

FAIR DATA & FAIR-by-design methodology

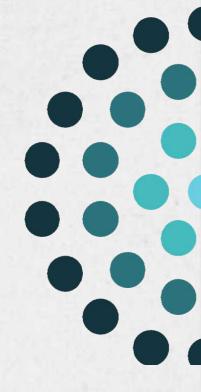
In the Open Science Orbit - Introduction to Open and Responsible Research

Workshop for experienced researchers

04 February 2025



Eva Méndez (@evamen)
Universidad Carlos III de Madrid











Funded by the European Union. Views and opinions expressed are however those of the auhor(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.





Stakeholders

Open Science Challenges

Shifts in the data landscape in the last 10 years

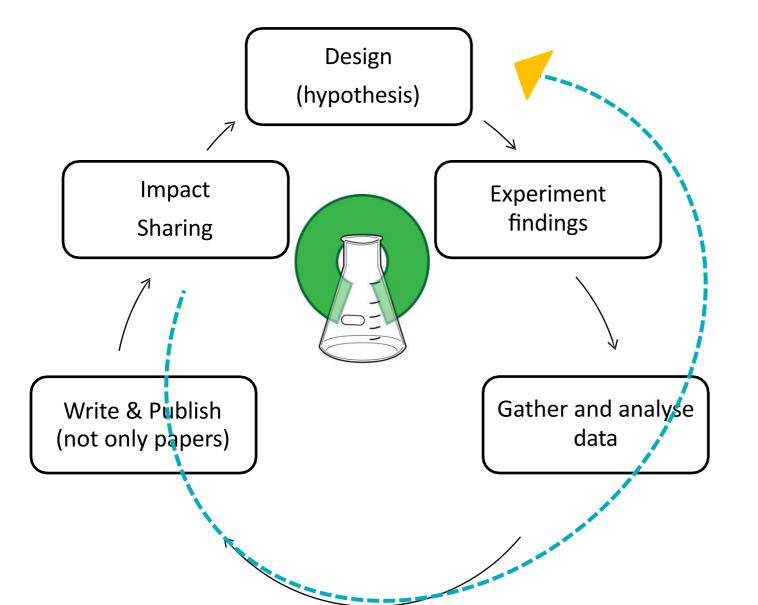
- From big data → to really really BIG data
- From USBs to Cloud and HPC
- Government (EU) open data policies and platforms
- Increase in compute power
- Data and software as a first-class research output
- FAIR Data Principles in practice
- Journal data policies (but...)
- More standards, schemas
- Increased use and value in PIDs, vocabularies, ontologies,
- More data discovery portals (Google Dataset search)
- OpenAI and LLM in the Research landscape
- ... and so much more







Research life cycle



Based on: Open science image CC BY-SA 3.0 by Greg Emmerich www.flickr.com/photos/gemmerich/6365692655

Change in the Research cycle

Open from the begining

Papers + Data

+

Methods + Code...

Foster reproducibility

FAIR





DATA in the new ORRI and RRA



- Immediate integration of research data in the research evaluation landscape
- Data as an object of research evaluation
- Good research outcomes (publications and data) include good metadata
- (Meta) Data as an input of curated research evaluation sources
- FAIR data as a proxy of good research data
- Al and ML to assist new research evaluation

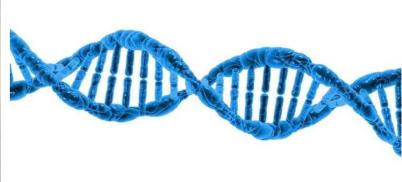
What kind of research data do you work with?

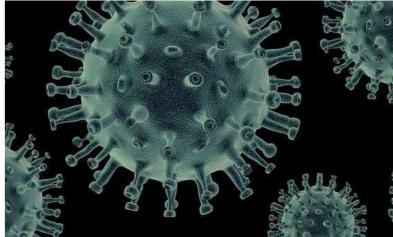












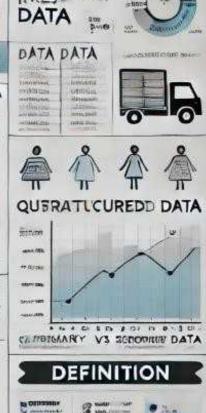
Guardar





DIVERSITY OF RESEARCH CLASSFICTION (350) CLASSFFICATION -CALITERATIVE STRATICURED VS QUALITTY DATA QUATTDATIVE VS DATA STEURETCURED CQUIATTUFED DATA STRATURED VY LINCORRETCORED STRATETURED VS SUPTONRED DATA STRATTURED DUITEDY DATA

PRIMINY VEUBOY DATA manuel obspective central Follow



DIVERSITY



Types of Research Data

By Nature

- Quantitative Data: Numerical and measurable (e.g., statistics, experiments).
- Qualitative Data: Descriptive and narrative (e.g., interviews, field notes).

By Format

- Structured Data: Organized in predefined formats (e.g., databases, spreadsheets).
- Unstructured Data: Not organized in a systematic way (e.g., videos, images).

By Source

- Primary Data: Collected directly by researchers for specific studies.
- Secondary Data: Pre-existing data used for new research purposes.

By Availability/Accessibility

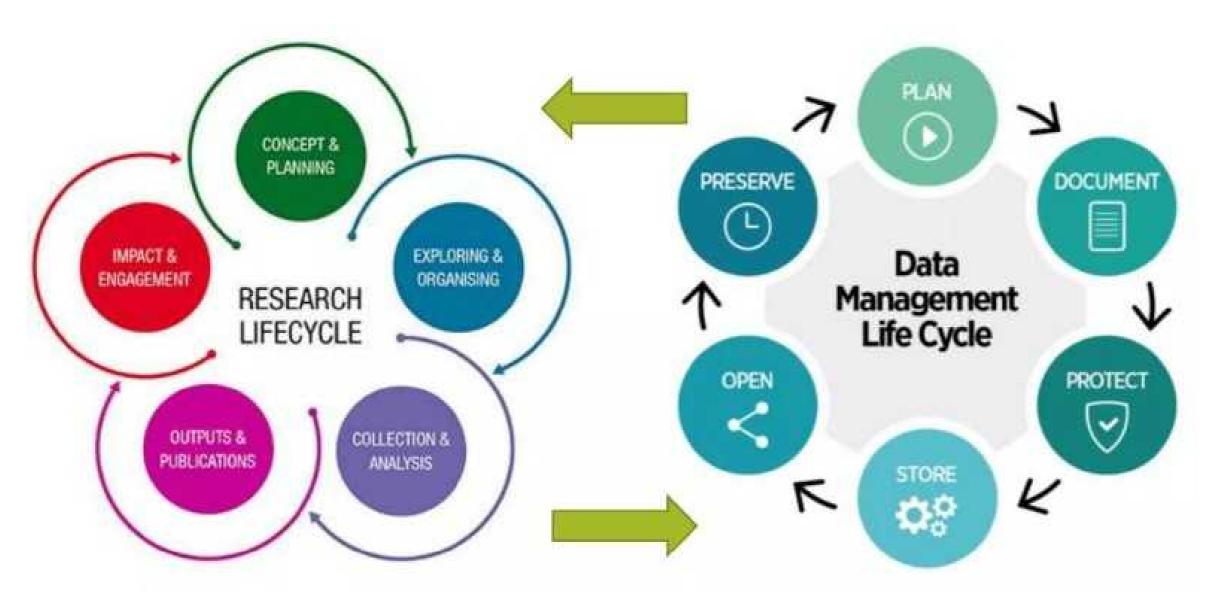
- Open Data: Freely available for use and sharing.
- Restricted Data: Access limited due to privacy, security, or proprietary concerns.



There are 4 types of research data:

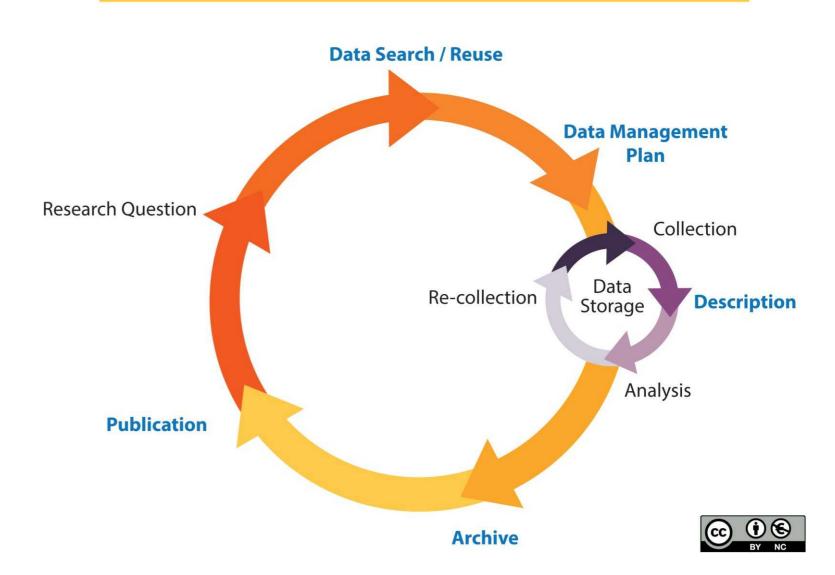


Research.com



The Research Data Management Lifecycle

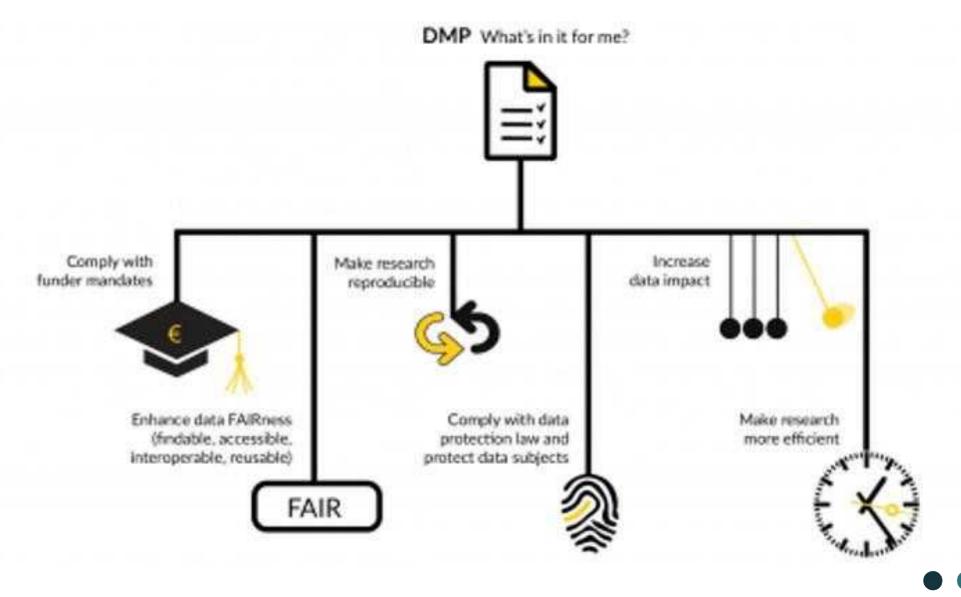
- structure
- documentation
- metadata
- storage
- backup
- long-term archive
- access
- publication





Why?

RDM DMPs FAIR







Where?

e Findeble

· Accessible

9[[G67999al] •

• Bedesple

Why?

Who?

How?





Findable

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

Accessible

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
 - A1.1 The protocol is open, free, and universally implementable
 - A1.2 The protocol allows for an authentication and authorisation procedure
- A2. 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.
- 12. (Meta)data use vocabularies that follow FAIR principles
- 13. (Meta)data include qualified references to other (meta)data

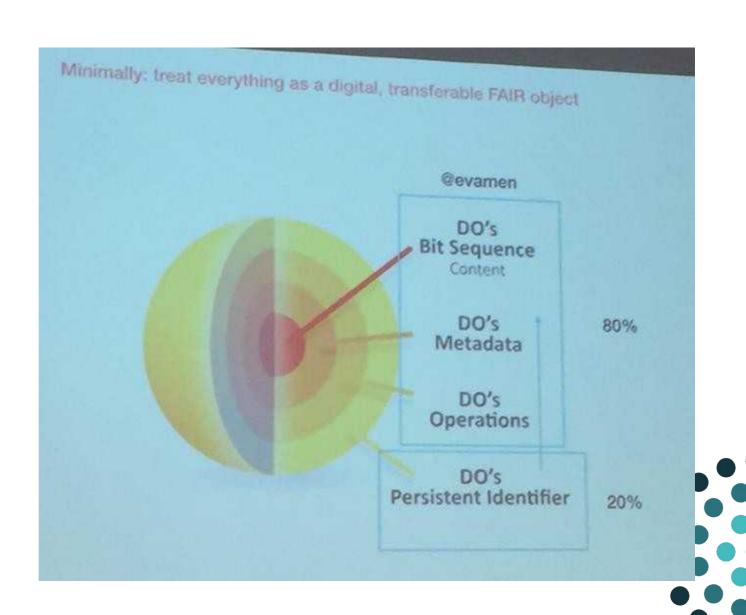


- R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (Meta)data are released with a clear and accessible data usage license
 - R1.2. (Meta)data are associated with detailed provenance
 - R1.3. (Meta)data meet domain-relevant community standards



FAIR Guiding Principles: all about metadata

80% metadata 20% PIDs





How FAIR are your data?

indable

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

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

Accessible

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

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

Interoperable

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

- Data is provided in commonly understood and preferably open formats
- The metadata provided follows relevant standards
- Controlled vocabularies, keywords, thesauri or ontologies are used where possible
- Qualified references and links are provided to other related data

Reusable

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

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

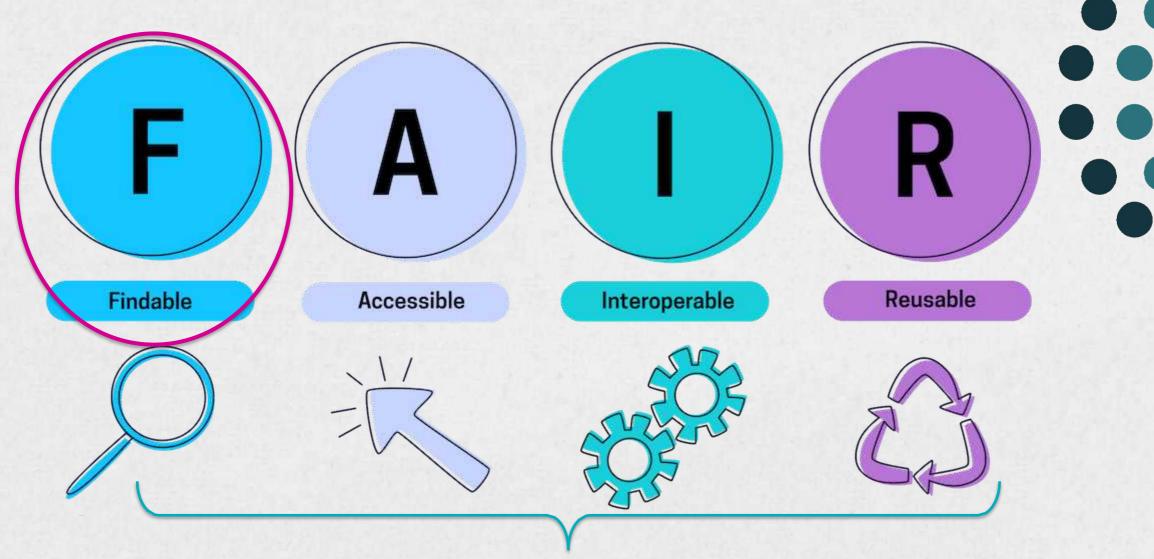


'How FAIR are your data?' checklist, CC-BY by Sarah Jones & Marjan Grootveld, EUDAT. Image CC-BY-SA by SangyaPundir

Metadata at the core of FAIR data

- Making data findable, including provisions for metadata
- What **metadata** will be created? In case
 - metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.
- Where will the data and associated metadata, ... be deposited?
- Interoperability of your data... What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
- The Research Data Alliance provides a Metadata Standards Catalogue that can be searched for discipline-specific standards and associated tools.

How to make your research data FAIR

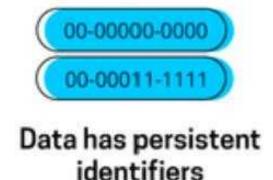


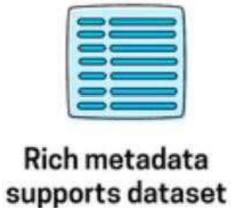
Good research data management

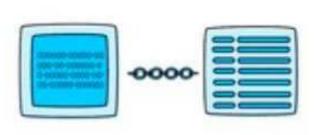


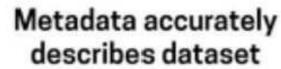
Metadata and data should be **easy to find** for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services.

https://www.go-fair.org/fair-principles











Metadata is indexed and searchable

Source: https://flur.ee/fluree-blog/making-data-fairest





Assign to your data a PID

(DOI or any other persistent identifier)

- OSF HOME

 My Quick Files My Projects Search Support Donate Rebecca Rosenblatt

 The Influence of Reaction Co... Files Wiki Analytics Registrations Contributors Add-ons Settings

 The Influence of Reaction

 Conditions on HMF Oxidation

 Rate

 Contributors: Rebecca Rosenblatt, Sara Bowman

 Affillated Institutions: None

 Date created: 2017-05-18 02:43 PM | Last Updated: 2017-06-28 05:11 PM

 Create DOI
 Category: Project
 Description: Add a brief description to your project

 License: Add a license
- Data should have unique and persistent identifiers (DOI, Handle, ARK, etc.).
- Ex: DOI is a unique alphanumeric string assigned by the International DOI Foundation to identify content and provide a persistent link to its location on the Internet.







Describe your data with (rich) metadata

Variable encoding?

Who collected the data?

Data version?

Time of data collection?

Purpose of data collection?



Dataset dimension?

Data processing?

Place of data collection?

License?

Observation unit?







Describe your data with (rich) metadata

Consider existing metadata standards

Metadata Standards Catalog

Search

Sign in

The Metadata Standards Catalog will be unavailable on 11 February 2025 from 09:00 to 12:00 UTC due maintenance.



https://rdamsc.bath.ac.uk

Metadata Standards Catalog

The RDA Metadata Standards Catalog is a collaborative, open directory of metadata standards applicable to research data. It is offered to the international academic community to help address infrastructure challenges.

Read more details about the scope of the Catalog

Read our terms of use





Find a home for your data

- Domain Agnostic Repositories (Figshare, Zenodo)
- Domain specific repositories: many data repositories are specialized on a field of research









Engineering

Archaeology

Geographic region / university (ej. eCienciaDatos (Madrid)







Find a home for your data

Find domain specific repositories <u>re3data.org</u>





Selecting a trustworthy subject- <u>s</u>
 <u>pecific repository</u> (download the
 report)





Find a home for your data

 What can you do if no specialized data repository exists for your research topic?





University data repositories



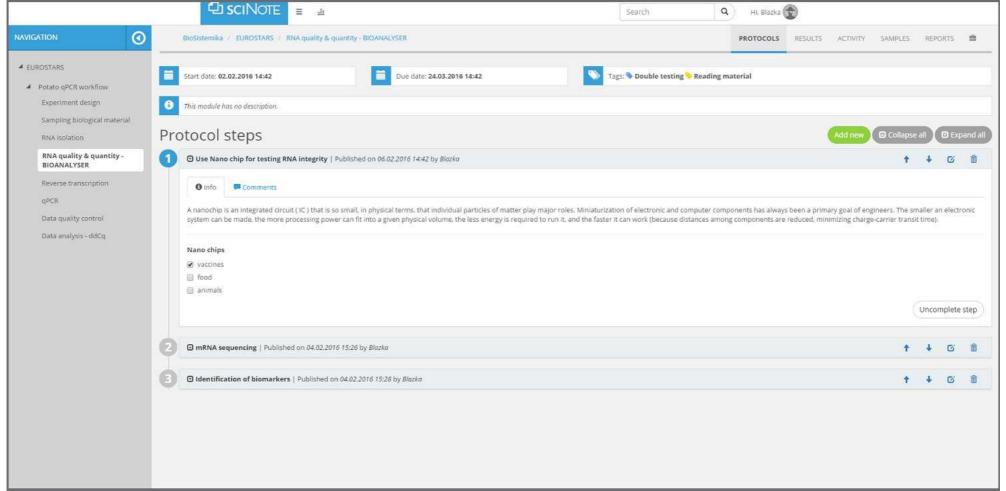
Generic data repositories





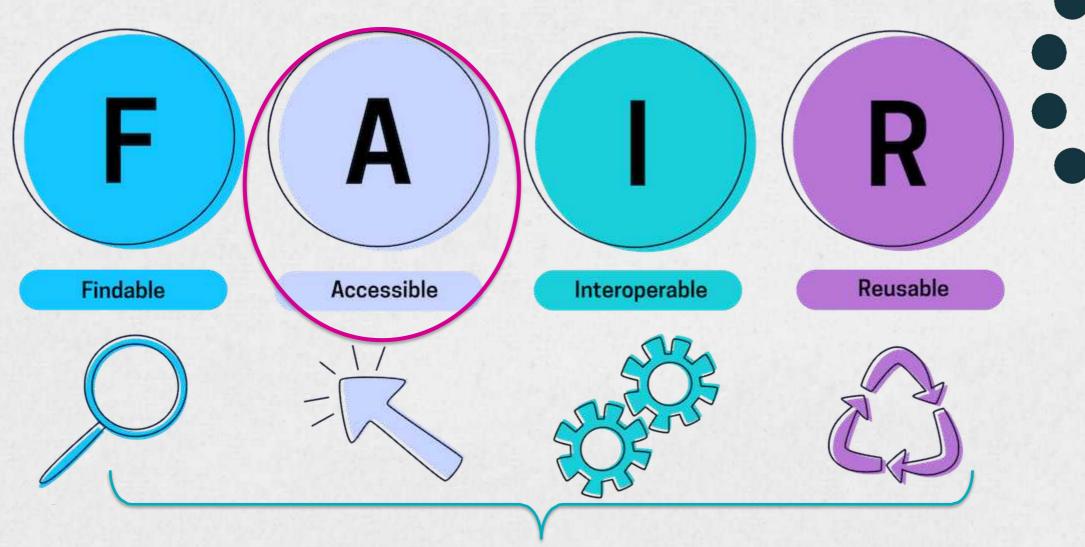


Use tools to improve the findability





How to make your research data FAIR

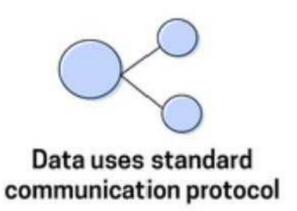


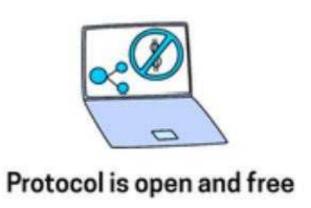
Good research data management



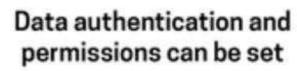
Once the user (researcher) finds the required data, she/he needs to know **how can they be accessed**, possibly including authentication and authorisation.

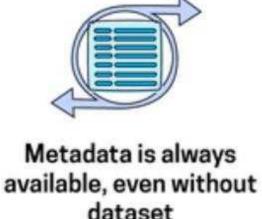
https://www.go-fair.org/fair-principles











Source: https://flur.ee/fluree-blog/making-data-fairest





Always available/accessible

Data should be **available to everyone** to access, use, and share \rightarrow Taking care of the personal information (GDPR) and other sensible information (as open as possible as closed as necessary)









political opinion

national security







Always available/accessible

As open as possible as closed as necessary:

Options for sensitive data

- Anonimisation / Pseudo-anonimisation
- Are all your data sensitive? Maybe you can share some of them.
- Restrict the access to your data to a relevant group (e.g., to researchers) and be clear and transparent about why you restrict the access and how people can gain access.
- Publish only metadata.



How to make your research data FAIR Reusable Interoperable Accessible **Findable**

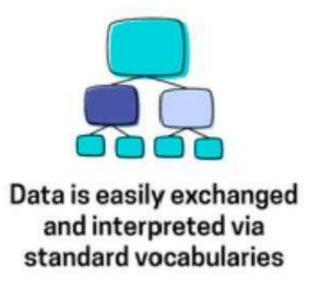
Good research data management

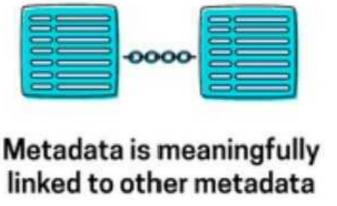


The data usually need to be **integrated with other data**. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

https://www.go-fair.org/fair-principles







Source: https://flur.ee/fluree-blog/making-data-fairest





Use standards to encode data

Do not use data proprietary formats

Format / Software	Proprietary	Open
Text files	Word (.doc), Pages (.pages)	Open Office (.odt), .txt, LaTeX
Spreadsheets	Excel (.xls), Numbers (.numbers)	Open Office (.ods), .csv
Video	.avi, .wmv, .mov, .qtvr, .rv	.mpg, .mp4
Audio	.wma, .asf, .ra, .wav	.mp3
Presentations	PowerPoint (.ppt), Keynote (.key)	PDF, HTML
Statistical Analyses	SPSS (.sav), Matlab (.m), SAS (.sas), Stata (.dta)	R, JASP (.jasp), Python
Experimental Software / Questionnaires	E-Prime, SurveyMonkey, UniPark	PsychoPy, Limesurvey, formr







Use standards to encode metadata

Metadata

are like toothbrushes...



...Everyone thinks that it is a good idea, but nobody wants to use someone else's.



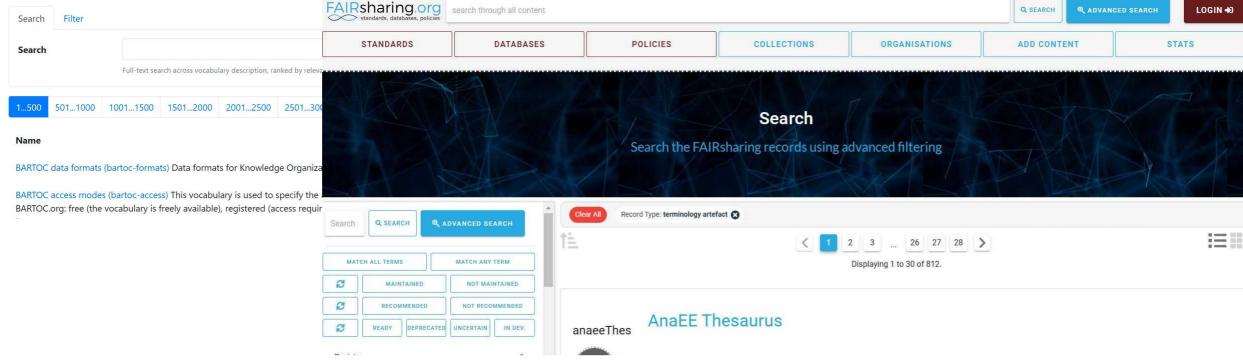


Use vocabularies (Semantic Artefacts)

BARTOC.org Vocabularies Registries Software About Contact

https://bartoc.org/vocabularies

Vocabularies

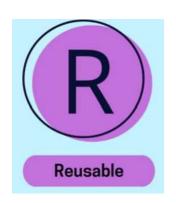


https://fairsharing.org/search?recordType=terminology_artefact

ary for the semantic description of the study of continental ecosystems and their ant main thematic areas: abiotic (e.g. atmosphere, climate, hydrosphere, litospher...

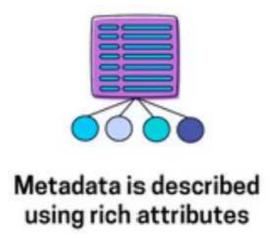
How to make your research data FAIR Interoperable Reusable Accessible **Findable**

Good research data management

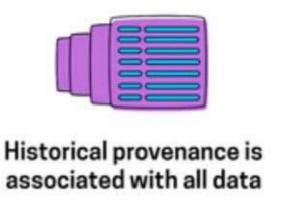


The ultimate goal of FAIR is to **optimise the reuse of data**. To achieve this, metadata and data should be well-described so that they can **be replicated** and/or combined in different settings.

https://www.go-fair.org/fair-principles









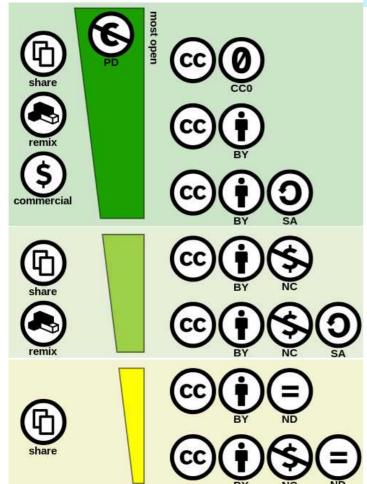
Source: https://flur.ee/fluree-blog/making-data-fairest



Make your data



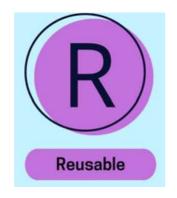
Choose and encode the right license



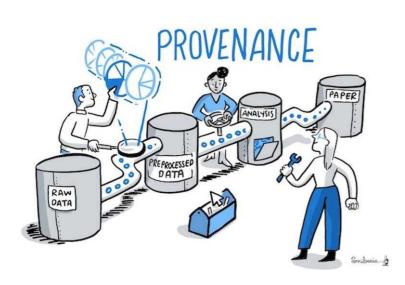
Regulate and describe clearly and appropriately, what others are allowed to do with the data



Make your data



Ensure data provenance and reproducibility



Always comment your code !!!

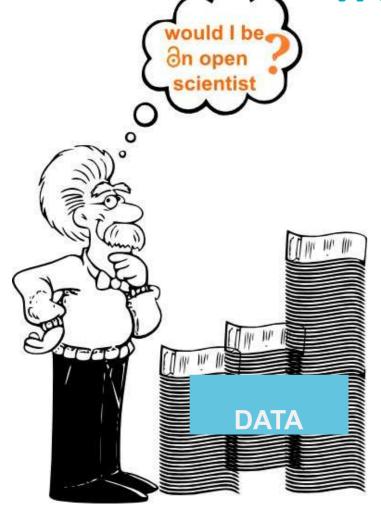
- Choose a coherent file / function naming system and coding style.
- Consider version control.
- Record the used packages and software.
- Write a README with details on the workflow if code fragments need to be combined.







FAIR data cost?



Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

Written by PwC EU Services

Research and Innovation



How to make your RD FAIR, FAIR (EST), CARE...





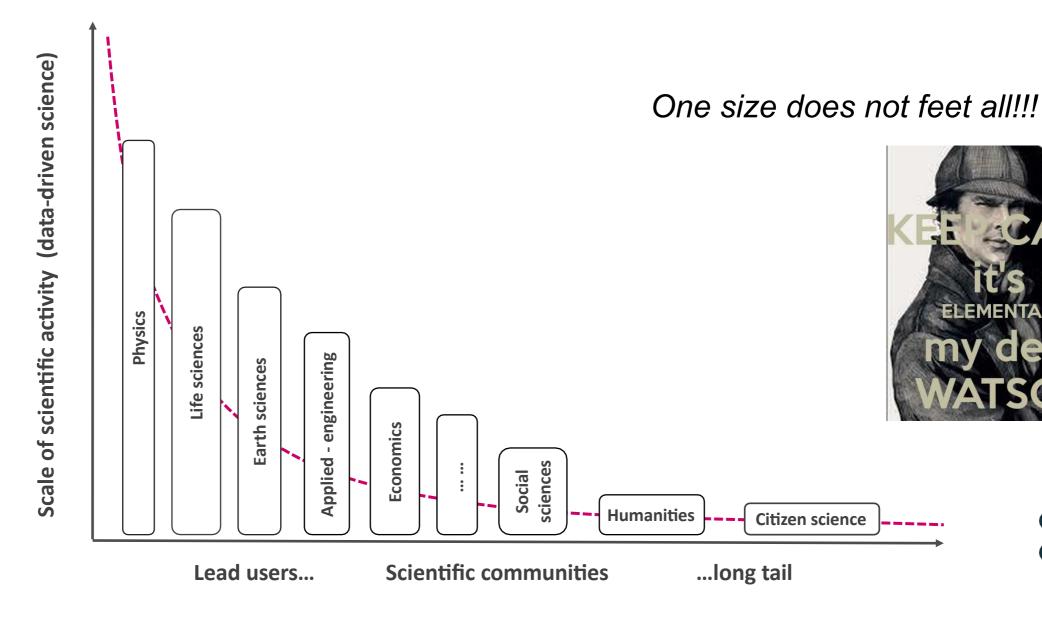
Better (research) data management

Research infrastructures for Open Science and FAIR DATA EOSC



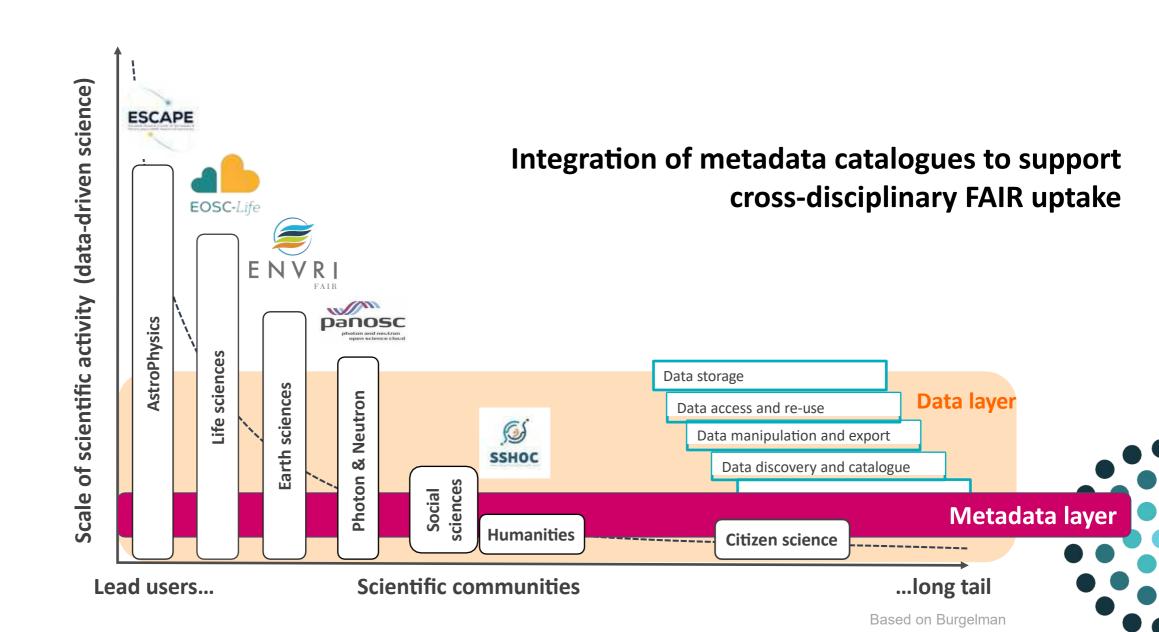


Scientific data landscape (EOSC)





EOSC and FAIR DATA



ENSURE DATA PROVENANCE AND REPRODUCIBILITY



Track what research data was processed by which application and version to ensure reproducibility and data integrity.



RESEARCH & CLINICAL TRIALS

RESEARCH FACILITY DISCOVERY & DISSEMINATION OF RESULTS

CITATION

INVESTMENT & NEW FUNDING

OPERATES UNDER EXISTING RESEARCH TOOLS



Mediaflux operates under existing research tools (electronic lab book, Laboratory Information Systems, etc.), transparently supporting and simplifying researchers work.

DESCRIPTIVE METADATA TAGGING



Collect and store data with descriptive metadata contained to enrich the roadmap, and provide more information to work with during analysis.

ANALYSIS/COLLABORATION



Access and discover data within any other file format with complete data transparency and the sharing of research data and associated metadata.

ALLOWS FINDINGS TO BE MADE FULLY AVAILABLE



Simplify how intellectual property is shared between owners and individuals and institutions.

LONG-TERM STORAGE



Maximise the value of the data throughout the lifecycle by leveraging more efficient data storage and eliminating unnecessary duplicated data.

speose EOSC vision in a nutshell

What

EOSC is the European web of FAIR data and related services for research

Research data that is easy to find, access, interoperate and reuse (FAIR)

Trusted and sustainable research outputs are available within and across scientific disciplines

Why

Unlock the full potential of research data to accelerate discoveries and innovation

How

- Ensure that Open Science practices and skills are rewarded and taught, becoming the 'new normal'
- Enable the definition of standards, and the development of tools and services, to allow researchers to find, access, reuse and combine results
- Establish a sustainable and federated infrastructure enabling open sharing of scientific results



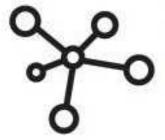
Strategic Research and Innovation agenda (SRIA) eosc.eu/sria-mar





coeosc From vision to implementation: "the EOSC federation"

The EOSC vision is to set up a 'Web of FAIR Data and Services' for science in Europe. Central to this ambition is the deployment of a trusted, virtual, federation of existing infrastructures in Europe to store, share and reuse FAIR research outputs across borders and scientific disciplines also called the "EOSC Federation".



- EOSC is NOT a new digital infrastructure
- The EOSC ambition is to federate existing data, research and e- infrastructures nodes to make them all available to European researchers across borders and across disciplines (distributed EOSC 'system of systems')
- In doing so, the federation will be augmented with new additional services and tools that will enable the EU web of FAIR data and related services (EOSC can be seen as a thin federation layer based on FAIR principles)
- The federation will provide coordinated entry points primarily for researchers in Europe (the so called "nodes") to find and access FAIR data and interoperable services that address elements of the whole research cycle (from discovery and mining to storage, management, analysis, publication, and re-use)
- The entry points for EU researchers will be via their traditional channels (e.g. via the national, regional, pan-European or thematic infrastructure nodes they are currently using) or via the EU EOSC node central instance (for the researchers that do not have existing access channels in place)
- EOSC rules of participation and access policies will be developed for the users and providers of the federation





speose Who is EOSC for?

EOSC aims to support all European researchers

How?

- Enhancing scientific research by providing access to a weath of data and research outputs from various fields and institutions
- Fostering innovation by making it easier for researchers to share, collaborate and build on each other's work
- Improving transparency and reproducibility of research by making data and methods more open and accessible
- Reducing data silos and duplication of efforts by promoting data sharing and reuse
- Facilitating interdisciplinarity research and cross-sectoral collaboration by making data and resources from different domains more easily available



Individual researchers will benefits from EOSC through their existing channels (e.g. universities, research institutes, research infrastructures, associations, science clusters, etc.) that will act as intermediaries.

specific education in the second education education in the second education edu

- EOSC as Open Science (OS) & FAIR culture booster: Working with policy makers and European initiatives (such as the CoARA coalition), EOSC will support the cultural change towards OS and FAIR principles in the EU countries and institutions.
 - All EU researchers will indirectly benefit from the cultural change brought by EOSC. The EOSC vision aims to have Open Science and FAIR practices fully encouraged, supported and recognised at institutional and national level.
- EOSC as point of reference for building the new OS / FAIR related skills and career paths:
 Working closely with universities and research institutes, EOSC will support the creation of the workforce of the future and of the necessary competence support centers.
 - All EU researchers will benefit from the training material, guidelines and support provided by EOSC.
- EOSC as web of FAIR data and related services: The EOSC federation will make available the resources of existing data, research and e-infrastructures across borders and disciplines
 - All the EU researchers will be able to access an extended offer in terms of FAIR data and related services.





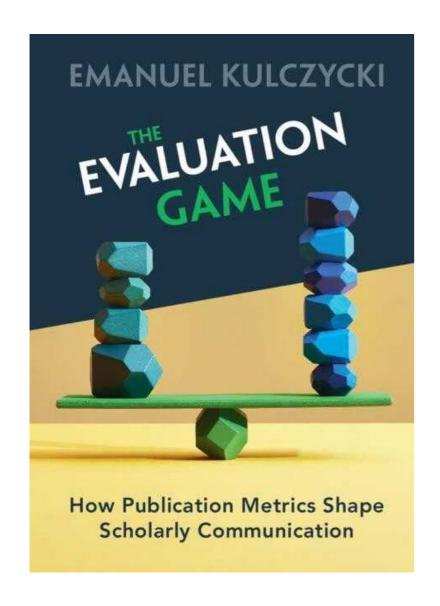


I want to recommend you several resources on Open Science for EARLY CAREER RESEARCHERS



- 1 Book
- 1 MOOC
- 1 Paper
- 1 Passport (extended)
- 1 Guide
- 1 Website
- 2 Communities
- + Several tools





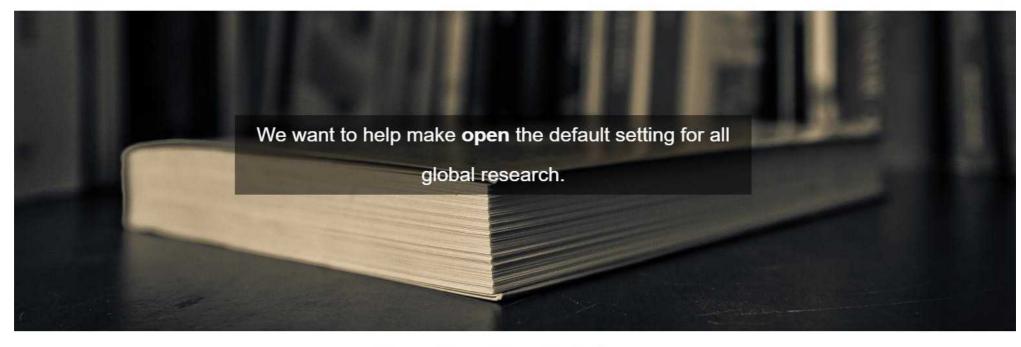
The use of metrics and the focus on increasing productivity impacts the weakest persons in academia most severely, notably young researchers and researchers from peripheral countries (Kulczycki, 2023)



https://opensciencemooc.eu



Home Blog Calendar Modules ▼ People Code of Conduct About



How Can You Help?

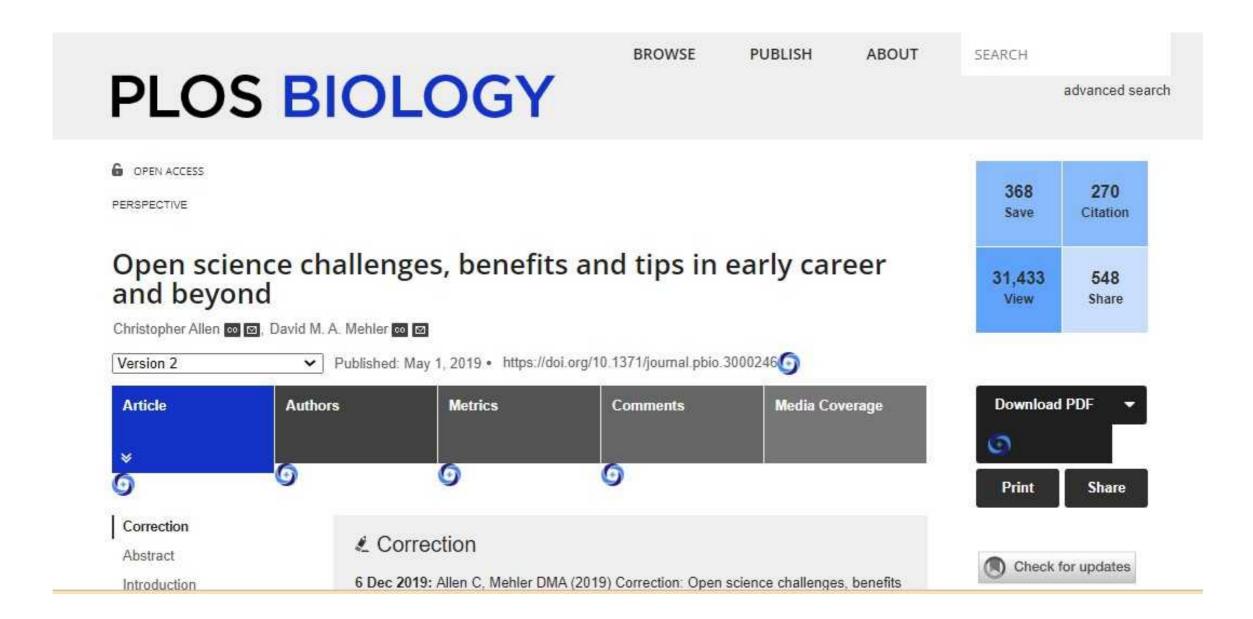






Join Our Team Become a Partner

https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000246



Passport for Open Science

(French OS policy 2020)

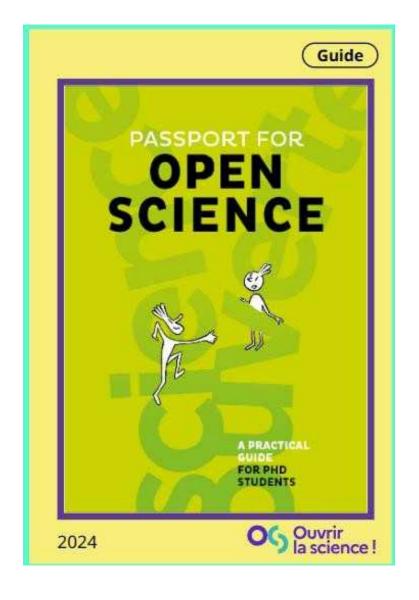


https://www.ouvrirlascience.fr/wp-content/uploads/2020/11/Passport-for-Open-Science-A-Practical-Guide-For-PhD-Students 30-10-2020 WEB.pdf



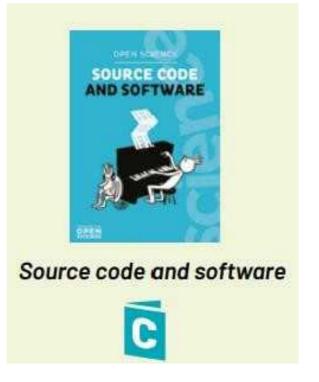
NEW EDITION IN 2024 https://www.ouvrirlascience.fr/passport-for-open-

science-a-practical-guide-for-phd-students/











NEW EDITION IN 2024

https://www.ouvrirlascience.fr/passport-for-open-science-a-practical-guide-for-phd-students/



Vidéos

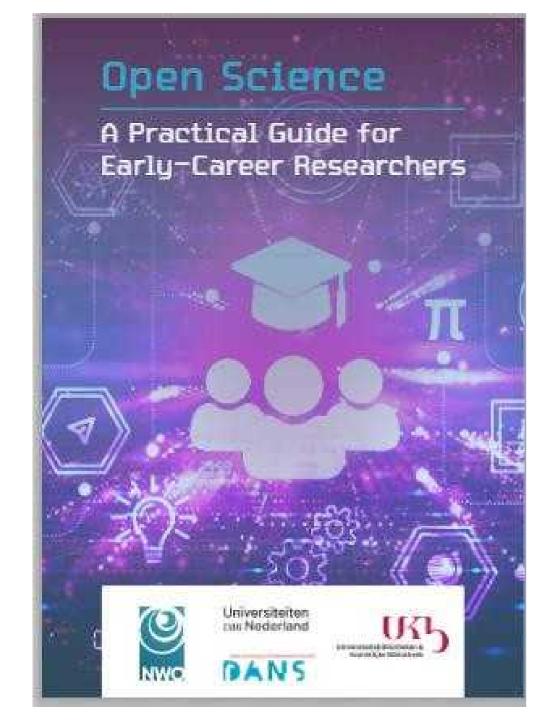








Collection of short videos for PhD students



https://doi.org/10.5281/zenodo.7716153
(Brinkman, et al. 2023)

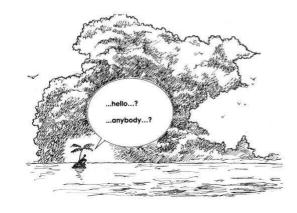
http://whyopenresearch.org

Why Open Research?

Advance your career by sharing your work.

Share your work. Be successful.

Open scholarship is good for the public and for you.









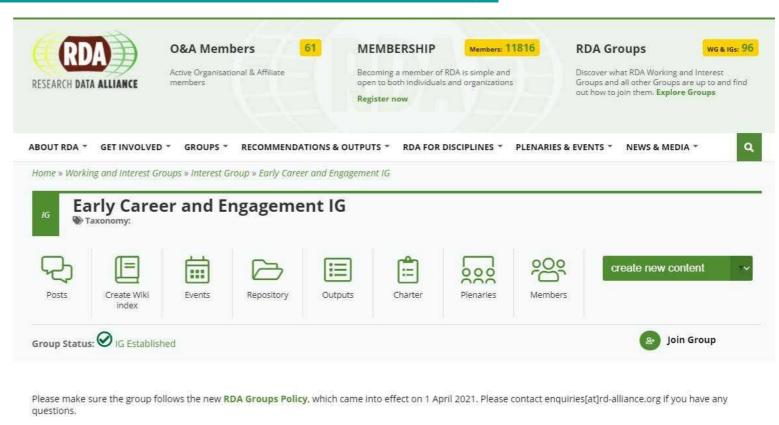
https://www.rd-alliance.org/groups/early-career-and-engagement-ig

6 Group details

0

Status: Recognised & Endorsed

Chair (s): Devan Ray Donaldson, Fotis Psomopoulos, Elli Papadopoulou, Maria Tsagiopoulou



Community 2: • COARA

https://coara.eu/working-groups/working-groups/wg-early-and-mid-career-researchers-emcrs-assessment-and-research-culture/

WG Objectives

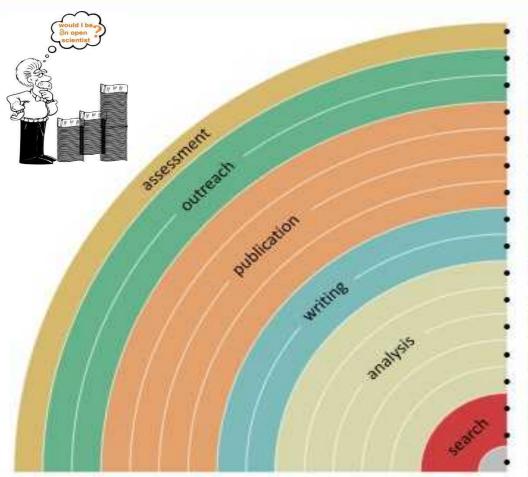
Working Group: Earlyand-mid-Career Researchers (EMCRs) -Assessment and Research Culture Objective 1: Gather information and exchange experiences, collect good and best practices – and bad practices to avoid – from a range of different countries and organisations with different levels of implementation of the reform of research assessment and diverse types of institutions as well as different institutional autonomy levels about the impacts of different assessment procedures/methods on EMCRs career paths and the cultures of research systems they are active in.

Objective 2: Develop pilot actions to:

- (i) monitor the outcomes and impacts,
- (ii) support EMCRs during the change through training and consultation,
- (iii) instruct assessors of careers towards novel research assessment practices, and
- (iv) implement an inclusive and positive culture (change)

Objective 3: Based on the gathered insights, develop guidelines, models and a toolbox for implementation, which will include monitoring templates, training recommendations, guidelines for EMCR assessment, and methodologies to drive an inclusive research culture change.

You can make your workflow more open by ...



adding alternative evaluation, e.g. with altmetrics communicating through social media, e.g. Twitter sharing posters & presentations, e.g. at FigShare using open licenses, e.g. CCO or CC-BY publishing open access, 'green' or 'gold' using open peer review, e.g. at journals or PubPeer sharing preprints, e.g. at OSF, arXiv or bioRxiv using actionable formats, e.g. with Jupyter or CoCalc open XML-drafting, e.g. at Overleaf or Authorea sharing protocols & workfl., e.g. at Protocols.io sharing notebooks, e.g. at OpenNotebookScience sharing code, e.g. at GitHub with GNU/MIT license sharing data, e.g. at Dryad, Zenodo or Dataverse pre-registering, e.g. at OSF or AsPredicted commenting openly, e.g. with Hypothes.is using shared reference libraries, e.g. with Zotero sharing (grant) proposals, e.g. at RIO





благодаря ви!



@evamen
Universidad Carlos III de Madrid
Member of the Steering Board of CoARA (2023-24)
Chair Open Science Policy Platform (2018-20)



Α







Funded by the European Union. Views and opinions expressed are however those of the auhor(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Skills



