema maps to or implements the CG Core metadata schema.

✓ **I3**. Metadata use standard vocabularies and/or ontologies.

Reusability

Data and metadata are sufficiently well-described to allow data to be reused in future research, allowing for integration with other compatible data sources. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines and humans.

✓ **R1.** Metadata are released with a clear and accessible usage license.

✓ **R2.** Metadata about data and datasets are richly described with a plurality of accurate and relevant attributes.

https://ccafs.cgiar.org/open-access-and-fair-principles

Data creation and organization

(file formats, file naming, documenting data and software, …)

- File naming: avoid special characters and give a meaningful name (audit trail)
    - File name e.g. IP02R0120180731.docx
    - File name components I P02 R01 20180731
    - I = interview (type of data)
    - P[n] = participant ID - participant 02
    - R[n] = researcher ID – researcher 01
    - Date of interview in form YYYYMMDD – 20180731
    - Compare with the file name interview02.docx
- Recommended file formats to allow interoperability

  Use:

  Common formats (may be de facto standards)

  Standard formats accepted in your field

  Interchangeable or open (published) formats for long-term preservation

  Avoid :

  Dependency on proprietary software to render your data

- Documentation & metadata

Documentation

The human-readable stuff that contextualises the research outputs and processes so your future self or others can understand how you got your findings and/or how the data can be repurposed.

- Data dictionary
- Readme file
- Study protocol
- Methodology statement
- Sampling frame description

Format: often text or PDF

Metadata

Machine-readable, standardised fields that allow discovery through search engines, or mark up the structure of a database, or show relationships between different digital objects.

- Dublin Core (DCMI)
- DataCite
- MI, Minimum Information Standards in Biosciences, e.g. MIAME, gene expression microarray.
- RDA | Metadata Directory: http://rd-alliance.github.io/metadata-directory/

Format: often XML or JSON

- Use of data management tools
  - ✓ such as Electronic Laboratory Notebooks (ELN)

When open data access is not possible

- Because potential risk / harm to research subjects is too great.
  - Information that can be used to discriminate requires extra protection.
- When it is not permitted by the data producer, funder, health authority etc.
  - Sometimes precautions are required even for anonymized data.
- Because anonymization is either not feasible or would negate the value of a dataset.
  - Population too small to be anonymous, e.g. those with a rare genetic condition.

6. IBEC’s Research Data Management Policy, procedures and tools

- IBEC’s Research Data Management Policy: https://ibecbarcelona.eu/wp-content/uploads/2021/11/IBEC-Data-Management-Policy.pdf

As a result of its activity, IBEC generates research data, defined as all information (independent of form or presentation) needed to support or validate the development, results, observations or findings of a research project, including contextual information. Research data include all materials which are created in the course of academic work, including digitization, records, source research, experiments, measurements, surveys and interviews. This includes software and code. Research data can take on several forms: during the lifespan of a research project, data can exist as gradations of raw data, processed data (including negative and inconclusive results), shared data, published data and Open Access published data, and with varying levels of access, including open data, restricted data and closed data.

Data management workshop - 6. IBEC’s Research Data Management Policy, procedures and tools

- Procedures, recommendations and guides - undergoing
- Service: knowledge manager from Strategic Initiatives Unit
  - Advice on DMP elaboration in collaboration with Projects Office
  - Assistance in data curation and publishing
- Open Science section at new IBEC website (by the end of April 2022)

7. Choosing a data repository

Data repositories

Most repositories will expect data to be deposited in preferred preservation formats (to enable reuse also in the long term) and for accompanying high-quality documentation to enable correct use of the data.
Good repositories will assign a **unique permanent identifier**, display a clear **reuse license** and data **citation format**.

Trusted cross-disciplinary repositories:

- Zenodo
- Figshare
- Open Science Framework
- Harvard Dataverse
- EUDAT's B2Share

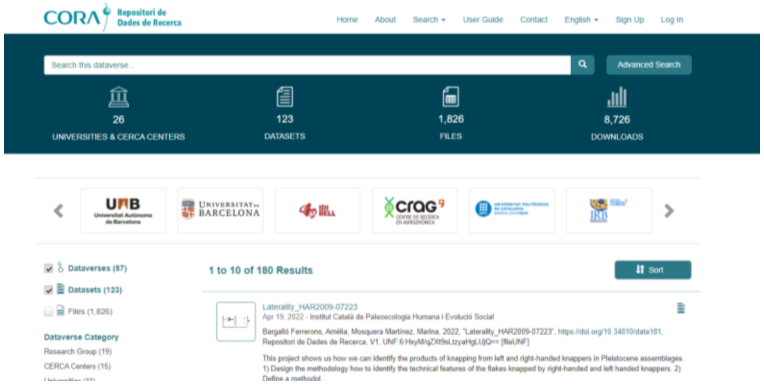
Discipline-specific repositories can be found via lists advised by publishers:

- Nature recommended data repositories
- PLoS One recommended repositories
- Springer Nature recommended repositories
- F1000Research approved repositories

or via re3data.org, a registry of over 2500 research data repositories.

CSUC Data repository: CORA. Repositori de Dades de Recerca

https://dataverse.csuc.cat/


CORA Repositori de Dades de Recerca
Home
About
Search ▾
User Guide
Contact
English ▾
Sign Up
Log In
Search this dataverse...
Advanced Search
26
UNIVERSITIES & CERCA CENTERS
123
DATASETS
1,826
FILES
8,726
DOWNLOADS
UAB Universitat Autònoma de Barcelona
UNIVERSITAT DE BARCELONA
IDIBELL
crag CENTRE DE RECERCA EN AGRIGENÒMICA
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
IRB
Sort

☑ Dataverses (57)
☑ Datasets (123)
□ Files (1,826)

Dataverse Category
Research Group (19)
CERCA Centers (15)
Universities (11)

1 to 10 of 180 Results

Laterality_HAR2009-07223
Apr 19, 2022 - Institut Català de Paleoecologia Humana i Evolució Social
Bargalló Ferrerons, Amèlia; Mosquera Martínez, Marina, 2022, "Laterality_HAR2009-07223", https://doi.org/10.34810/data181, Repositori de Dades de Recerca, V1, UNF:6:HxyM/qZXt9sLtzyaHgLUjQ== [fileUNF]
This project shows us how we can identify the products of knapping from left and right-handed knappers in Pleistocene assemblages. 1) Design the methodology how to identify the technical features of the flakes knapped by right-handed and left handed knappers. 2) Define a methodol...

The CORA. Repositori de dades de Recerca is a repository of open, curated and FAIR data that covers all academic disciplines. CORA. Repositori de dades de Recerca is a shared service provided by participating Catalan institutions (Universities and CERCA Research Centers). The repository is managed by the CSUC and technical infrastructure is based on the Dataverse application, developed by international developers and users led by Harvard University (https://dataverse.org).

Data management workshop

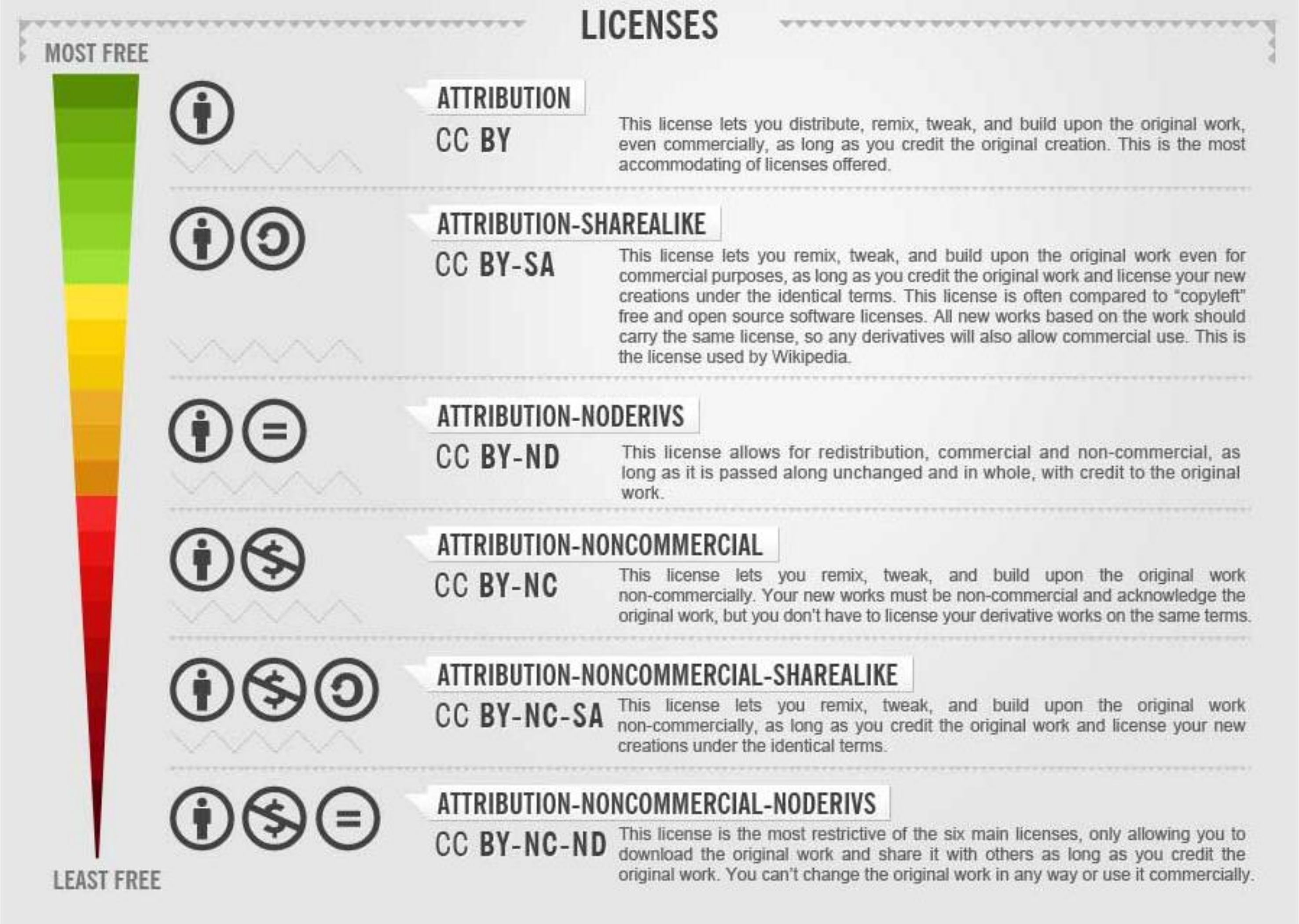
8. Licenses and copyright

A piece of content or data is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and share-alike: conformant licenses: https://opendefinition.org/licenses/

- Open Data Commons is the home of a set of legal tools and licenses to help you publish, provide and use open data. https://opendatacommons.org/

- Copyright law gives creators certain kinds of control over their creative work. If people want to use copyrighted work, they often have to ask for permission from the creator. Creative Commons works within copyright law. It allows creators to grant permission to everyone in the world to use their work in certain ways. https://creativecommons.org/about/cclicenses/

Data management workshop - 8. Licenses and copyright



LICENSES

MOST FREE

ATTRIBUTION

CC BY

This license lets you distribute, remix, tweak, and build upon the original work, even commercially, as long as you credit the original creation. This is the most accommodating of licenses offered.

ATTRIBUTION-SHAREALIKE

CC BY-SA

This license lets you remix, tweak, and build upon the original work even for commercial purposes, as long as you credit the original work and license your new creations under the identical terms. This license is often compared to "copyleft" free and open source software licenses. All new works based on the work should carry the same license, so any derivatives will also allow commercial use. This is the license used by Wikipedia.

ATTRIBUTION-NODERIVS

CC BY-ND

This license allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole, with credit to the original work.

ATTRIBUTION-NONCOMMERCIAL

CC BY-NC

This license lets you remix, tweak, and build upon the original work non-commercially. Your new works must be non-commercial and acknowledge the original work, but you don't have to license your derivative works on the same terms.

ATTRIBUTION-NONCOMMERCIAL-SHAREALIKE

CC BY-NC-SA

This license lets you remix, tweak, and build upon the original work non-commercially, as long as you credit the original work and license your new creations under the identical terms.

ATTRIBUTION-NONCOMMERCIAL-NODERIVS

CC BY-NC-ND

This license is the most restrictive of the six main licenses, only allowing you to download the original work and share it with others as long as you credit the original work. You can't change the original work in any way or use it commercially.

LEAST FREE

(Creative Commons Infographic from: Technology Enhanced Learning Blog) www.ibecbarcelona.eu

Generalitat de Catalunya

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA