

Data Management Plans for Doctoral Students

Ignasi Labastida i Juan
CRAI Universitat de Barcelona



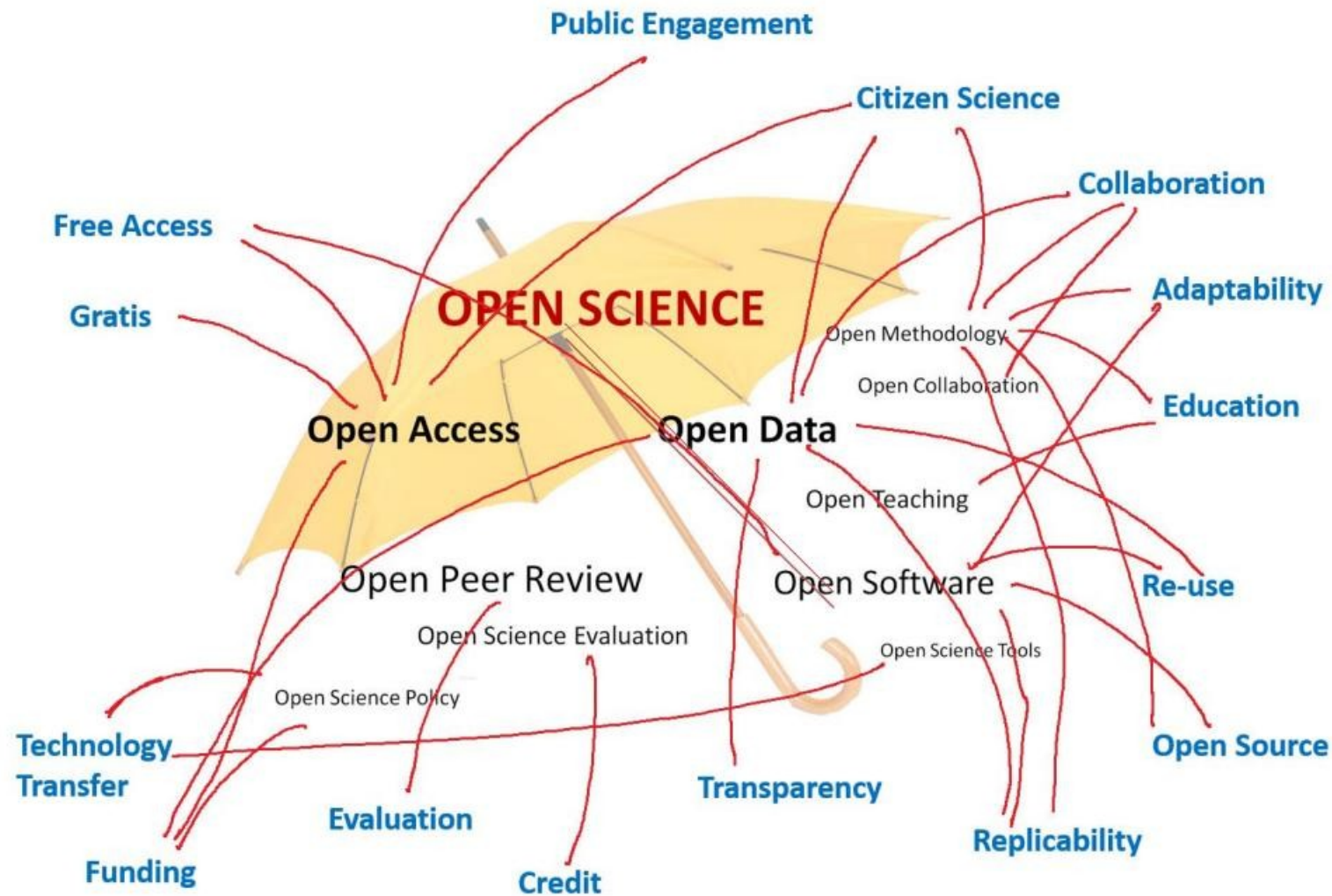
The context

Open Science



J. C. Burgelman, Comissió Europea, Science 2.0

The «complexity» of Open Science



Koen Vermeier

https://www2.helsinki.fi/sites/default/files/atoms/files/koen_vermeir_open_science_helsinki_newest2.pdf

Roadmaps, policies, commitments...



PUSHING
THE FRONTIERS
OF INNOVATIVE
RESEARCH

ADVICE PAPER
NO.24 - MAY 2018

Open Science and
its role in universities:
A roadmap for cultural change

LEAGUE OF EUROPEAN RESEARCH UNIVERSITIES

University of Amsterdam • Universitat de Barcelona • University of Cambridge • University of Copenhagen • Trinity College Dublin •
University of Edinburgh • University of Freiburg • Université de Genève • Universität Heidelberg • University of Helsinki •
Universiteit Leiden • KU Leuven • Imperial College London • University College London • Lund University • University of Milan •
Ludwig-Maximilians-Universität München • University of Oxford • Sorbonne University • Université Paris-Sud •
University of Strasbourg • Utrecht University • University of Zurich

Compromisos de las
universidades ante
la **Open Science**



 **crue**
Universidades
Españolas

Research Data are key for Open Science



This Call for Action is the main result of the Amsterdam conference on 'Open Science-From Vision to Action' hosted by the Netherlands' EU presidency on April 2016

A multi-actor approach was formulated to reach two important pan-European goals for 2020:

1. Full open access for all scientific publications

This requires leadership and can be accelerated through new publishing models and compliance with standards set.

2. A fundamentally new approach towards optimal reuse of research data

Data sharing and stewardship is the default approach for all publicly funded research. This requires definitions, standards and infrastructures.

LERU Roadmap for Cultural Change (2018)

	Topic	Question	Assesment of progress	RAG Status
FAIR data				
10.	Institutional policy	Has your institution a research data policy or strategy?		
11.	Institutional policy	Does your institution research data policy or strategy include FAIR principles?		
12.	Institutional support	Has your institution established a dedicated service to provide data stewardship to its researchers?		
13.	Infrastructure	Does your institution provide access to an infrastructure storage and publication of research data ? If it does not, does your institution inform its researchers of available infrastructures that follow the FAIR principles?		
14.	Data	Does your institution gather information about the data archived and published by its research community?		
15.	Metadata	Does your institution publish all metadata about research data generated or obtained within its research community?		
16.	Assessment	Does your institution include research data as a valuable output in research assessments?		

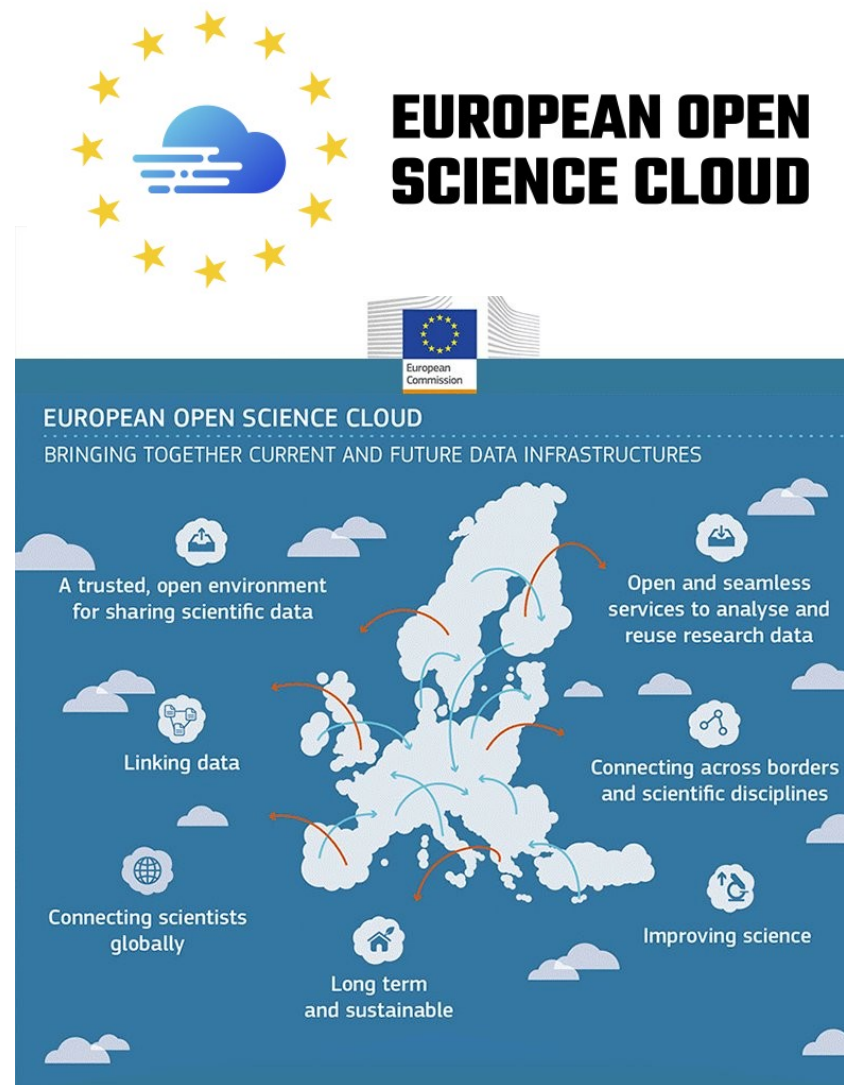
<https://www.leru.org/publications/open-science-and-its-role-in-universities-a-roadmap-for-cultural-change>

Commitments from CRUE (2019)

Promoting collaboration with the relevant national entities to establish a national infrastructure, shared by universities and research centers, and integrated into EOSC, for the storage, management and publication of scientific data from thematic areas not covered by European infrastructures already integrated into EOSC.

<https://www.crue.org/documentos-openscience/>

Ultimate goal: EOSC



Research Data Management

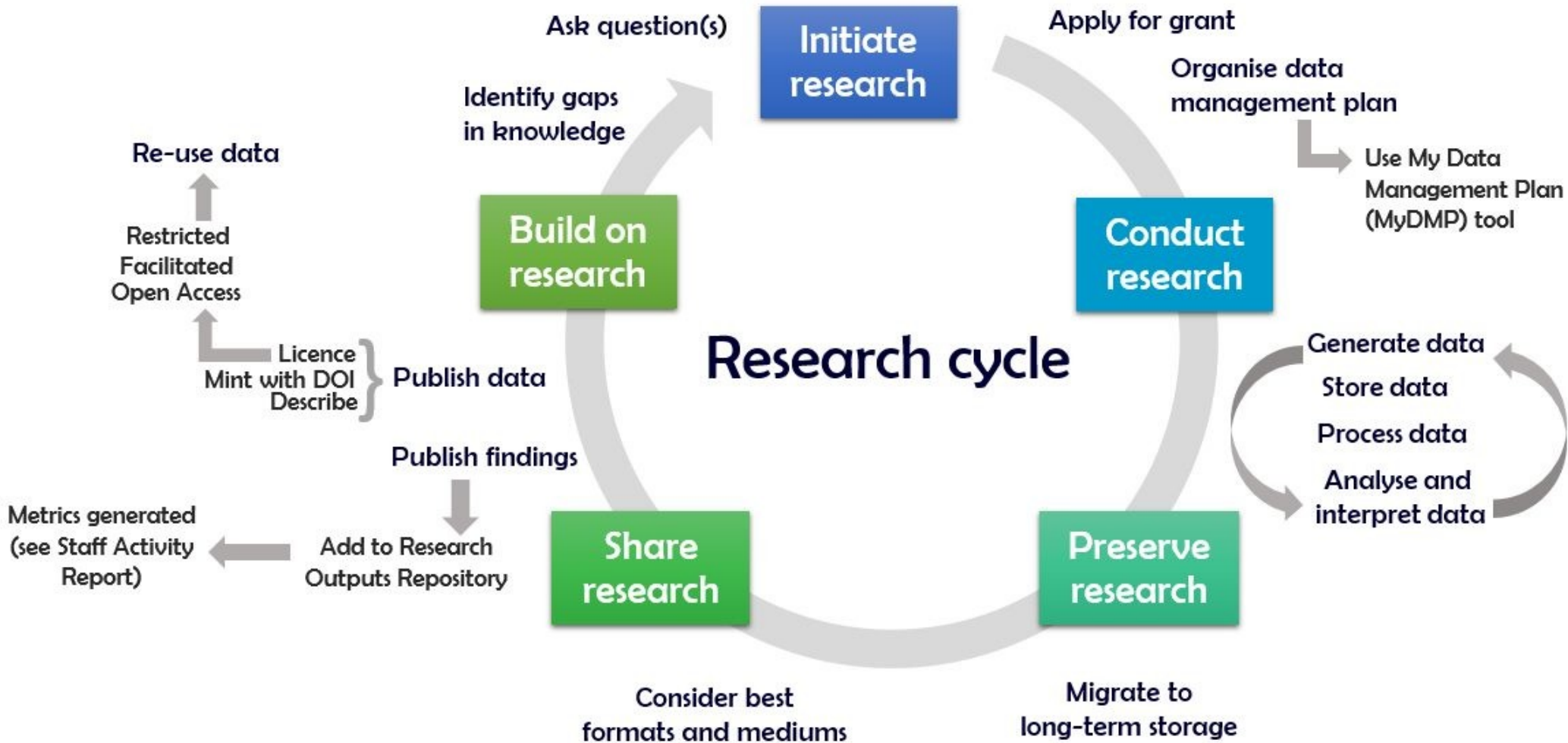
What do we understand as research data?

Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. All other digital and non-digital content have the potential of becoming research data.

Research data may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data, or repurposed data.

<https://casrai.org/term/research-data/>

Managing data during the research cycle



<http://guides.library.unisa.edu.au/ResearchDataManagement>

What is a Data Management Plan?

A Data Management Plan

A DMP is a

- Dynamic document

- Useful document

A DMP is not another bureaucratic burden

A DMP is aimed at facilitating research, foreseeing possible problems, analysing future situations ...

A Data Management Plan

A DMP needs to include

Which data will be produced, gathered, observed, reused and in which formats or how existing data will be reused

How data will be processed, stored and preserved

When data will be accessible and who could access to data

The responsible for the data and the rights holder

How ethical questions have been taken into account

How data will be reusable

You need to answer to these questions

Do you use personal or sensible data?

Is your storing facility safe?

Are you using standard formats?

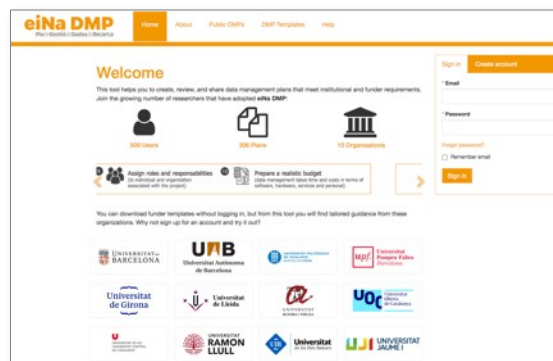
Will your data be understood?

Are you authorised to reuse existing data?

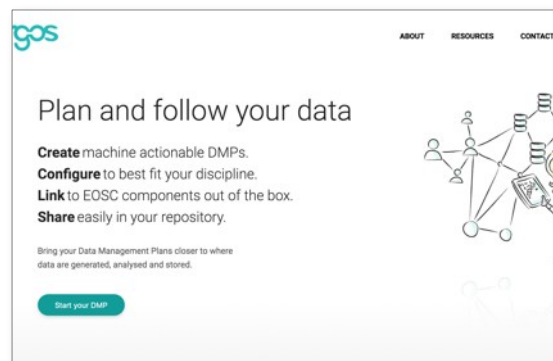
Are you sharing your data with your colleagues?

Are you going to publish your data? When, where and how?

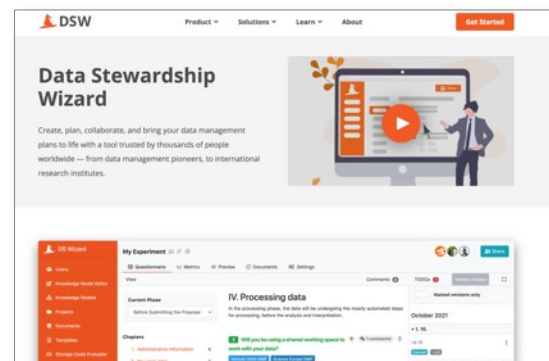
Tools for creating and sharing DMP



EINA DMP
<https://dmp.csuc.cat>

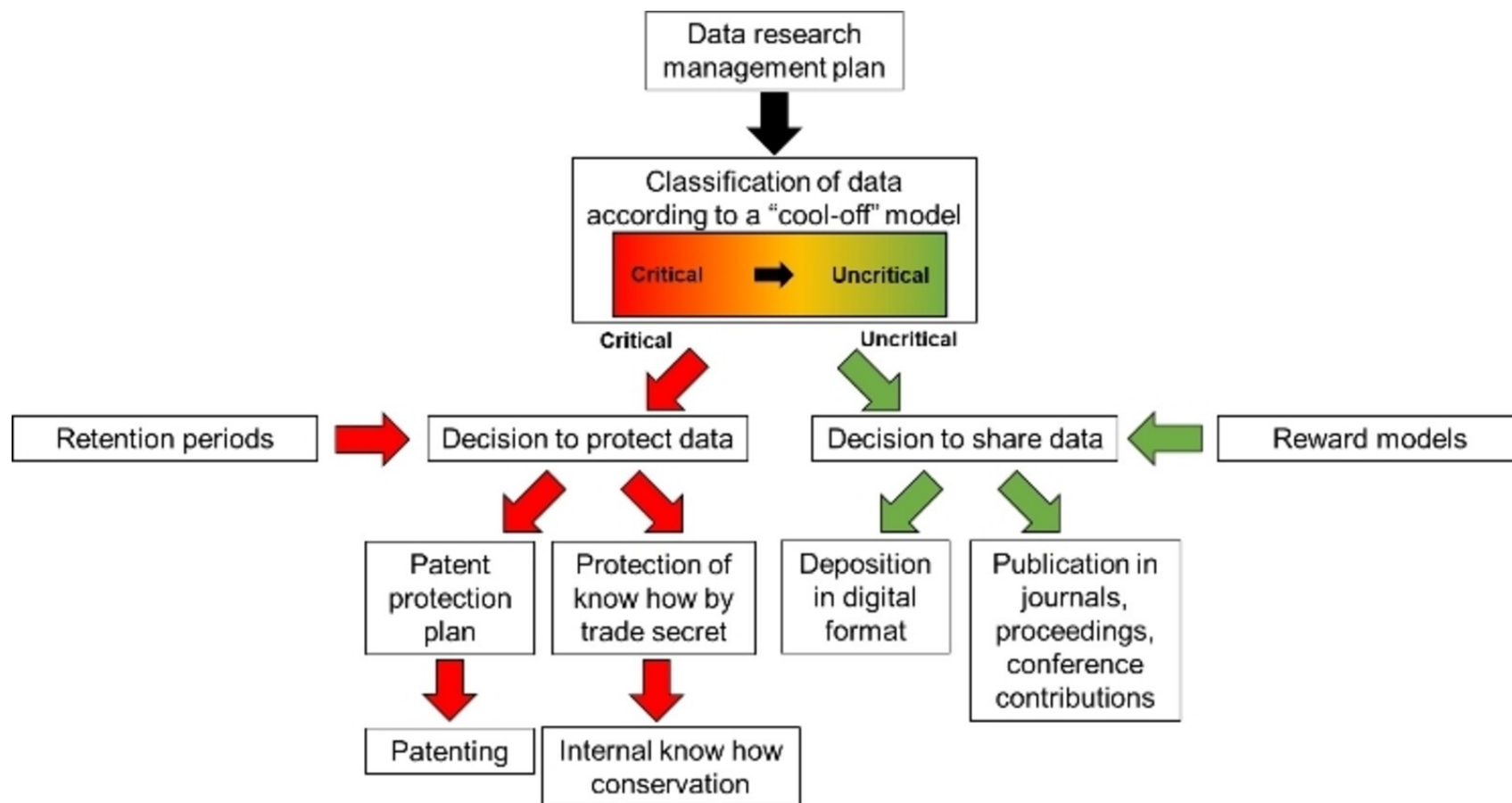


ARGOS
<https://argos.openaire.eu/splash/>



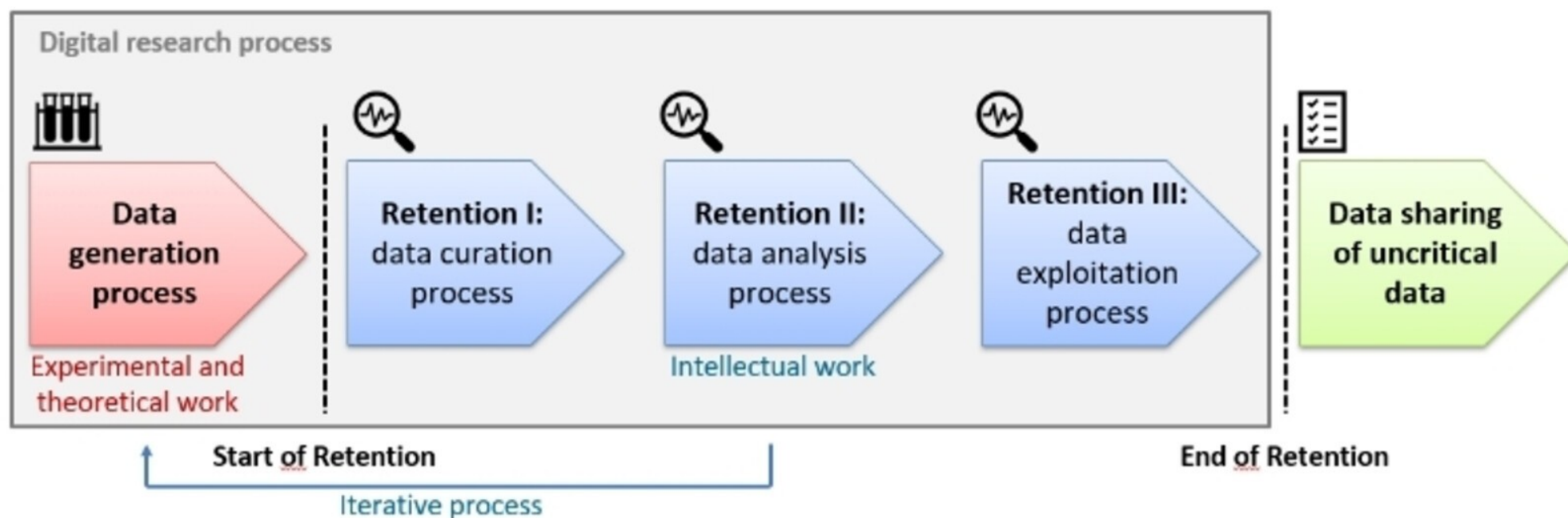
DATA STEWARD WIZARD
<https://ds-wizard.org>

How to create a Data Management Plan



How Research Data Management Plans Can Help in Harmonizing Open Science and Approaches in the Digital Economy
Chemistry A European J, Volume: 29, Issue: 9, First published: 29 December 2022, DOI: (10.1002/chem.202202720)

Examples of retention of data



How Research Data Management Plans Can Help in Harmonizing Open Science and Approaches in the Digital Economy
Chemistry A European J, Volume: 29, Issue: 9, First published: 29 December 2022, DOI: (10.1002/chem.202202720)

Institutional requirements

Institutional Policy (2019)

Política de gestió de dades de recerca pel aprovada Consell de Govern el 22 de juliol de 2019 (<http://hdl.handle.net/2445/142043>)

La gestió de les dades de recerca a la Universitat de Barcelona assegurarà que aquestes siguin trobables, accessibles, interoperables i reutilitzables, segons els principis **FAIR**

Les dades s'emmagatzemaran tan aviat com siguin obtingudes en un espai d'accés restringit als investigadors que hi treballen. Les dades s'enviaran a un dipòsit reconegut com a màxim **sis mesos** després de finalitzar-se la recerca i caldrà fer-les públiques quan es publiquin els resultats de la recerca desenvolupada.

Llevat dels casos on calgui restringir l'accés a les dades per qüestions de **confidencialitat, seguretat, privacitat o per acords amb terceres parts**, les dades que es publiquin es faran de manera **oberta**

Duties and responsibilities

Towards the institution

To establish a specific service for supporting the management of research data, to facilitate the needed infrastructure to manage data and to train the members of the university

Towards the researchers

Elaboration of data management plans

Collecting, documenting, archiving and publishing research data following the principles of this policy

To follow the regulations and legislation and the ethical principles

Doctoral Regulations

Normativa reguladora del Doctorat a la Universitat de Barcelona

<https://seu.ub.edu/normativaMilloradaAnunciPublic/showPublicacion/515737>

El pla de recerca ha d'incloure, com a mínim, el títol provisional de la tesi, els objectius i/o les hipòtesis del projecte, el marc teòric, la metodologia que s'emprarà, els mitjans, els resultats esperats, la bibliografia rellevant, la planificació de la tesi i el calendari previst per elaborar-la. **També ha d'incloure un pla de gestió de dades** i les consideracions ètiques de la recerca, si escau..

A proposal for Doctoral Students

Data Management Plan for Doctoral Students

Type of data to be worked with and methods used for data collection, processing or re-use

Where and how the data will be stored during the doctoral research process

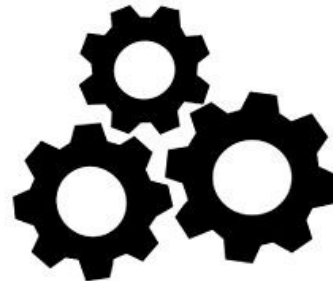
Long-term availability and storage of data

<https://crai.ub.edu/en/services-and-resources/research-data-management-support>

A bit more of information

FAIR Data?

F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}



CC BY-SA SangyaPundir
https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg

FAIR Principles

Findable

- Metadata, identifiers, names, version control

Accessible

- Open by default? Where will be deposited? Needed software

Interoperable

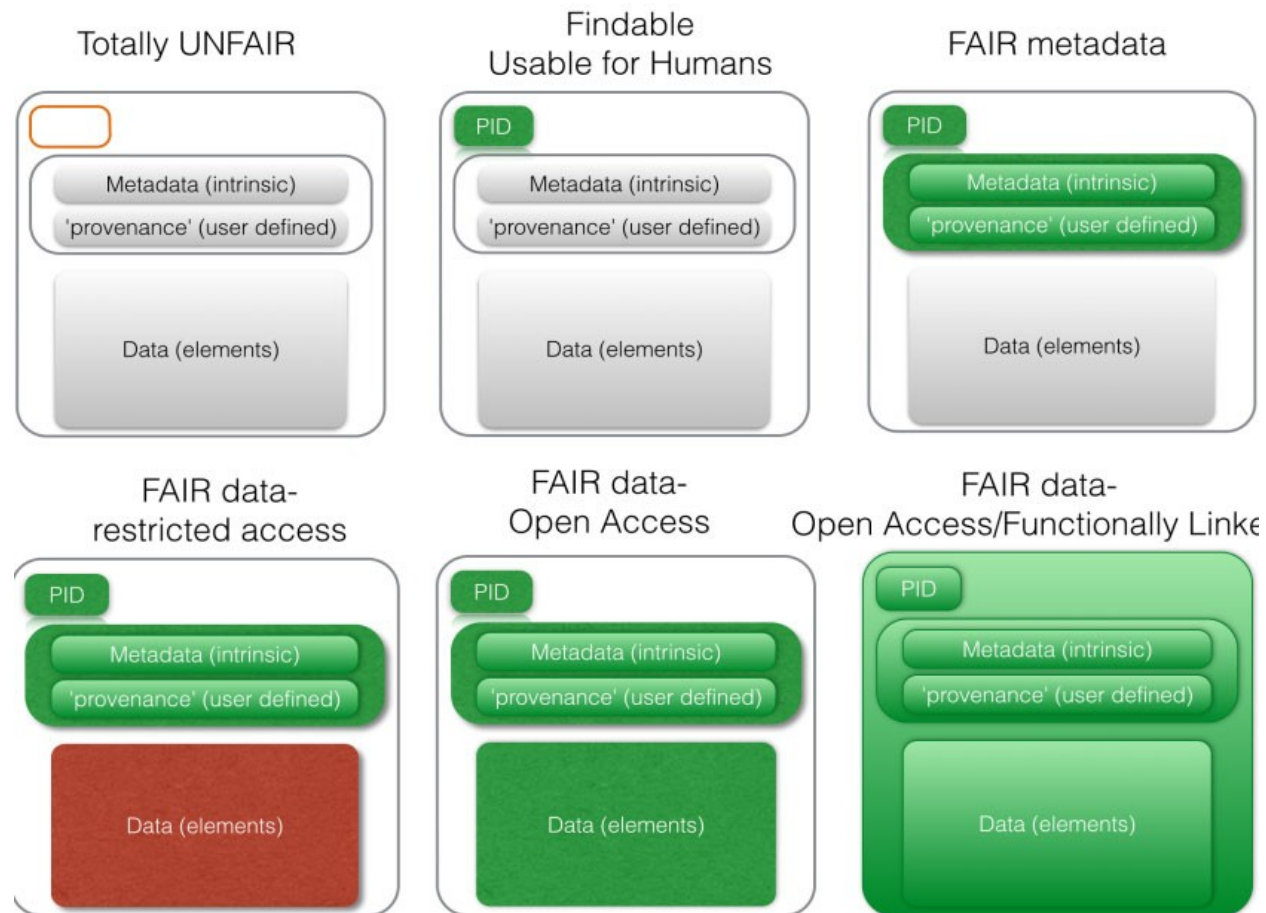
- Data sharing among people and machines, standards

Reusable

- Licenses, embargo periods, conditions

Degrees of FAIR Data

Data as increasingly FAIR Digital Objects



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

Sharing data publicly

Subject based repositories (<http://www.re3data.org/>)

Institutional repositories

National repositories (<http://easy.dans.knaw.nl>)

Zenodo (<http://zenodo.org>)

Figshare (<http://figshare.org>)

Mendeley Data (<https://data.mendeley.com/>)

The institutional solution for research data

The screenshot shows the CORA Repositori de Dades de Recerca website. The header includes the CORA logo and navigation links: Home, About, Search, User Guide, Contact, Sign Up, and Log In. A search bar is present with the text "Search this dataverse..." and an "Advanced Search" button. Below the search bar, four statistics are displayed: 23 UNIVERSITIES & CERCA CENTRES, 92 DATASETS, 1,088 FILES, and 4,109 DOWNLOADS. A horizontal carousel shows logos of partner institutions: UNIVERSITAT DE BARCELONA, UAB (Universitat Autònoma de Barcelona), UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH, upf. (Universitat Pompeu Fabra Barcelona), Universitat de Girona, and Universitat de Lleida. On the left, a filter sidebar shows "Dataverses (40)", "Datasets (92)", and "Files (1,088)". Below this, a "Dataverse Category" section lists: CERCA Centres (16), Universities (10), Research Group (6), Organization or Institution (3), and Research Project (3). The main content area shows "1 to 10 of 132 Results" with a "Sort" button. The first result is "Data of Hepatitis A virus and Norovirus titers from heat-inactivation experiments in clams" by Nov 12, 2021 - Universitat de Barcelona. The authors listed are Fuentes, Cristina; Pérez-Rodríguez, Francisco Javier; Sabrià Sunyé, Aurora; Beguiristain, Nerea; Pintó Solé, Rosa María; Guix Armau, Susana; Bosch, Albert. The result includes a DOI link and a brief description of the data measured during viral inactivation experiments.

<https://dataverse.csuc.cat/>

Real cases

An European Project

#	Data type	Origin	WP#	Format
1	Stakeholder contacts collection	Publicly available data	2	.xls
2	Public security services collection	Publicly available data	2	.xls
3	Digital security solutions collection	Publicly available data	2	.xls
4	Quantitative survey data	Primary data	3	.xls + .csv
5	Expert interview data	Primary data	3	.mp3 + .docx + .txt
6	Focus groups data	Primary data	3	.docx + .txt
7	Workshops data	Primary data	3	.docx + .txt
8	Validation cycles data	Primary data	7	xls + .csv

Table 1: Data sets overview

https://www.takedownproject.eu/wp-content/uploads/2018/04/TAKEDOWN_700688_D1.2-Data-management-plan_update_M18.pdf

An European Project

#	Data type	Data openly available (y/n)	Justification	Alternative solution
1	Stakeholder contacts collection	No	Although the contacts of the collection are professionals' contacts that are publicly available, the consortium can't publish them due to potential misuse caused by automated Spam programs.	The statistical information on the stakeholder data (such as how many, from which countries, which professions etc.) will be integrated in the public report D2.6. In case an external institution is looking for contacts in a specific and the coordinator doesn't see any privacy concerns, relevant contacts might be forwarded.
2	Public security services collection	Yes	(not relevant)	(not relevant)
3	Digital security solutions collection	Yes		
4	Quantitative survey data	Yes		
5	Expert interview data	No	The data from the expert interviews (recordings, protocols and transcriptions) will not be published as primary data due to privacy and security concerns. Anonymization is not considered as an alternative, because the sample size allows drawing conclusions on the respondents.	The categorization, analysis and interpretation of the primary data will be accessible in the public report D3.6 (and others) that can be accessed on the TAKEDOWN project
6	Focus groups data	No	The data from the focus groups (recordings, protocols and transcriptions) will not be published as primary data due to privacy and security concerns.	

An European Project

#	Data type	Location	Level of accessibility	Type of availability and required software tools	Information on metadata and additional data information
2	Public Security Services	TD Solutions Platform	Public	Filterable and searchable database; can be accessed with a state-of-the-art webbrowser	No metadata needed; additional information will be provided on the platform
3	Digital security solutions collection	TD Solutions Platform	Validated professionals	Filterable and searchable database; can be accessed with a state-of-the-art webbrowser	No metadata needed; additional information will be provided on the platform
4	Quantitative survey data	https://zenodo.org/	Registered ZENODO users	Cleaned primary data; can be accessed with SPSS, Excel or	Metadata as well as a codebook will be deposited in the data repository Zenodo

A DMP from a Dutch PhD Student

Type of research data	Specification	Software choices	File extension
Model parameter values	I will need to gather information to use as model parameter values, for example on pathogen removal rates for different types of sewage treatment, or the die-off under specific environmental conditions. I plan to make an overview file of different parameter values found in literature or from other sources.	Excel	.csv
Model input data - gridded	Existing datasets on for example climate and hydrology	Depending on how they are available. Runoff data are .grd files, for example, most others I don't know yet	.grd and others
Model input data - country data	Existing datasets on for example population and livestock density, land use	Depending on how they are delivered, perhaps a spread sheet format. I may all convert these to csv files	.csv
Model input data - metadata	For all input data, one document containing all metadata will be created, specifying at least source, time period, region, measurement method, type of data, unit of measurement, access rights, date downloaded. All data will be checked for consistency, and any changes made to input data will be documented.	Excel	.csv
Data for model validation - waterborne pathogen measurements	Various mostly not yet identified sources: different institutes who perform waterborne pathogen measurements	Depending on how they are delivered, probably a spread sheet format. I may all convert these to csv files	.csv
Data for model validation - metadata	For all data for validation, one document containing all metadata will be created, specifying at least source, time period, region, measurement method, type of data, unit of measurement, access rights, date downloaded. All data will be checked for consistency, and any changes made to input data will be documented.	Excel	.csv
Original model scripts	Created by me	R	.R
Adapted model scripts	Created by me	R	.R
Model output - tables	Created by me	R	.csv
Model output - tests	Created by me	R	.csv / .txt
Model output - graphs	Created by me	R	.tiff / .eps
Model output - maps	Created by me	R	undetermined
Manuscripts	Created by me: manuscripts of papers	Word	.docx
Other	Administrative documents, coursework-related documents	Adobe, Word, Excel	.pdf / .docx / .xlsx

https://www.wur.nl/upload_mm/d/4/9/fec82c96-d901-49c2-8205-c40a2e819ee7_Data_Management_Plan_Lucie_Vermeulen.pdf

Gràcies

Ignasi Labastida i Juan
CRAI Universitat de Barcelona

ilabastida@ub.edu