FAIR DATA MANAGEMENT

Innovation Acta, November 8th, 2023

Elena Giglia

University of Turin

elena.giglia@unito.it

Innovation acta, Nov. 8, 2023

FAIR data management + Horizon Europe

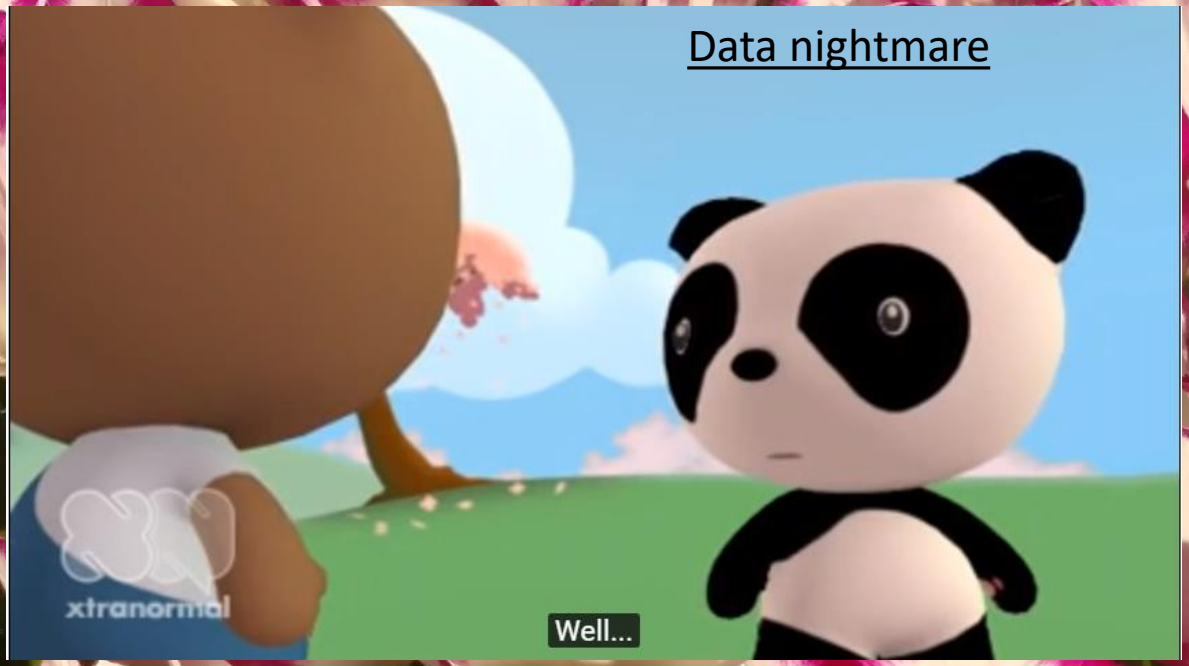
Elena Giglia

elena.giglia@unito.it

@egiglia

Why should we care about data?

Why should you take care of your data?



… THIS IS THE DATA STEWARD'S NIGHTMARE:

- NO BACKUP
- NO SOFTWARE
- - NO DATA LEGEND

… AND:

- DATA GENERATED WITH PUBLIC FUNDS
- PUBLISHED IN «SCIENCE» (DATA POLICY)
- REQUESTED FROM A DIFFERENT DISCIPLINE

Why should we care about data?

LOST VALUE IF DATA ARE MISSING:

- AT BEST: EXPENSIVE RESEARCH IS OF LITTILE OR NO VALUE
- AT WORST: RESULTS OF INVALID RESEARCH

ARE PUT INTO CLINICAL USE

Great values lost by not sharing data

Lack of reproducibility well known problem in medical research.

Investigations in the US: Up to 50% of studies not reproducible. 25% of this caused by unavailability of data.

At best: Expensive research is of little or no value.

At worst: Results of invalid research are put into clinical use.

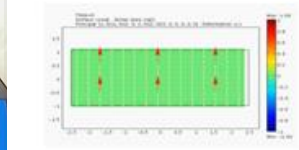
Katrine Weisteen Bjerd, EOAS Symposium 2022

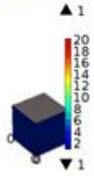
Why should we care about data? A personal view

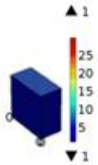
Past scientific interests

Mathematical models for soft-active materials

- Elasticity within large deformation framework (non-linear models)
- Deformation of active-smart materials (swelling materials, nematic elastomers, ...)

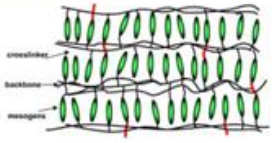


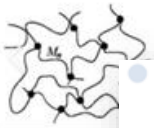




M. de Luca, A. DeSimone. Elastomeric Gels: A Model and First Results. Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems. Lecture Notes in Applied and Computational Mechanics, vol 81. Springer, Cham. (2016) https://doi.org/10.1007/978-3-319-39022-2_4

M. de Luca, A. Petelin, M. Copic and A. DeSimone, "Sub-stripe pattern formation in liquid crystal elastomers: Experimental observations and numerical simulations", JMPS, 61 (2013) 2161 – 2177 https://doi.org/10.1016/j.jmps.2013.07.002





10 YEARS ON…

- DO I HAVE ACCESS TO MY OWN PUBLICATIONS?
- WHERE ARE MY DATA?
- CAN I REPRODUCE MY SIMULATIONS?

[M.R. DE LUCA, PhD]

Research (FAIR) data management

2023

| Mariarita de Luca |
https://orcid.org/0000-0002-5507-968X
mariarita.deluca@areasciencepart.it
Institute for Research and Innovative Technologies (RIT)
AREA SCIENCE PARK

1° Workshop for National PhD in "Theoretical and Applied Neuroscience", Bertinoro 18.10.2023

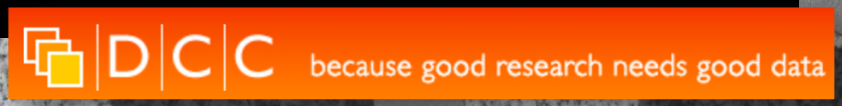
What about my data and my publications?

- Do I have access to my publications?
- Where are my data?
- Can I reproduce my numerical simulations?

Image by Elisa from Pixabay

Why should we care about data?

1. DATA ARE THE FOUNDATION OF GOOD RESEARCH



2. COVID SHOWED THAT WE NEED DATA, AND WE NED THEM AS SOON AS POSSIBLE

3. DATA ARE FRAGILE. THEY GET LOST

4. SOME DATA ARE UNIQUE AND NOR REPRODUCIBLE (ATMOSPHERIC, EARTHQUAKES…)

5. DATA CAN BE MANIPULATED, DATA MANAGEMENT PRESERVES INTEGRITY

6. TO ALLOW FOR CHECKS AND REPRODUCIBILITY

7. DATA CAN BE REUSED (IN UNEXPECTED WAYS)

Why should we care about data?

8.1 WE HAVE TO. OPEN DATA DIRECTIVE

8.3 WE AVE TO. WE HAVE EOSC

L 172/56 EN Official Journal of the European Union 26.6.2019

DIRECTIVE (EU) 2019/1024 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 20 June 2019

on open data and the re-use of public sector information

(recast)

Open data directive

EOSC Association

Advancing Open Science in Europe

DIRECTIVE ENLARGED TO INCLUDE RESEARCH DATA

8.4. WE HAVE TO. A GROWING NUMBER OF JOURNALS IS ASKING FOR DATA TO BE DEPOSITED UPON PUBLICATIONS (TRASPARENCY AND E REPRODUCIBILITY)

8.2 WE HAVE TO. IN HORIZON EUROPE YOU HAVE TO RESPONSIBLY MANAGING RESEARCH DATA ACCORDING TO FAIR PRINCIPLES (MANDATORY PRACTICE)

V.1 Feb 2021

Horizon Europe (HORIZON)
Euratom Research and Trai
(EURATOM)
General Model Grant A
EIC Accelerator Co

ANNEX 5

COMMUNICATION, DISSEMINATION, OPEN SCIENCE AND VISIBILITY (— ARTICLE 17)

Open science: research data management

The beneficiaries must manage the digital research data generated in the action ('data') responsibly, in line with the FAIR principles and by taking all of the following actions:

Why should we care about data?

Data creates a bridge between traditional disciplines, spawning discovery and innovation from the humanities to the hard sciences. Data dissolves barriers, opening up new channels of communication, lines of research, and commercial opportunities. Data will be the engine, the spark to create a better world for all.

World Economic Forum 2012

Sept. 29, 2021

9. DATA CREATES BRIDGES…

…REMIND: HORIZON EUROPE AND THE MISSIONS…

Why should we care about FAIR data?

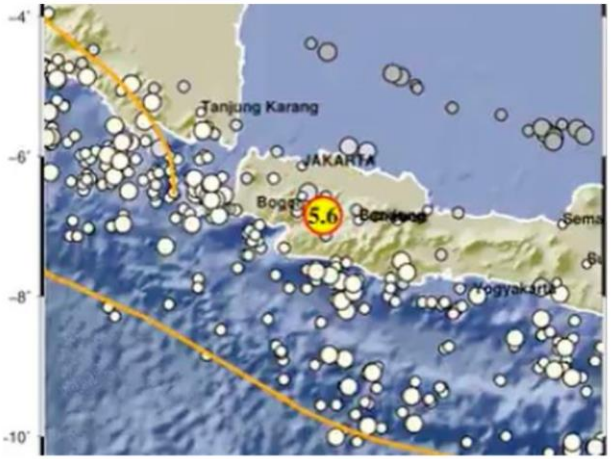
…the selfie…

How we can get those data

This was the best map that we can get (cited by the media)

Those data points are not really data points. They're just a selfie of data points.

They're not reusable.



IN «FAIR» THE STRESS IS ON «R»

BEWARE…
IF DATA ARE NOT REUSABLE THEY ARE JUST A SELFIE OF DATA
[USELESS]
[Dasapta Erwin Irawan]

FAIR are the pillar of EOSC

The Vienna Declaration on the European Open Science Cloud

Vienna, 23 November 2018

BECAUSE EOSC IS HERE TO STAY

Vienna, Nov.23, 2018

We, Ministers, delegates and other participants attending the launch event of the European Open Science Cloud (EOSC):

1. Recall the challenges of data driven research in pursuing excellent science as stated in the "EOSC Declaration" signed in Brussels on 10 July 2017.

2. Reaffirm the potential of the European Open Science Cloud to transform the research landscape in Europe. Confirm that the vision of the European Open Science Cloud is that of a research data commons, inclusive of all disciplines and Member States, sustainable in the long-term.

3. Recognise that the implementation of the European Open Science Cloud is a process, not a project, by its nature iterative and based on constant learning and mutual alignment. Highlight the need for continuous dialogue to build trust and consensus among scientists, researchers, funders, users and service providers.

4. Highlight that Europe is well placed to take a global leadership position in the development and application of cloud services for Science. Rea and open to the world, reaching out over time to

SEAMLESS ACCESS TO OPEN BY DEFAULT FAIR DATA

5. Recall that the Council roadmap and the federated

9. Call for the European Open Science Cloud to provide all researchers in Europe with seamless access to an open-by-default, efficient and cross-disciplinary environment for storing, accessing, reusing and processing research data supported by FAIR data principles.

Science Cloud a reality, hinting at the need to further strengthen the ongoing dialogue across institutions and with stake-holders, for a new governance framework to be launched in Vienna, on 23 November 2018.

What is EOSC?

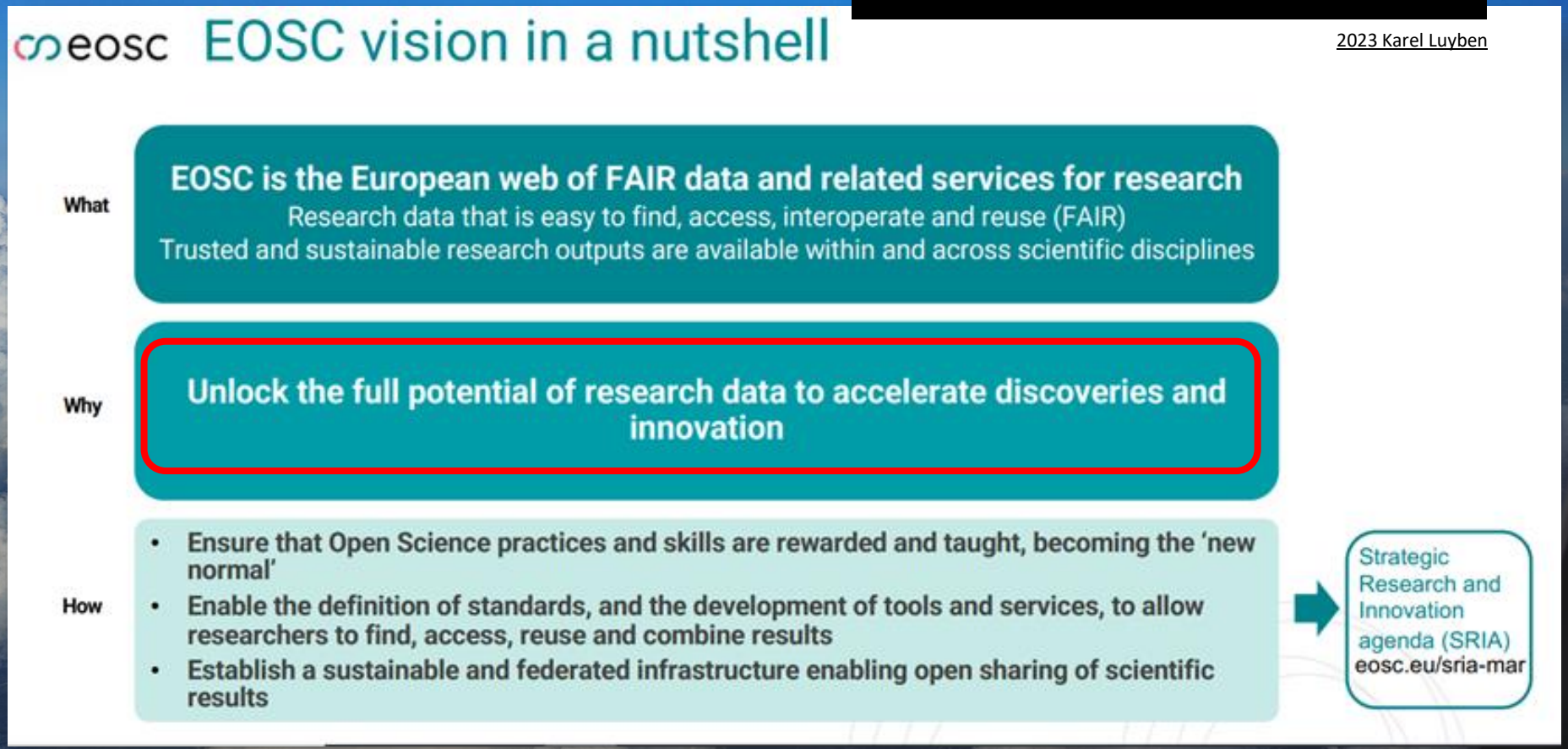
OSC

"A web of scientific insight"

- Web of FAIR Data and related Services
- Federation of relevant existing and future data sources
- Virtual space where science producers and consumers come togethe
- An open-ended range of content and services
- Based on the FAIR principles
- Meeting all European data requirements
- In interaction with other regions of the world

What is EOSC?

EU WEB OF
FAIR DATA AND SERVICES
TO UNLOCK THE FULL POTENTIAL
OF RESEARCH DATA



What EOSC is NOT

2023 Karel Luyben

EOSC is not …

1. …a cloud infrastructure
Despite the word "cloud" part of its name, EOSC is not a new cloud computing platform

2. … a new research data repository or research data management system
The federation of existing infrastructures, i.e. EOSC, is a new infrastructure which does not exist today

3. … a new pan-European e-infrastructure
EOSC is not building a new e-infrastructure. EOSC is building i. the components to enable the federation of existing data, research and e- infrastructures nodes and ii. the additional services needed to enable the Web of FAIR data and related services.

4. … synonymous of Open Science
EOSC is the enabler that will support the deployment of Open Science in Europe. EOSC does not substitute any existing Open Science networks.

5. … the EOSC Association
The EOSC Association as representative of the various stakeholders in Europe is the legal entity established to work together with the European Commission to support the realisation of the EOSC strategy.

6. … substituting any existing national, regional, pan-European, agnostic nor thematic Research Infrastructures o e-infrastructures
EOSC will enable the federation of existing data, research and e-infrastructures nodes. The new developments are focused on components enabling the federation and on the additional services needed to enable the Web of FAIR data and related services

7. … the EOSC Portal
The EOSC Portal is one of the results of the EOSC Future EC funded project (2019-2023). The EOSC Portal is piloting the EOSC AAI and the idea of a European marketplace for services supporting researchers.

8. … owning any data or services
EOSC is an enabler. The ownership of the federated elements (data, services, research infrastructures, e-infrastructures, etc.) will remain with the providers.

9. … engaging directly individual researchers.
Individual researchers will benefit from EOSC through their existing channels (e.g. universities, research institutes, research infrastructures, associations, etc.) that will act as intermediaries.

EOSC IS NOT A BIG BOX]

THE EUROPEAN OPEN SCIENCE CLOUD? SOME NUANCES AND DEFINITIONS

Imagine a federated, globally accessible environment where researchers, innovators, companies and citizens can publish, find and re-use each other's data and tools for research, innovation and educational purposes. Imagine that this all operates under well-defined and trusted conditions, supported by a sustainable and just value for money model. This is the environment that must be fostered in Europe and beyond to ensure that European research and innovation contributes in full to knowledge creation, meet global challenges and fuel economic prosperity in Europe. This we

EOSC IS NOT A REPOSITORY NOR A «CLOUD»

YOU MAKE YOUR DATA FAIR SO THAT EOSC *SERVICES* CAN «FIND» THEM…

A SUPPORTING ENVIRONMENT FOR OPEN SCIENCE AND NOT AN «OPEN CLOUD» FOR SCIENCE

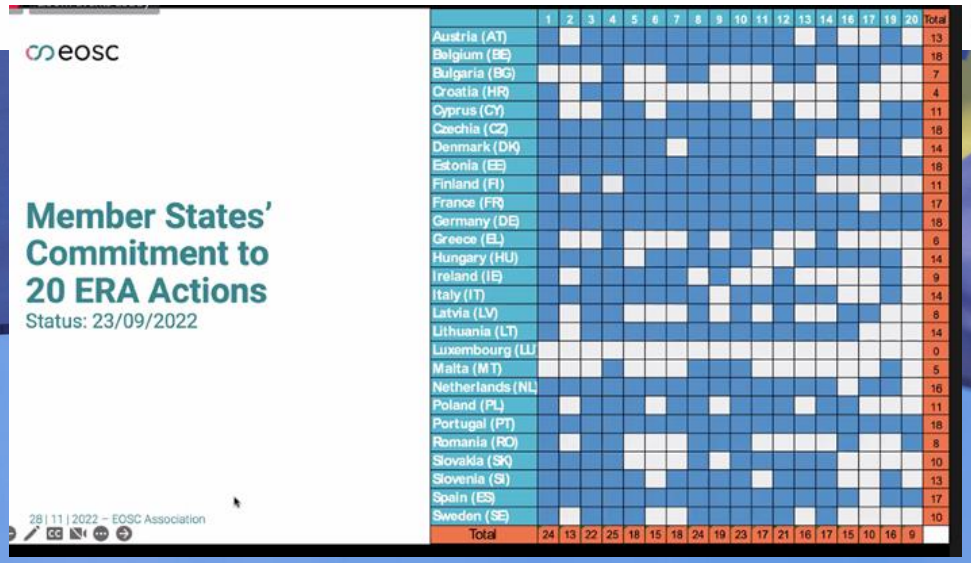
YOU DON'T «UPLOAD» YOUR DATA INTO EOSC

AND GIVE SEAMLESS ACCESS TO 20 M EU RESEARCHERS

[ACTION 1 OF THE ERA AGENDA]

- In particular ERA Action 1: "***Enable the open sharing of knowledge and the re-use of research outputs**, including through the **development of the EOSC***", targeting to:
  - Deploy Open Science principles and identify Open Science best practices
    - Mainstreaming OS across nat'l programmes, catalogue of OS practices, tools and services, data scientists and data stewards, nat'l EOSC tripartite events …
  - Deploy the core components and services of EOSC and federate existing data infrastructures in Europe, working towards the interoperability of research data
    - Horizon Europe support to EOSC Partnership, connection of nat'l/regional research infrastructures to EOSC federation, community frameworks for interoperability and quality control …
  - Establish a monitoring mechanism to collect data and benchmark investments, policies, digital research outputs, open science skills and infrastructure capacities related to EOSC
    - Co-development of EOSC national surveys, roll-out of key layers of monitoring mechanism …

European Commission



[EOSC is based on data stewardship]

Report, 2016

The number of people with these skills needed to effectively operate the EOSC is, we estimate, likely exceeding half a million within a decade. As we further argue below, we believe that the implementation of the EOSC needs to include instruments to help train, retain and recognise this expertise, in order to support the 1.7 million scientists and over 70 million people working in innovation[9]. The success of the EOSC depends upon it.

- WE NEED 500.00 DATA STEWARDS
- DATA STEWARDS ARE ONE OF THE CRITICAL SUCCESS FACTORS OF EOSC

SRIA 1.0

7.4. Critical success factors

The developments and expected impacts described above will not happen spontaneously. For these benefits to materialise a number of critical success factors (CSFs) must be in place. The following CSFs have been identified for EOSC:

- Researchers performing publicly funded research make relevant results available as openly as possible;
- Professional data stewards are available in research-performing organisations in Europe to help implement FAIR principles and support Open Science;

[competence profi

KØBENHAVNS UNIVERSITET 06/04/202 0

Education core content

This 1-year degree should build upon students' educational/job background through domain specific data knowledge and leverage with theoretical and practical competences.
The education can be viewed as a Data Steward specialisation within the domain of their previous degree/jobs. The education contains **60 ECTS** and is expected to finish with a 15 ECTS project.

Preliminary Content

The 60 ECTS should be distributed among the following main areas:

- 22,5-30 ECTS: IT competences – including computational thinking, data modelling, data management, data harvesting, cleaning, and storing, infra-structure (storage & compute). An introduction to data science, machine learning, and their derived data needs.
- 7,5-15 ECTS: Legal and ethical competences – including GDPR, FAIR, data security, and data & AI ethics.
- 7,5-15 ECTS: Domain specific data competences – including knowledge about data, infrastructure, and practice within the students primary domain, e.g., health, life-science, finance/fintech, or the public sector.
- 15 ECTS: Graduate project (possibly in collaboration with academia, industry, or the public sector)

Competences such as project management, communication skills, and change management should be

KØBENHAVNS UNIVERSITET

Competence Profile

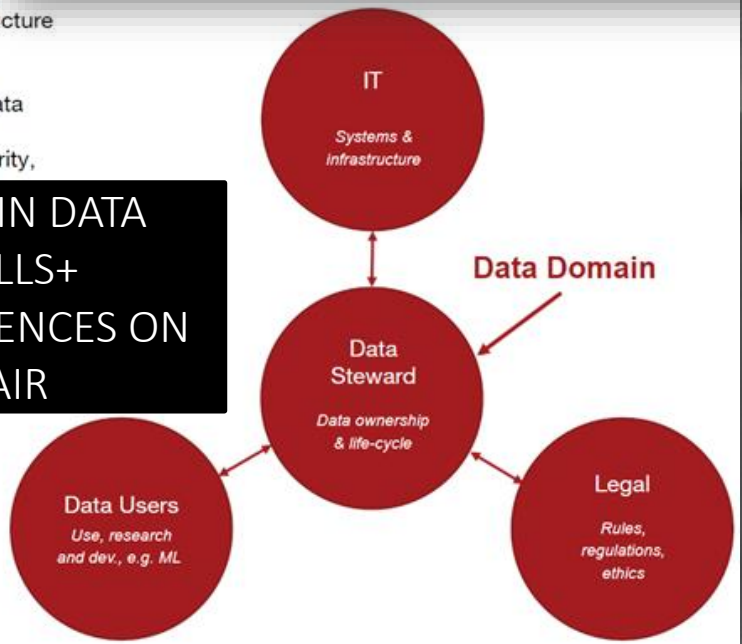
A data steward is a data specialist with strong domain-specific knowledge who understands and appreciates the relevance of data, data sources, data infrastructure and constraints within a scientific or other application domain.

The future Data Steward must assume ownership and responsibility for data, data quality, and the data life-cycle as their primary function. They should ensure collaboration and coherence between IT competences, quality assurance, security, rules & regulations, and facilitate the application and use of [illegible] externally in the organisation.

Competence profile examples

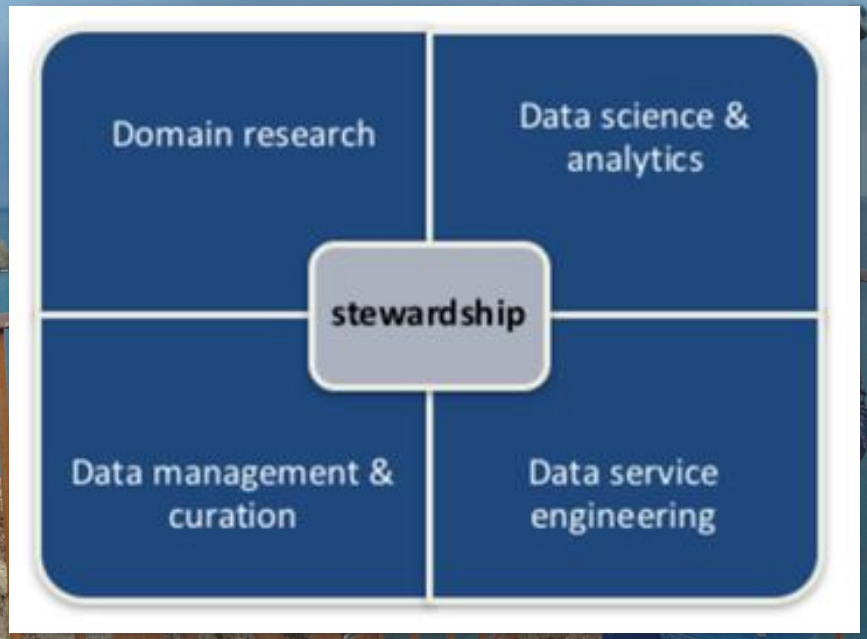
- Domain-specific data understanding
- Ability to ensure that structured and unstructured data data is modelled, harvested, stored, and maintained in documented, and regulated fashion with focus and findability, accessibility, interoperability, and reusability.
- Competences to facilitate HPC (High Performance Computing) during development and research through handling of large-scale data in public and private enterprises.
- Understanding of and competences within legal, ethical and security aspects of data handling, data sharing, e.g., integrity and GDPR.

DOMAIN DATA SKILLS+ COMPETENCES ON FAIR



Copenhagen Univ. June 17 2020

[competence profile]



2018

EOSC pilot
The European Open Science Cloud for Research Pilot Project

D7.3: Skills and Capability Framework

Author(s)	Angus Whyte, Jerry de Vries, Rahul Thorat, Eileen Kuehn, Gergely Sipos, Valentino Cavalli, Vasso Kalaitzi, Kevin Ashley

[BTW, time to rethink…]

nature

Explore content · About the journal · Publish with us · Subscribe

nature › career column › article

CAREER COLUMN | 14 April 2022

Time to re-think the divide between academic and support staff

Apr. 2022

Research professionals should not be split into two categories, say Marta Teperek, Maria Cruz and Danny Kingsley.

In recent years, we have seen 'support' jobs become more important at research organizations, including roles such as data stewards, research software engineers, scientific community managers and programme managers. We have seen how a diversity of roles and contributions drives progress and success in research and innovation.

We have come to see the sharp distinction between 'academics' and 'support staff' as a barrier to effective research because it discourages a culture of collaboration and appreciation of a diversity of roles and contributions.

drives rift between academic and non-academic staff

As professionals, we make a significant contribution alongside conventional academics. Like many of our colleagues in 'support' roles, we are well connected with the academic community. We work in partnership with researchers, contributing unique expertise and skills. We have academic credentials. We write papers, books, grant proposals, reports and manuals. We train students and academic staff; manage projects; organize and present at conferences and workshops; and lead developments in our areas of expertise. We are knowledge brokers, able to translate generic infrastructure, tools and policies into practical solutions that make research more efficient.

- DIVERSITY OF CONTRIBUTIONS IS A SUCCESS FACTOR
- CULTURE OF COLLABORATION

What is data stewardship?

Data stewardship is the responsible planning and executing of all actions on digital data before, during and after a research project, with the aim of optimising the usability, reusability and reproducibility of the resulting data.

It differs from data management, in the sense that data management concerns all actual, operational data-related activities in any phase of the data lifecycle, while data stewardship refers to the assignment of responsibilities in, and planning of, data management.

DATA STEWARDSHIP IS THE RESPONSIBLE PLANNING AND EXECUTING OF ALL ACTIONS ON DIGITAL DATA BEFORE, DURING AND AFTER A RESEARCH PROJECT, WITH THE AIM OF OPTIMISING THE USABILITY, REUSABILITY AND REPRODUCIBILITY OF THE RESULTING DTAA

What is data stewardship? / 2

experts and research roles. Three different, partly overlapping stakeholder fields (or working areas) of the data steward were characterised, which all have their own focus and thus different data steward role: policy, research and infrastructure. Together they form the data stewardship landscape. Each data steward role has eight competence areas:

- Policy/strategy
- Compliance
- Alignment with FAIR data principles
- Services
- Infrastructure
- Knowledge management
- Network
- Data archiving

The responsibilities, tasks and KSAs were defined per competence area and differ between the data steward roles. The data steward role is often experienced as a role that is 'in between' different disciplines and professionals. Translation between different stakeholders and professionals is seen as a key element of the function of a data steward.

DATA STEWARD HAS 8 COMPETENCE AREAS - ONE OF THE KEY AREAS IS ACTING AS A BRIDGE AMONG DIFFERENT PROFESSIONALS (DATA ENGINEER, LEGAL ADVISOR…)

Table 3. Overview of the eight defined competence areas for all data steward roles

2019

Competence area	This concerns
Policy/strategy	Development, implementation and monitoring of research data management policy and strategy for the research institute
Compliance	Compliance to the Netherlands Code of Conduct for Academic Practice, the Netherlands Code of Conduct for Research Integrity, the General Data Protection Regulation (GDPR), and other relevant legal and ethical standards
Alignment with FAIR data principles	Alignment to the FAIR data principles and the principles of Open Science
Services	Availability of adequate support on research data management, in staff or services
Infrastructure	Availability of adequate data infrastructure for research data management
Knowledge management	Adequate level of knowledge and skills on research data management within the institute, department or project
Network	Obtaining and maintaining a network of aligned expertise areas and relevant departments and organisations inside and outside the institute, department or project
Data archiving	Adequate support and data infrastructure for FAIR and long-term archiving of data of the institute, department or project

Data stewards - profile

DIFFERENT APPROACHES

2019

[Curricula]

Data Steward is a data handling and management professional whose responsibilities include planning, implementing and managing (research) data input, storage, search, and presentation. Data Steward creates a data model for domain specific data, support and advice domain scientists/ researchers during the whole research cycle and data management lifecycle.

DATA STEWARD IS A DATA HANDLING AND MANAGEMENT PROFESSIONAL WHOSE RESPONISIBLITIES INCLUDE PLANNING, IMPLEMENTING AND MANAGING DATA INPUT, STORAGE, SEARCH, AND PRESENTATION. DATA STEWARDS CREATE DATA MODEL AND ADVICE IN EACH STEP OF THE CYCLE

Research Data Management and Data Stewardship Competences in University Curriculum

Yuri Demchenko
University of Amsterdam, The Netherlands
y.demchenko@uva.nl

Lennart Stoy
EUA, Belgium
lennart.stoy@eua.eu

May 2021

***Abstract*— Skills for data governance and management are critical for wide adoption of the Open Science practices and effective use of the data in research, industry, business and other economy sectors. The FAIR (Findable – Accessible – Interoperable - Reusable) data management principles and data stewardship provide a foundation of effective research data management. The 2018 "Turning FAIR into Reality" report and other documents recommend that data skills should be more widely included in university curricula and that a concerted effort should be made to coordinate and accelerate the pedagogy for professional data roles. Throughout Europe, and beyond, many organisations, projects and initiatives work on providing training on FAIR data competences. However wider adoption of the FAIR data culture can be achieved by including FAIR competence into university curricula. This paper presents the ongoing work of the FAIRsFAIR project to develop Data Stewardship competence framework and provide recommendations for implementing in the university curricula by defining the Data Stewardship Body of Knowledge Model Curricula. The proposed approach and identified competences and knowledge items are supported by the job market analysis. The presented work is actively using the EDISON Data Science Framework as a basis for Data Stewardship competences definition and methodology for linking competences, skills, knowledge, and intended learning outcomes when designing curricula.**

D7.5 Good Practices in FAIR Competence Education

2022

FAIRSFAIR
Fostering Fair Data Practices in Europe

Data steward

2018

Data Stewardship: Addressing Disciplinary Data Management Needs

Marta Teperek
Research Data Services
TU Delft Library

Maria J. Cruz
4TU.Centre for Research Data
TU Delft Library

Data Stewards: Disciplinary Experts Who Look After Research Data

Data stewards are disciplinary experts with knowledge of data management who are employed at faculties in order to advise researchers and faculty members on the various aspects of research data management. Specifically, the data stewards are tasked with the following:

In addition, we believed that disciplinary expertise, reflected in a PhD degree (or equivalent experience) in the area of faculty's research, was necessary for the stewards to provide relevant and tailored advice to their communities.

Is One Data Steward Per Faculty Enough?

Finally, in order to be truly discipline-specific, one data steward per faculty might not be enough. There is substantial diversity in the research topics and disciplines within the faculties themselves. For example, research groups at the Faculty of Applied Sciences

Data ste

Data Stewardship: Addressing Disciplinary Data Management Needs 2018

Marta Teperek
Research Data Services
TU Delft Library

Maria J. Cruz
4TU.Centre for Research Data
TU Delft Library

- **Analyse data management needs** – through undertaking a mixture of semi-structured qualitative interviews and quantitative surveys;
- **Provide advice and consultancy** – meet with researchers, discuss their data management practices, make suggestions for possible improvements and become the trusted person for any questions about data management;
- **Liaise with key faculty stakeholders** – ensure that the various faculty service providers (such as contracts managers or faculty information coordinators) are aware of good data stewardship and that requirements of good data stewardship are aligned with their workflows (for example, budgeting for data management in grant applications);
- **Train and inspire** – advocate for good data management, deliver information sessions, analyse training needs, develop and deliver workshops to ensure that researchers have the skills necessary for responsible data stewardship;
- **Help comply with funders' and journals' policies** – assist researchers with drafting their data management plans, preparing their research data for deposit and advise them on changes to data policies;
- **Develop faculty research data policies** – organise and facilitate policy consultations across the faculty, help faculty define roles and responsibilities of the different faculty-level stakeholders, and drive policy implementation, evaluation and revision;
- **Prepare the faculty for the future** – keep the faculty up to date with new developments and policy changes related to data stewardship, and keep abreast of new developments in the faculty's research area to ensure that researchers have the right skills to manage their data, despite of evolving research methodologies;
- **Liaise with the Data Stewardship Coordinator and other stewards** – liaise with other members of the Data Stewardship programme to exchange practice and to discuss relevant issues;
- **Deliver regular reports** – regularly evaluate, monitor and report on data management practices within the faculty.

You need skilled people

Skills 4 eosc

ABOUT KERS NEWS

https://www.skills4eosc.eu/

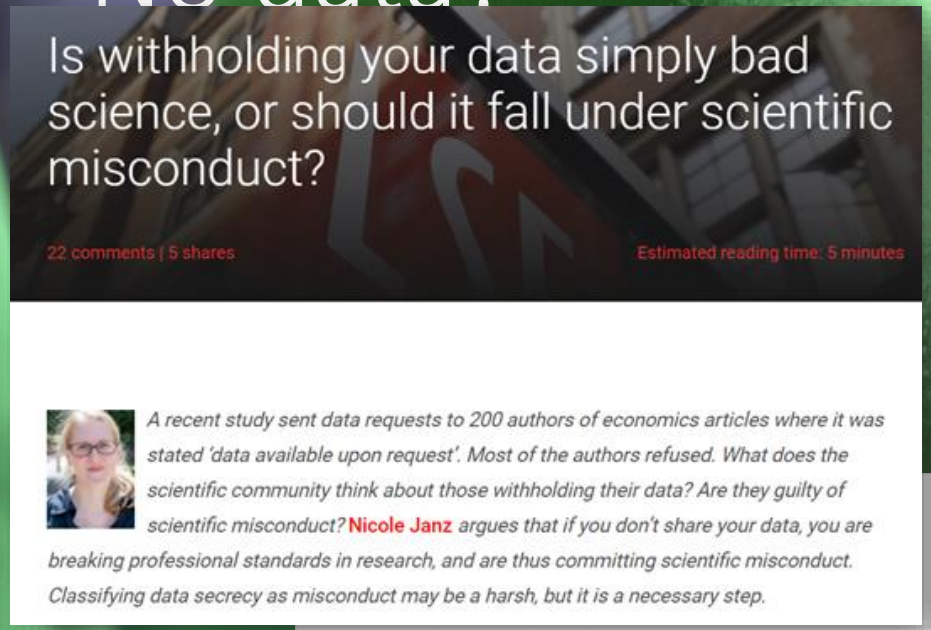
- Skills for the European
- Open Science
- Commons

SKILLS4EOSC PROJECT
(UNITO PARTNER)
CURRICULA FOR DATA STEWARDS AND COMPETENCE CENTER COORDINATON

Objectives of the project are:

1. **Map career profiles related to Open Science** and define, through co-creation the "**Minimum Viable Skillset**" (MVS) for each of them; create a shared framework for the recognition of competencies acquired by university students, trainers and new professionals as a part of an academic path or a lifelong learning process.
2. **Define a methodology and a Quality Assurance process** to ensure the quality and relevance of OS learning materials and the management of their life-cycle, thus enhancing their re-usability.
3. **Offer training on OS and the usage of data in evidence-based policy for civil servants** and policymakers and empower CCs, researchers and "honest brokers" through the offering of resources to carry out training for this target.
4. **Define "OS and data-intensive science essentials"** for inclusion in generic undergraduate, postgraduate and PhD curricula as a key skill that anyone doing research is expected to acquire.
5. Design and implement **a collaboration model between national and regional CCs and international Research Infrastructures** and communities to provide specialised OS competencies targeting the needs of researchers and thematic RI professionals.
6. **Support lifelong learning** through professional networks as an enabling environment to discuss, cocreate and exchange best practices and solutions among OS professionals and researchers.
7. **Coordinate national, regional and thematic Competence Centres on OS and EOSC** in Europe and leverage their expertise to create a widespread user support network and an environment that fosters and harmonises training and skills activities.
8. Create and implement a strategy for engaging with **relevant stakeholders to co-create and promote the project outputs** (Curricula, shared certification and QA frameworks, human networks), building partnerships to embed project activities and results among the broadest network of stakeholders.
9. **Establish synergies with key actors within the Member States and in the EOSC arena**, and with human capital and training programmes at the national, regional and European levels to maximise the impact of the project activities and results and pave the way for their long-term sustainability.

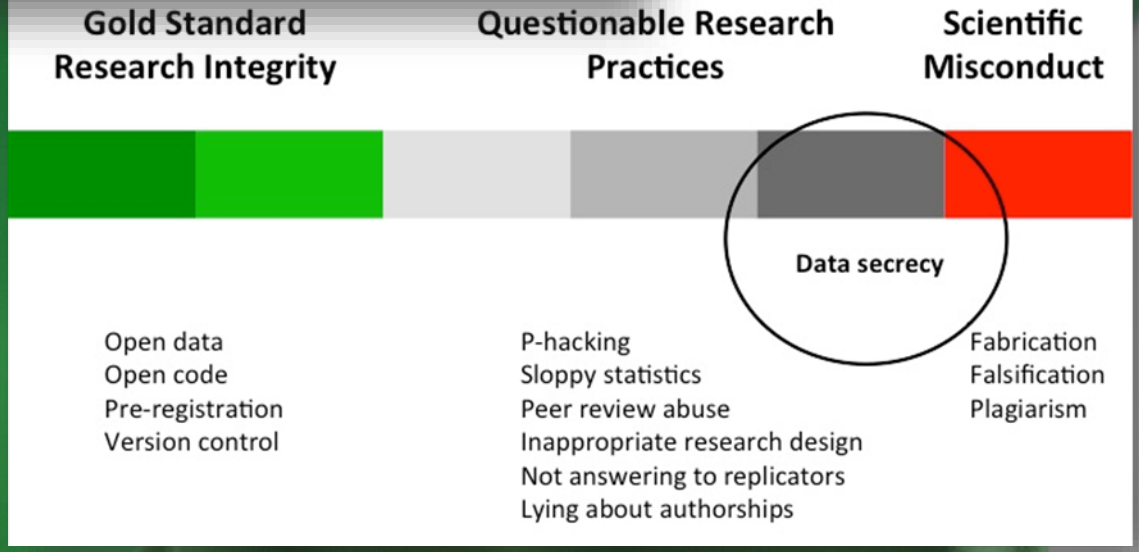
No data?



https://twitter.com/alastairdunning/status/968453078218395648

NO DATA?
LAZINESS OR FRAUD?

2015



Wilma van Wezenbeek
@wvanwezenbeek

Following

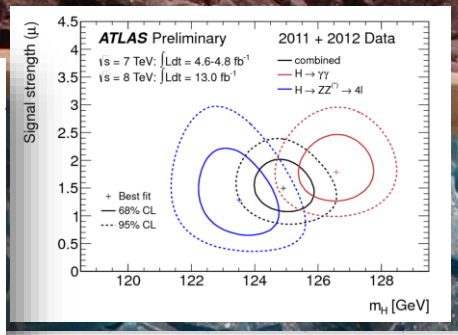
#osc2018 Wolfram Horstmann wants us to talk about datadiversity, like we do with biodiversity #openscience

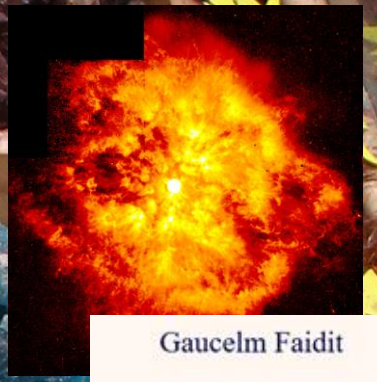
Traduci il Tweet

12:51 - 13 mar 2018

3 Retweet **1** Mi piace

https://twitter.com/wvanwezenbeek/status/973527086685093893





karyotic operational taxonomic unit (OTU) and sample from the cohort.

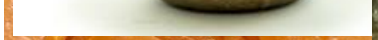
	A01_TP3	A03_TP1	A03_TP3	A04_TP1	A04_TP2	A04_TP3	A05_T
	1206	523	2131	25707	64473	60665	
	117	43035	206	119	1152	539	
					22858	1898	
					1457	29	
					19	85	
					2646	214	
					292	37	
					18	170	
					6	4	

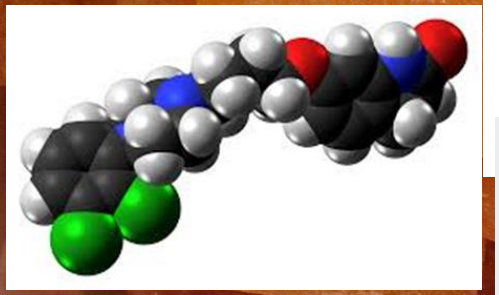
OTU_5	601	16
OTU_16	65	17
OTU_33	0	0

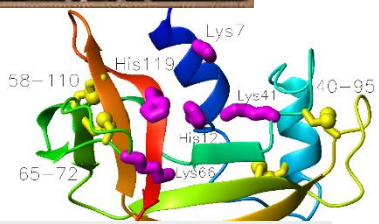
Gaucelm Faidit

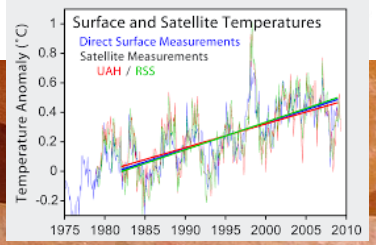
I.

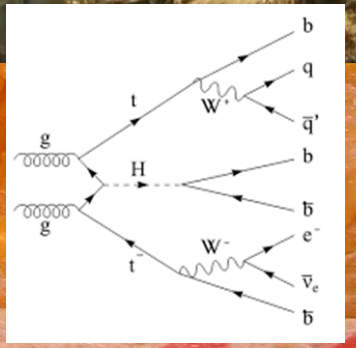
Ara nos sia guitz
lo vers dieus Iesu Cristz,
car de franca gen gaia
soi per Lui partitz,
on ai estat noiritz
et onratz e grazitz;
per so·l prec no·ill desplaia
s'ieu m'en vauc marritz.
A! gentils lemozis,
el vostr'onrat pais
lais de bella paria
seignors e vezis
e domnas ab pretz fis,
pros, de gran cortesia,
don planc e languis
e sospir nueg e dia.











Data

ALLEA 2020

> *We could then define data in the humanities broadly as all materials and assets scholars collect, generate and use during all stages of the research cycle. In this report we focus on digital assets.*

2022

PLOS ONE

Browse Topics

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Seeing oneself as a data reuser: How subjectification activates the drivers of data reuse in science

Marcel LaFlamme, Marion Poetz, Daniel Spichtinger

Published: August 18, 2022 • https://doi.org/10.1371/journal.pone.0272153

DATA=ALL MATERIALS AND ASSETS COLLECTED, GENERATED AND USED DURING THE RESEARCH CYCLE

THINK OF ALL YOUR RESEARCH ASSETS AS RESEARCH DATA THAT COULD POTENTIALLY BE REUSED

RECOMMENDATIONS

» Think of all your research assets as research data that could be potentially reused by other scholars. Consider how useful it would be for your own work if others shared their data.

Data basics

[DMP]

5 WAYS TO THINK OF DATA :

- THE WAY DATA ARE COLLECTED
- THEIR FORM
- THEIR FORMAT
- THEIR SIZE/VOLUME
- THE WORKFLOW PHASE THEY ARE IN

THEY MIGHT REQUIRE DIFFERENT TOOLS

- **The way the data is collected**.
  - By experimenting, simulations, observations, derived data, reference data.
- **The data forms**.
  - For example text documents, spreadsheets, lab journals, logs, questionnaires, software code, transcripts, code books, audio and video recordings, photos, samples, slides, artefacts, models, scripts, databases, metadata, etc.
- **The formats for electronic storage of the research data**.
- **The size (volume) of the data files**.
- **The *research lifecycle* phase the data is in.**

Univ. Southampton 2016

Data are not static: the lifecycle

PLANNIG DATA MANAGEMENT IN EVERY STEP OF THE CYCLE IS CRUCIAL

2023


…a step behind…

[the foundation]

Information Guide: Introduction to Ownership of Rights in Research Data. CREATe, University of Glasgow, 2018

Burrow, S., Margoni, T. and McCutcheon, V. (2018) Information Guide: Introduction to Ownership of Rights in Research Data. CREATe, University of Glasgow, 2018. Documentation. University of Glasgow.

http://eprints.gla.ac.uk/171314/

OpenAIRE

Guides for Researchers

How do I know if my research data is protected?

Learn more about what is research data and their protection by intellectual property rights

OpenAIRE

RESEARCH DATA ARE NOT «MINE»
NO COPYRIGHT AS THERE IS NO CREATIVITY ON DATA PER SE

Wainer Lusoli
@w_lusoli

Following

repeat with me: #researchdata is NOT mine. I was paid to get it, I'll get a #nobel 4 it, but it's NOT mine linkedin.com/pulse/repeat-m … #opendata

Traduci dalla lingua originale: inglese

Repeat with me: research data is not mine

Seldom do I see something that truly shakes me at work. You know, work is work, I am no neurosurgeon, no médecin sans frontières nor am I a social

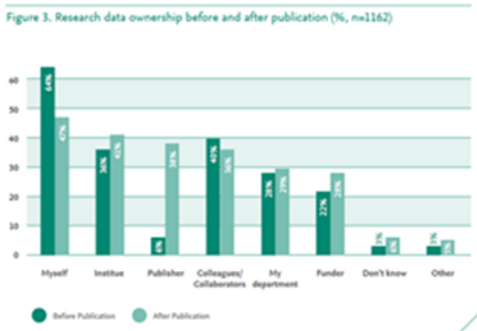
linkedin.com

11:18 - 12 apr 2017

14 Retweet **18** Mi piace

Lusoli, Apr.2017

This time though it happened. What it was: **64% of researchers believe they own the data they generated for their research.**



The result comes from a solid piece of academic research based on equally solid (open) data. The study and the report 'Open Data - the Researcher Perspective' were done by CWTS / Leiden and Elsevier. Credit giving, check.

Of course, the study reports

[DMP]

[webinar]

2020

- NO COPYRIGHT

BUT THERE MIGHT BE OTHER LEGAL PROTECTION

- UNDER GDPR, IF YOU DEAL WITH SENSITIVE DATA YOU ALWAYS MUST STATE THE LEGAL GROUND OF YOUR RESEARCH

FAIR principles

FAIR traning

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identif

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for kr

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

TO BE RE-USABLE:

R1. meta(data) have 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 their provenance.

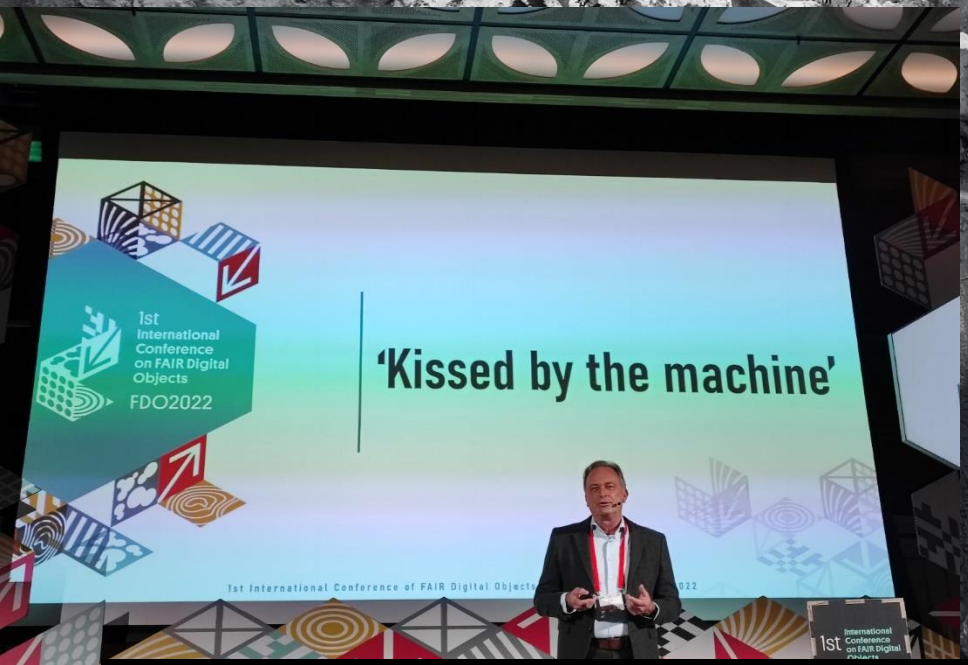
R1.3. (meta)data meet domain-relevant community standards.

Force 11

«ACCESSIBLE» DOES NOT MEAN «OPEN».

DATA CAN BE CLOSED, PROVIDED YOU – AND MACHINES - KNOW WHERE TO FIND THEM AND UNDER WHICH ACCESS CONDITIONS

Kissed or missed?



FAIR PRINCIPLES ARE
«MACHINE ACTIONABLE»
(MORE THAN READABLE)
FAIR = FULLY AI READY
IF NOT… YOU'LL BE MISSED (INSTEAD OF KISSED) BY THE MACHINE

RDA

Decision making procedures in data management and data stewardship for Open Science

Connie Clare, PhD

RDA

Data-centric AI

Automated decision making using data.

Data is fundamental for training and deploying AI models.

Data management and/or curation is a crucial step to feed into AI model.

'Machine learning models are only as good as the data they're trained on' - https://fairmlbook.org/datasets.html (Chapter 8)

Clearbox AI

Clearbox

We are on a mission to harness powerful AI technologies t improve businesses and society in a trustworthy and human-centered way.

s flexible product / Rea

clearbox AI

Your

Synthetic Data

provider

RDA

Data stewardship challenges & AI ethics

Black box AI - Model inputs and operations remain a mystery. Unknown input data provenance and quality. Automated data retrieval lead to inconsistent results.

AI bias due to generalisation (insufficient representative input data), or unsuitable data collection, processing (cleaning), quality, mislabelling and model design. Synthetic (output) data generated inherits and propagates bias affecting scientific validity.

Data misuse - Using data as input for an AI model that causes harm.

Lack of standards, tools and mechanisms to evaluate data quality ar whether datasets are fit for purpose.

ARTIFICIAL INTELLIGENCE

- WORKS IF DATA ARE GOOD
- THERE ARE ETHICAL ISSUES

FAIR research software

16 Mar 2022

FAIR Principles for Research Software (FAIR4RS Principles)

By Neil Chue Hong

2022

FAIR for Research Software (FAIR4RS) WG

Group co-chairs: Michelle Barker, Paula Andrea Martinez, Leyla Garcia, Daniel S. Katz, Neil Chue Hong, Jennifer Harrow, Fotis Psomopoulos, Carlos Martinez-Ortiz, Morane Gruenpeter

The FAIR4RS Principles are:

F: Software, and its associated metadata, is easy for both humans and machines to find.
F1. Software is assigned a globally unique and persistent identifier. • F1.1. Components of the software representing levels of granularity are assigned distinct identifiers. • F1.2. Different versions of the software are assigned distinct identifiers. F2. Software is described with rich metadata. F3. Metadata clearly and explicitly include the identifier of the software they describe. F4. Metadata are FAIR, searchable and indexable.
A: Software, and its metadata, is retrievable via standardized protocols.
A1. Software is retrievable by its identifier using a standardized communications protocol. • A1.1. The protocol is open, free, and universally implementable. • A1.2. The protocol allows for an authentication and authorization procedure, where necessary. A2. Metadata are accessible, even when the software is no longer available.
I: Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.
I1. Software reads, writes and exchanges data in a way that meets domain-relevant community standards. I2. Software includes qualified references to other objects.
R: Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software).
R1. Software is described with a plurality of accurate and relevant attributes. • R1.1. Software is given a clear and accessible license. • R1.2. Software is associated with detailed provenance. R2. Software includes qualified references to other software. R3. Software meets domain-relevant community standards.

Table 1: The FAIR Principles for Research Software

FAIR RESEARCH SOFTWARE

[the 3 steps]

1. DATA SHOULD BE AS OPEN AS POSSIBLE

2. BUT IF DATA ARE NOT «FAIR», OPENING IS RISKY (MISUSE, MISINTERPRETATION, …)

3. IF DATA ARE NOT PROPERLY MANAGED FROM THE BEGINNING, IT'S ALMOST IMPOSSIBLE TO MAKE THEM «FAIR» [WITH EOSC MANAGED/FAIR INCREASINGLY OVERLAPPING, «FAIR BY DESIGN»]

AND MANAGING DATA PROPERLY IS IN THE PRIMARY INTEREST OF ANY RESEARCHER, AS THE WHOLE RESEARCH PROCESS RESULTS STREAMLINED AND MORE EFFECTIVE

1) Data management

DESCRIPTION FOR DISCOVERABILITY (metadata)

ORGANIZATION (file naming, folders, versioning…)

BACKUP AND STORAGE

LONG TIME PRESERVATION

LEGAL ASPECTS

ALONG THE ENTIRE LIFE CYCLE

2) Make them FAIR

FINDABLE

ACCESSIBLE [≠OPEN]

INTEROPERABLE

REUSABLE

Metadata standards catalog

Metadata Standards Catalog

Search Sign in

tandards Catalog is a collaborative, open directory of metadata standards
:h data. It is offered to the international academic community to help address
nges.

DataCite
FIND, ACCESS, AND REUSE DATA

ORCID
Connecting Resea
and Researchers

ROR

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

https://www.re3data.org/

What are data journals?

Data journals are scholarly journals that publish datasets or data papers. According to Geoscience Data Journal, "a **data paper describes a dataset**, giving details of its collection, processing, software, file formats etc, without the requirement of novel analyses or ground breaking conclusions. It allows the reader to understand the when, how and why data was

Data journals

If your data are stored in other formats than those mentioned below, please contact DANS.

DANS formats

Type	Preferred format(s)	Non-preferred format(s)
Text documents	• PDF/A (.pdf) • ODT (.odt)	• Microsoft Word (.doc) • Office Open XML • Rich Text File (• PDF other than

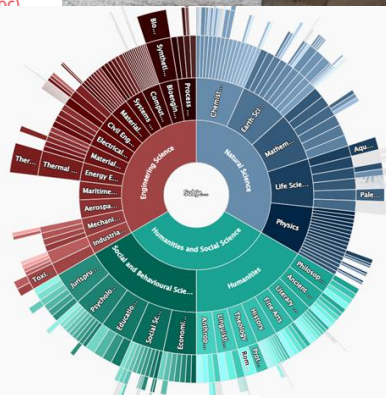
FAIRsharing.org
standards, databases, policies

STANDARDS DATABASES POLICIES COLLECTIONS ADD CONTENT STATS

https://fairsharing.org/

A curated, informative and educational resource on data and metadata standard
inter-related to databases and data policies.

We guide consumers to discover, select and use these resources with confidence, and producers to make thei
resource more discoverable, more widely adopted and cited



OpenAIRE

SERVICES SUPPO

Guides for Researchers

How do I know if my research data is protected?

Learn more a
intellectual pr

MIT Press Direct 2020

Data Intelligence

Volume 2, Issue 1-2
Winter-Spring 2020

January 01 2020

Licensing FAIR Data for Reuse

Ignasi Labastida, Thomas Margoni

CC Factsheet

creative commons

FACT SHEET ON
CREATIVE COMMONS & OPEN SCIENCE v.0.1

This information guide contains questions and responses to common concerns surrounding open science and the implications of licensing data under Creative Commons licences. It is intended to aid researchers, teachers, librarians, administrators and many others using and encountering Creative Commons licences in their work.

Project-level documentation

The project-level documentat
researcl
instrum
that you

Data-level documentation

Data-level or object-level
umentation provides information
he level of individual objects such

cessda

Data Management Expert Guide

3) Whenever possible, Open

YOU SAVE LIVES.

Digital Science Report

The State of Open Data 2021

The longest-running longitudinal survey and analysis on open data

Foreword by Natasha Simons, Australian Research Data Commons (ARDC)

November 2021

Nov. 29, 2021

Open data saves lives. The global pandemic has highlighted beyond anything that came before it the importance of data sharing in solving the big challenges of our time. COVID-19 data may be the most visualized data in history and it was made publicly available on a daily basis to people all over the world. The urgent need to better understand and treat the virus in 2020 brought unprecedented collective and collaborative action from all research stakeholders on an international scale to bring down barriers to research and speed up analysis and testing. These efforts, combined with support from governments and industry, resulted in not one but many vaccines made available by the end of the year. This gives us a glimpse of what incredible research outcomes are possible when we start with collaboration to address a common threat. Imagine how much more we could do, how many more lives we could save, if research data was routinely made open and shared. So, why isn't data sharing the norm? The answers lie in the harmony needed between policies, infrastructure, and practices.

Sharing Data

Why share data

2. Why share data?

PORT postgraduate online research training | SCHOOL OF ADVANCED STUDY UNIVERSITY OF LONDON

Better research

- Demonstrates research integrity, as there is transparency and accountability in the production of the d
- Encourages research enquiry and debat
- Promotes innovation and potential new
- Encourages the improvement of researc
- Prevents research fraud

BETTER RESEARCH
- INTEGRITY
- DEBATE
- REUSE

Better impact

- Enables peer scrutiny of the research findings, validating the work carried out
- Increases the visibility of the research
- Provides credit for the creation of the da
- Can lead to new collaborations
- Produces a public record of the research

BETTER IMPACT
- VISIBILITY
- CREDIT
- COLLABORATIONS

Better value

- Avoids duplication of effort in data creation
- Provides resources for use in teaching and learning
- Meets funder requirements
- Ensures data can be re-visited for future
- Maximises return on research investmen
- Preparing data for sharing also prepares

BETTER VALUE
- AVOID DUPLICATIONS
- MAX RETURN ON INVESTMENTS

FAIR/Open

"Open data is like a renewable energy source: it can be reused without diminishing its original value, and reuse creates new value."

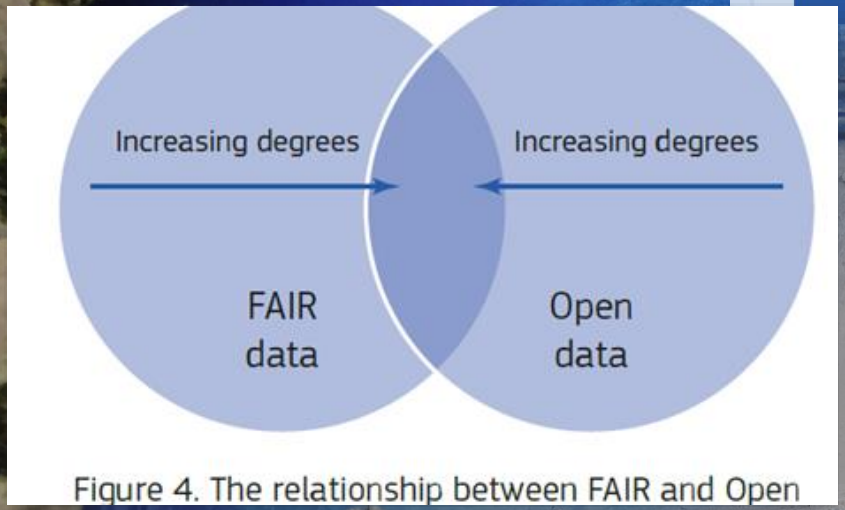
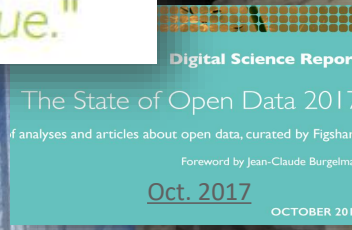


Figure 4. The relationship between FAIR and Open



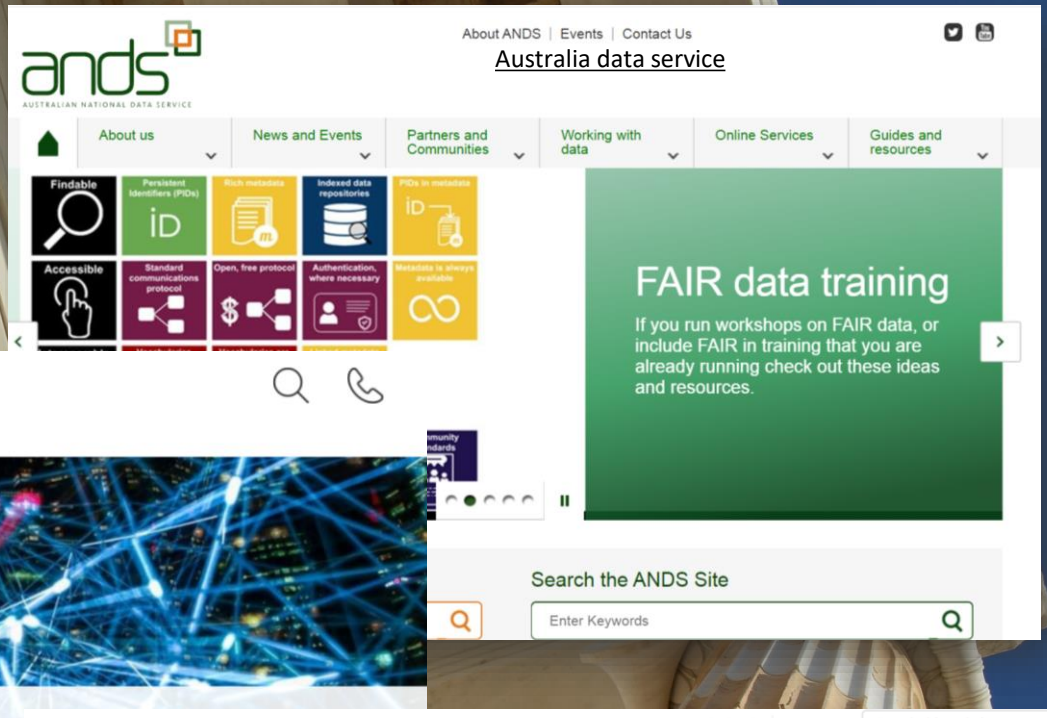
THERE WILL BE AN INCREASING DEGREE IN OVERLAPPING.
BUT WE'LL ALWAYS HAVE PERFECTLY FAIR CLOSED DATA

[BEWARE]

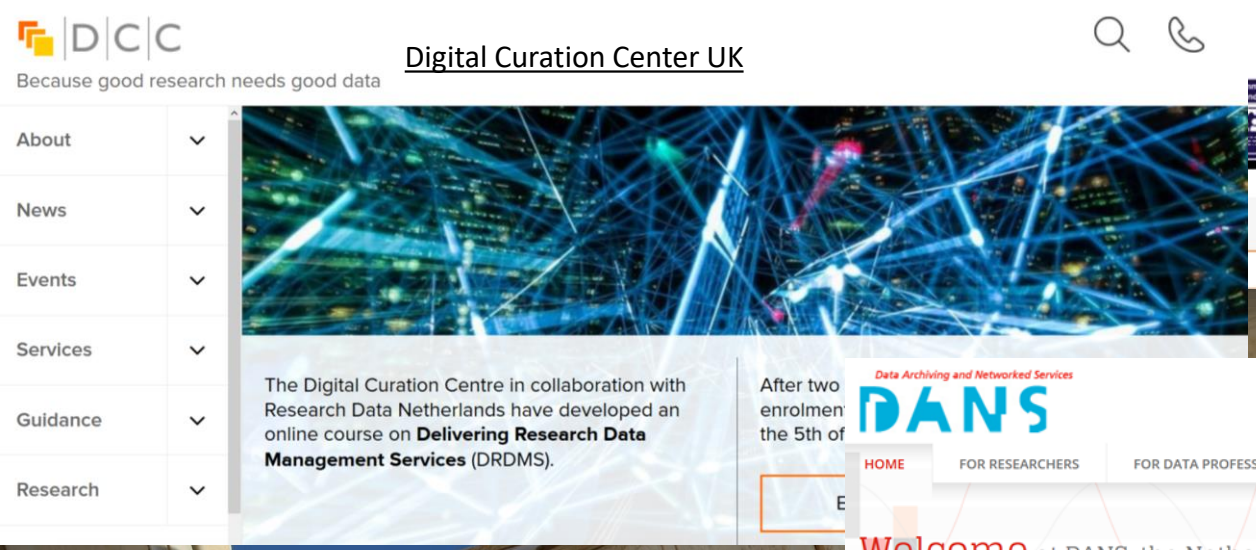
- FROM NOW ON, WE'LL SEE A LOT OF TOOLS
- **YOU NEED TO «TASTE» THEM**
- AND SEE IF THEY ARE SUITABLE FOR YOUR RESEARCH OR NOT
- BUT YOU NEED TO PRACTICE.

[DMP]

4 pillars



Australia data service



Digital Curation Center UK

Dutch data service

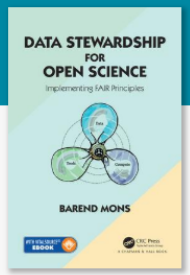
Dutch consortium

… and a master

Taylor & Francis Group
an informa business

Search for keywords, authors, titles, ISBN

2019



Data Stewardship for Open Science
Implementing FAIR Principles

the worst way imaginable to communicate the outcome of the scientific process. If science has become indeed data driven and *data is the oil of the 21st century*, we better put data centre stage and publish data as first-class research objects, obviously with supplementary narrative where needed, steward them throughout their life cycle, and make them available in easily reusable format.

Yet another recent study claimed that only about 12% of NIH funded data finds its way to a trusted and findable repository. Philip Bourne, when associate director for data science at the U.S.A. National Institutes of Health coined the term *dark data* for the 88% that is lost in amateur repositories or on laptops. When we combine the results of the general reproducibility related papers and the findability studies,

GET ACCESS PREVIEW PDF

In conclusion to this paragraph, my statement in 2005: Text-mining? Why bury it first and then mine it again? [Mons, 2005] is still frighteningly relevant.

A good data steward publishes data with a supplementary article(Data(+)).

FROM ARTICLE+
TO DATA +
[BOOK CHAPTERS WILL OPEN
IN THE DATA WIZARD]

The 5

nature Feb. 25, 2020 Subscribe

WORLD VIEW · 25 FEBRUARY 2020

Invest 5% of research funds in ensuring data are reusable

It is irresponsible to support research but not data stewardship, says Barend Mons.

Barend Mons

I tell research institutions that, on average, 5% of overall research costs should go towards data stewardship. With €300 billion (US$325 billion) of public money spent on research in the European Union, we should expect to spend €15 billion on data stewardship. Scientists, especially more experienced ones, are often upset when I say this. They see it as 5% less funding for research.

Bunk. First, taking care of data is an ethical duty, and should be part of good research practice. Second, if data are treated properly, researchers will have significantly more time to do research. Consider the losses incurred under the current system. Students in PhD programmes spend up to 80% of their time on 'data munging', fixing formatting and minor mistakes to make data suitable for analysis – wasting time and talent. With 400 such students, that would amount to a monetary waste equivalent to the salaries of 200 full-time employees, at minimum. So, hiring 20 professional data stewards to cut time lost to data wrangling would boost effective research capacity. Many top universities are starting to see that the costs of not sharing data are significant and greater than the associated risks. Data stewardship offers excellent returns on investment.

- TAKING CARE OF DATA IS ETHICAL
- HIRING DATA STEWARDS OFFERS HUGE RETURNS ON INVESTMENT
- FAIR=FULLY ARTIFICIAL INTELLIGENCE READY

Funders hold the stick: they should disburse no further funding without a properly reviewed and budgeted data-stewardship plan. The carrot is that FAIR data allow much more effective artificial intelligence (FAIR can also mean 'fully AI ready'), which will open up unprecedented research opportunities and increase reproducibility.

C

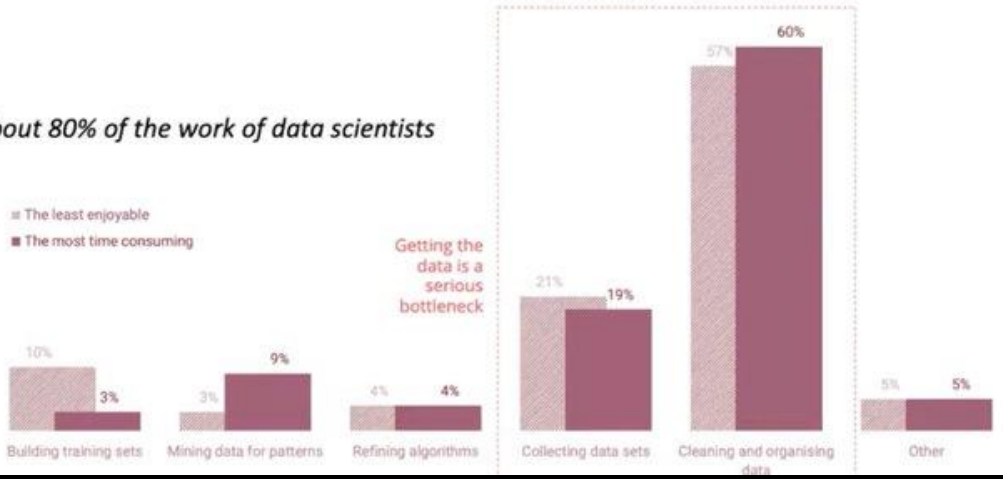
The problem

Data science report, 2016, cit. by Susanna Sansone Apr. 27, 2021

Discoveries are made using shared data and this requires data that are:

- Retrievable and structured in standard format(s)
- Self-described so that third parties can make sense of it

***Data preparation** accounts for about 80% of the work of data scientists*



Forbes article on 2016 Data Scientist Report
https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#276a35e6f637

OF COURSE MANAGING AND PRESERVING DATA COSTS
BUT PLEASE THINK ABOUT

- HOW MUCH WOULD IT COST **NOT TO MANAGE AND PRESERVE**
- HOW MUCH TIME DO YOU SPEND IN «PREPARING» DATA COMING FROM DIFFERENT SOURCES (79%)]

Costs

COSTS OF NOT HAVING FAIR DATA

Following this approach, we found that the annual cost of not having FAIR research data costs the European economy at least €10.2bn every year. In addition, we also listed a number of consequences from not having FAIR which could not be reliably estimated, such as an impact on research quality, economic turnover, or machine readability of research data. By drawing a rough parallel with the European open data economy, we concluded that these unquantified elements could account for another €16bn annually on top of what we estimated. These results relied on a combination of desk research, interviews with the subject matter experts and our most conservative assumptions.

10,2 bn DIRECT
16 bn INDIRECT
26,2 bn TOTAL

Cost of not having FAIR data, 2018

Costs

UK Data Service — UK Data service costing tool — ukdataservice.ac.uk

Data management costing tool and checklist

CHECKLIST OF ANY ASPECTS YOU NEED TO BE TAKEN INTO ACCOUNT FOR DATA MANAGEMENT COSTS

How to use the costing tool

Step 1: Check

Check the data management activities in the table and tick those that may apply to your proposed research.

Step 2: Estimate

For each selected activity, estimate the additional time and/or other resources needed and cost this, e.g., people's time or physical resources needed such as hardware or software. Find out which resources are available to you from your institution. Consider whether you need a dedicated data manager.

Step 3: Implement

Add these data management costs to your research application. Coordinate resourcing and costing with your institution, research office, and institutional IT services.

Step 4: Plan

Plan the data management activities in advance to avoid them competing with the need to focus on research excellence.

The costing tool

Activity	Comments and suggestions	✓	Cost
Data description • Are data in a spreadsheet or database clearly marked with variable and value labels, code descriptions, missing value descriptions, etc? • Are labels consistent? • Do textual data like interview transcripts need description of context, e.g., included as a heading page?	• If data descriptions are implemented as part of data creation, data input or data transcription - low or no additional cost. • If needed to be added afterwards - higher cost. • Codebooks for datasets can often be easily exported from software packages.		
Data cleaning • Do quantitative data need to be cleaned, checked, or verified before sharing, e.g., check validity of codes used, check for anomalous values?	• If carried out as part of data entry and preparation before data analysis - low or no additional cost. • If needed afterwards - higher cost.		
Formatting and organising • Are your data files, spreadsheets, interview transcripts, records, etc. all in a uniform format or style? • Are files, records and items in the collection clearly named with unique file names and well organised?	• If planned beforehand by developing templates and data entry forms for individual data files (transcripts, spreadsheets, databases) and by constructing clear file structures - low or no additional cost. • If needed afterwards - higher cost. • Free software exists for batch file renaming to harmonise file names.		
Transcription • Will you transcribe qualitative data (e.g., recorded interviews or focus group sessions) as part of your research; or will you need to do this specifically so data can be more easily shared and reused? • Is full or partial transcription needed? • Is translation needed? • Will you need to develop a	• If transcription is part of research practice – very low or no additional cost. • If transcription not planned as part of research practice - potentially high cost. • Is additional hardware /software needed? • Consider cost of time needed for developing procedures, templates, and guidance for transcribers.		

1. MANAGING DATA

Good data management facilitates the reuse of data, which helps avoid duplication of effort, and mitigates against data loss. It also supports collaboration, facilitates continuity across projects, and improves the visibility and impact of research outputs.

2020

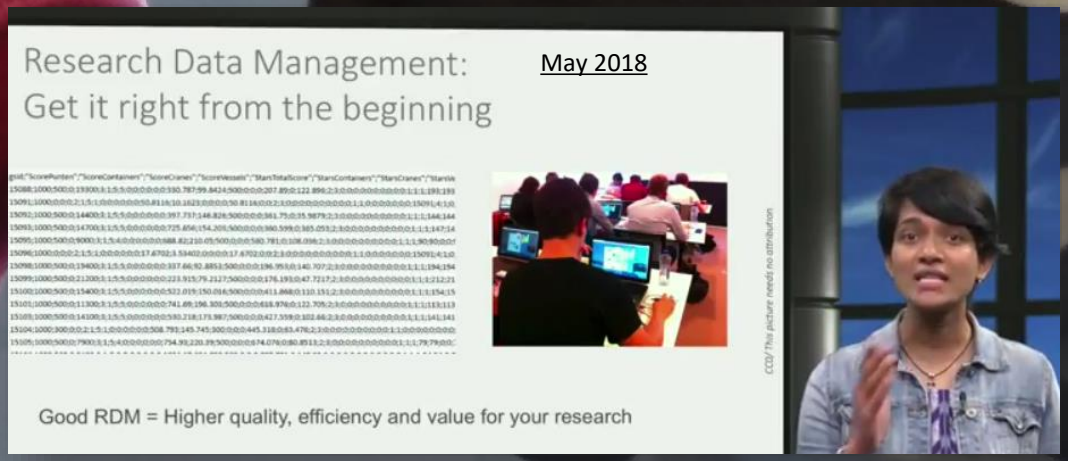
Why managing data

SAVING TIME
PRESERVE
DATA ARE A RESEARCH OUTPUT
INTEGRITY

- **Save Time** – By spending a little bit of up-front time and planning and organising the data you produce you will save time and resources in the long run.
- **Increase your efficiency** – If you document your data properly whenever you or someone else comes to it they will be able to understand it quickly and without difficulty. Thus saving time and increasing efficiency.
- **Preserve and protect your data** – It is relatively easy to produce data that will be useful only the once and for a very specific purpose. Learn how to ensure that the data can be useful again and again, and how to make sure that it is never lost.
- **Data is an output in its own right** – that's right; data itself is increasingly being seen as an important output of research. If shared, it can better enable researchers. The REF (Research Excellence Framework) now takes note of it.
- **Meet grant requirements** – Many funding bodies now require that researchers archive data as well as the resulting publications as part of their project. Good data management will make this easy rather than a last minute chore.
- **Open Access** – In the UK government policy has moved to an open access framework. Producing and making available data is a vital part of this process. Journals are increasingly making room for data alongside articles, for example.
- **Transparency/research integrity** – If required you have all the documents and materials easily available making your research more transparent if questioned.

Why data management

… from the beginning



AS YOU WILL SAVE TIME AND YOUR RESEARCH WILL BE SMOOTHLESS AND MORE EFFICIENT

Add a "version management" tab to your spreadsheet.

Now, let me expand on this idea.

Start by adding an extra "version management" tab to a new spreadsheet. In this sheet, carefully write down a version name (name of the file, typically) in the first column, in the second column the date, and in a third column an explanation of all changes you made to the sheet. Carefully fill out this sheet every single time you move something around, or tinker with the sheet.

If you're a starting PhD student, start doing this the very next time you build a new sheet. Thank me later.

If you already have multiheaded monstrous sheets: start by managing them in this way, and take a few extra hours to redefine the logic behind what you did earlier. Your dissertation writing self will thank you.

PhD Talk 2013 Eva Lantsoght

AS THE FIRST RE-USER IS YOUR FUTURE SELF!!!
SO, A DMP IS NOT A VASTE OF TIME

Main Points for Good Data Management

Data acquisition

- Check the type, source of the data and how to gather/collect it
  - Data types (to help define sensitivity of
  - Data format (to help define the tools an
  - Data size (to help define storage and in
- Check the ownership of the collected and proce
  - Check with the data source about conditions (e.g. licence)
  - Check the need to make a data proce on the ownership / access control
  - Are there (own) institutional policies tha
  - Can the data be shared with other parti
- Confidentiality of the data (if applicable):
  - Register crucial information regarding d
  - Ensure security of confidential data (p data, or data that would harm society w
  - Ensure compliance with General Dat verordening gegevensbescherming whe
  - Ensure there are procedures in place to of a privacy advisor/data protection offic

Data collection

- Establish a workflow for data collection
  - How will the data be collected?
  - Who has access to which data in short / l
  - What resources are needed for data analysis?
  - How will the data be exchanged / transferred among relevant stakeholders?
- Storage arrangement
  - Check available storage capacity and backup strategy

Data storing / backup

- Create a clear folder structure and consistent file naming convention
- Make a backup strategy where data is stored at least two different physical locations and preferably automatically backed up
- Access control to confidential data
- Apply encryption at disk or folder level if needed
- Create a consistent and standard versioning of the data files
- Determine the minimal documentation of the data that is required to find it, understand it and use it

Data sharing

- Create proper data sharing procedures
  - Consider agreements established in the Data acquisition phase, and evaluate/assess data sharing with other parties
  - Be aware of the permission and consequence of sharing confidential data
- Copyright / Licensing
  - How should others use the data
  - Who should be attributed for creating/gathering the data

Organizational Implications

In addition to the above mentioned actions, there are also a few things to consider to make data management a standard practice in daily operations.

ASK THE RIGHT QUESTIONS

Before boarding / 2

EXERCISE TWO

USING THE DATA MANAGEMENT CHECKLIST FOR YOUR RESEARCH PLANNING

Use the data management checklist to help point to relevant data management topics you need to consider when planning your research project.

1/2

DATA MANAGEMENT CHECKLIST	NOTES
DATA MANAGEMENT PLANNING	
Who is responsible for which part of data management?	
Do you need extra resources to manage data, such as people, time or hardware?	
DOCUMENTING YOUR DATA	
Are your structured data self-explanatory in terms of variable names, codes and abbreviations used?	
Which descriptions and contextual documentation can explain: what your data mean, how they were collected and the methods used to create them?	
How will you label and organise data, records and files?	
Will you apply consistency in how data are catalogued, transcribed and organised, e.g. standard templates or input forms?	
DATA FORMATTING	
Are you using standardised and consistent procedures to collect, process, check, validate and verify data?	

UK Data service p. 24

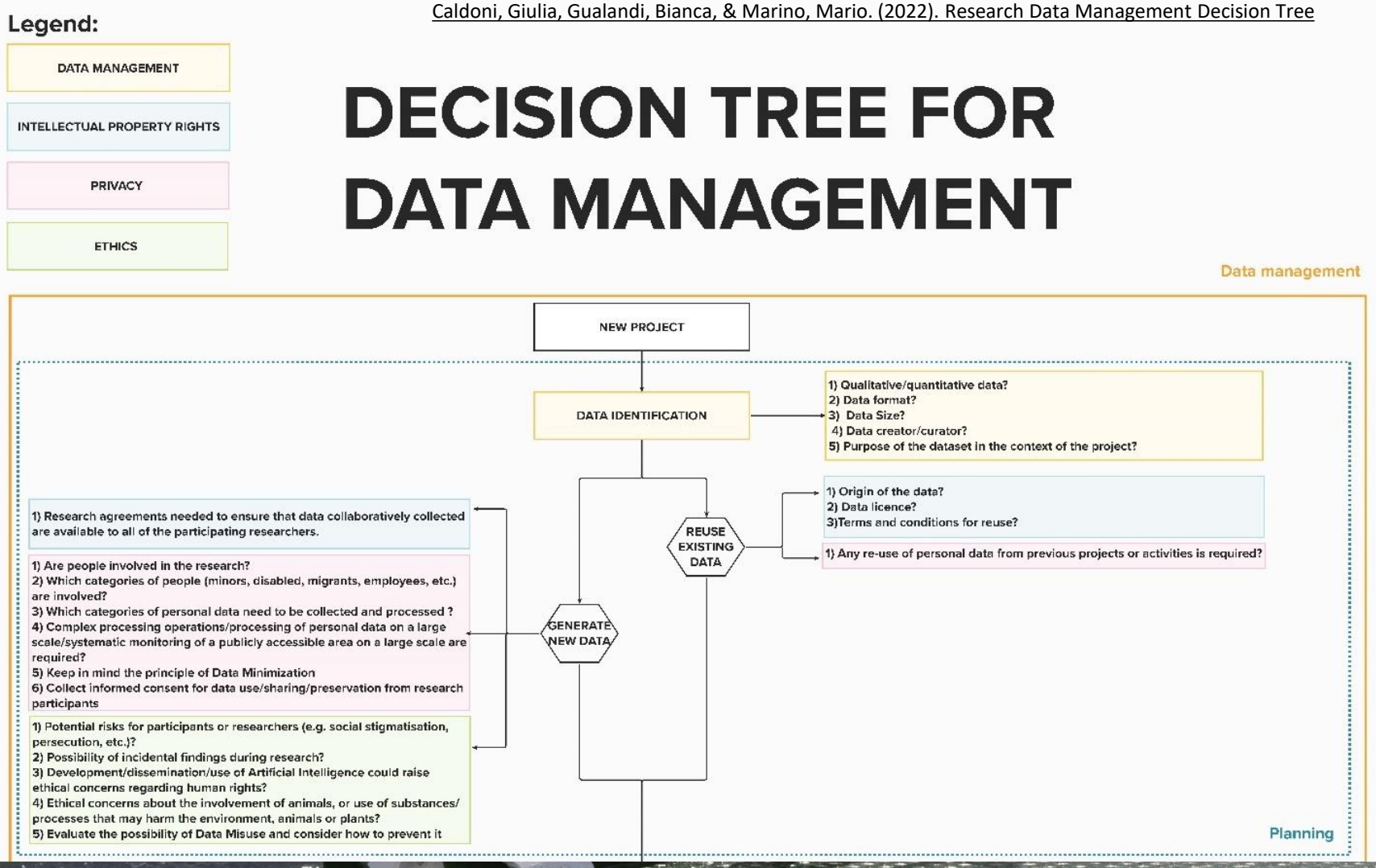
ASK THE RIGHT QUESTIONS

DATA MANAGEMENT CHECKLIST	NOTES
STORING YOUR DATA	
Are your digital and non-digital data, and any copies, held in a safe and secure location?	
Do you need to securely store personal or sensitive data?	
If data are collected with mobile devices, how will you transfer and store the data?	
If data are held in various places, how will you keep track of versions?	
Are your files backed up sufficiently and regularly and are back-ups stored safely?	
Do you know what the master version of your data files is?	
Who has access to which data during and after research? Are various access regulations needed?	
ETHICS AND CONSENT	
Do your data contain confidential or sensitive information? If so, have you discussed data sharing with the respondents from whom you collected the data?	
Are you gaining (written) consent from respondents to share data beyond your research?	
Do you need to anonymise data, e.g. to remove identifying information or personal data, during research or in preparation for sharing?	

Before boarding / 3

USEFUL TOOL AS A FIRST APPROACH TO DATA MANAGEMENT [PLAN]

Caldoni, Giulia, Gualandi, Bianca, & Marino, Mario. (2022). Research Data Management Decision Tree



[remind: it's not open/close at th

…THE ISSUE IS NOT JUST OPEN/CLOSED AT THE END.

DURING MY RESEARCH, WHERE CAN I SAFELY STORE THE DATA?

WHO CAN ACCESS THEM?

WHAT ABOUT SECURITY?

Level	Data Classification and Examples (abridged version)
5	***Information that would cause severe harm to individuals or the University if disclosed.***
	• Research information classified as Level 5 by an IRB or otherwise required to be stored or processed in a high security environment and on a computer not connected to the Harvard data networks • Certain individually identifiable medical records and genetic information, categorized as extremely sensitive
4	***Information that would likely cause serious harm to individuals or the University if disclosed.***
	• High Risk Confidential Information (HRCI) and research information classified as Level 4 by an IRB • Personally identifiable financial or medical information • Information commonly used to establish identity that is protected by state, federal, or foreign privacy laws and regulations • Individually identifiable genetic information that is not Level 5 • National security information (subject to specific government requirements) • Passwords and Harvard PINs that can be used to access confidential information
3	***Information that could cause risk of material harm to individuals or the University if disclosed.***
	• Research information classified as Level 3 by an IRB • Information protected by the Family Educational Rights and Privacy Act (FERPA) to the extent it is not covered under Level 4 including non-directory student information and directory information about students who have requested a FERPA block …mes or any other information that could identify individuals …ds (employees may discuss terms and conditions of employment with each other and third …ectory student information and directory information about students who have requested a …mes or any other information that could identify individuals …ds (employees may discuss terms and conditions of employment with each other and third …ords …tion …on protected under state, federal and foreign privacy laws not classified as Level 4 or 5
	… of which would not cause material harm, but which the University has chosen to
	…ork and intellectual property not in Level 3 or 4 …ssified as Level 2 by an IRB …work papers, drafts of research papers • Building plans and information about the University physical plant
1	***Public information.***
	• Research data that has been de-identified in accordance with applicable rules • Published research • Published information about the University • Course catalogs • Directory information about students who have not requested a FERPA block • Faculty and staff directory information

Harvard security

Data management = security

Sweeney L, Crosas M, Bar-Sinai M. Sharing Sensitive Data with Confidence: The DataTags System. *Technology Science*. 2015101601. October 16, 2015. http://techscience.org/a/2015101601

JOTS Technology Science

Sharing Sensitive Data with Confidence: The Datatags System 2015

Latanya Sweeney, Mercè Crosas, and Michael Bar-Sinai

Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

Training?

[DMP]

cessda Tools Training Strategy & Expertise Development & Impact About

Training

- CESSDA Training
- Training Resources
- Event Calendar
- Data Management Expert Guide
- Data Archiving Guide
- Journals Outreach

Home / Training / Data Management Expert Guide

Data Management Expert Guide (DMEG)

The CESSDA Data Management Expert Gu
scientists at the heart of making their rese
understandable, sustainably accessible an

You will be guided by different European e
basis - busy ensuring long-term access to valuable social
discovery and reuse at one of the CESSDA social science

The core version of the DMEG has been created for CESS
service providers' experts at ADP, AUSSDA, CSDA, DANS,
So.Da.Net and UKDS. DANS has led the creation of the ex

Webinar - CESSDA Data Management Expert Guide: W

Data Management Expert Guide

1. Plan
2. Organise & Document
3. Process
4. Store
5. Protect
6. Archive & Publish
7. Discover
8. Contributors

Are you here for the first time?

Take the quiz below and find out which chapters of DMEG will be most useful for you.

Take the DMEG quiz

Target audience and mission

This guide is written for social science researchers who are in an early stage of practising research data management. With this guide, CESSDA wants to contribute to professionalism
data management and increase the value of research data.

What is the CESSDA Data Management Expert Guide &...

Data Management Expert Guide
https://dmeg.cessda.eu/

Data Management Expert Guide (DMEG)

The DMEG is designed by European experts to help social science researchers make their research data Findable, Accessible, Interoperable and Reusable (**FAIR**).

You will be guided by different European experts who are - on a daily basis - busy ensuring long-term access to valuable social science datasets, available for discovery and reuse at one of the CESSDA social science data archives.

You can download the full DMEG for your personal study offline (DOI: 10.5281/zenodo.3820473). PDFs for every single chapter are also available for being printed as handouts for training.

See also the pilot interactive game version of the guide!

Some [practical] support

AT THE END OF EACH STEP, THERE IS A SECTION «ADAPT YOUR DMP» ACCORDING TO WHAT YOU HAVE JUST LEARNT

Adapt your DMP: part 6

This is the sixth 'Adapt your DMP' section in this tour guide. To adapt your DMP, consider the following elements and corresponding questions:

cessda TRAINING

⊕ Versioning

⊖ Interoperability

In order to be able to link your work to other research, it might be useful to build on established terminologies as well as commonly uses coding and soft- and hardware wherever this is possible.

- Which *software and hardware* will you use? How does this relate to other research?

If applicable:

- Will established *terminologies/ontologies* (i.e. structured controlled vocabularies) be used in the project? If not, how does yours relate to established ones?
- Which *coding* is used (if any)? How does this relate to other research?

⊖ Deposit your data

- Will the data you produce and/or used in the project be useable by third parties, in particular after the end of the project?
- Which data and associated metadata, documentation and code will be deposited?
- 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)?
- What data quality assurance processes will you apply?

CESSDA Guide

Training

RDM kit

RDMkit | Data management | About | Contribute | GitHub | Search RDMkit

Data management

Data life cycle

Collecting

- What is data collection?
- Why is data collection important?
- What should be considered for data collection?
- Related pages
- More information

What is data collection?

Data collection is the process where information is gathered about specific variables of interest either using instrumentation or other methods (e.g. questionnaires, patient records). While data collection methods depend on the field and research subject, it is important to ensure data quality.

You can also reuse existing data in your project. This can either be individual earlier collected datasets, reference data from curated resources or consensus data like reference genomes. For more information see Reuse in the data life cycle.

Why is data collection important?

Apart from being the source of information to build your findings on, the collection phase lays the foundation for the quality of both the data and its documentation. It is important that the decisions made regarding quality measures are implemented, and that the collect procedures are appropriately recorded.

Your role

Your domain

Your tasks

Tool assembly

National resources

All tools and resources

All training resources

RDMkit | Data management

Data management

Data life cycle

Reuse | Plan | Collect | Process | Analyse | Preserve | Share

Your role

Your domain

Your tasks

Tool assembly

National resources

All tools and resources

All training resources

Data life cycle

In this section, information is organised according to the stages of the research data life cycle. You will find:

- A general description and introduction of each stage.
- A list of the main considerations that need to be taken into account during each stage.
- Links to training materials related to each stage.
- Links to related data management tasks that can be performed at each stage.
- Links to a Data Stewardship Wizard for your DMP and to step-by-step instructions to make your data FAIR.

GUIDANCE IN
ANY STEP
YOUR ROLE
#YOUR DOMAIN
#YOUR TASKS

Training

[DMP]

rdnl research data netherlands

Essentials 4 Data Support

Essentials4data

ABOUT THE COURSE › | START THE COURSE › | LOGIN ›

I - A bird's-eye view

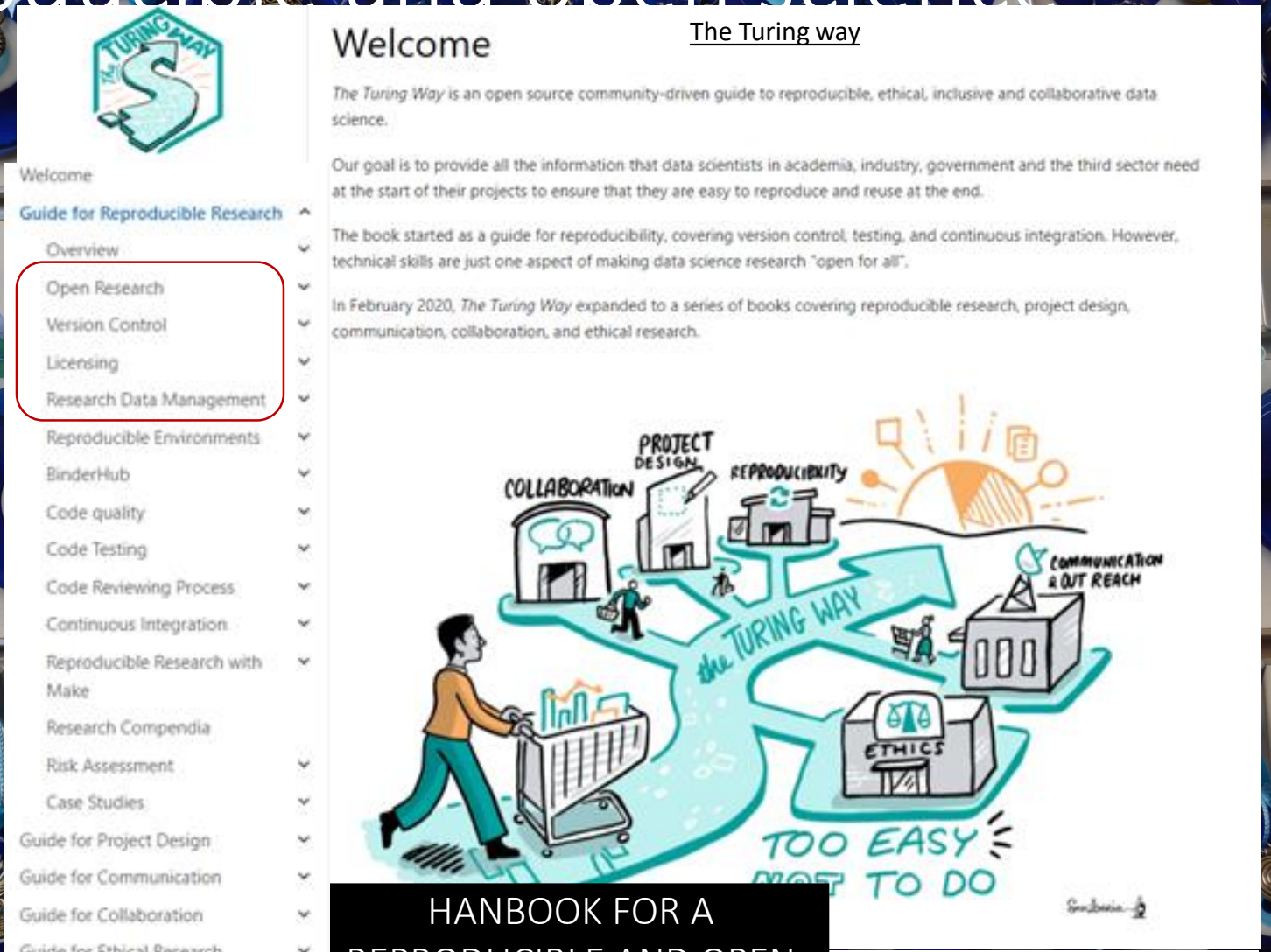
Data jargon

- DOI
- FAIR data
- GDPR
- Integrity
- Linked data
- Metadata
- Open data
- Open science
- Persistent identifier (PID)
- Preferred format

- I - A bird's-eye view ›
- II - Planning phase ›
- III - Research phase ›
- IV - Harvest phase ›
- V - Legislation and policy ›
- VI - Data support ›
- Closing remarks

…reproducible and open science

The Turing way



HANBOOK FOR A REPRODUCIBLE AND OPEN SCIENCE

Training

Data stewards training

Home
Calendario
Sezioni del corso
Impostazioni di accessibilità

Data Steward Training

Home › Corsi › Data › Data Steward Training

Welcome | 1. RDM, FAIR and Open Science | 2. Design training in easy steps
3. Open and responsible research | 4. Data management plans | 5. RDM service delivery | Completion
Authors and contributors

Welcome to the course!

This is an introductory level course targeted at professionals working in research data support roles or individuals with a research background considering a change to a data support role. In this course you will learn the skills and gain knowledge of how to be a **successful data steward**. Five self-paced modules prepared by a **team of international experts** will guide you step by step through relevant topics for your daily work. By watching a series of recorded videos and completing practical assignments, you will be **up to speed with data**

EOSC synergy cou

- What is Open Science?
- What is European Open Science Cloud (EOSC)?
- EOSC in practice: EOSC Synergy
- EOSC in practice: Facilitating software quality across EOSC services
- EOSC in practice: Integrating resources into EOSC
- **Research data management**
- FAIR principles

Accessibility settings

Bring
data
rese

Home
Researc

Welcome!
What is E
Completio

Introduc

The course consists of the following modules:

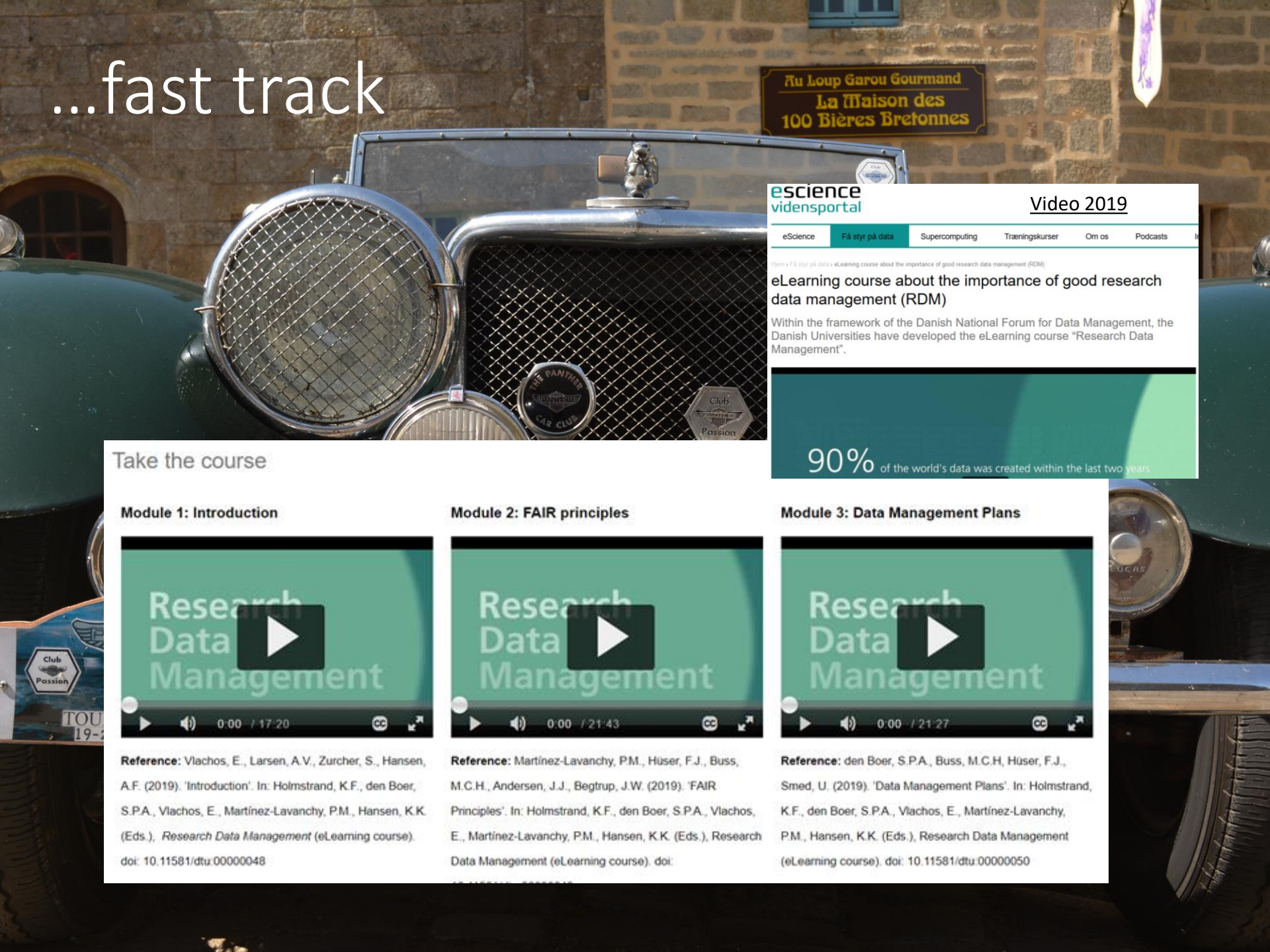
1. RDM, FAIR and open science
   - 1.1. The role of a data steward
   - 1.2. How FAIR aware are you?
2. Responsible and open research
3. Design training in easy steps
4. Data management plans
5. RDM service delivery

By the end of this course you will:

- Be able to explain the difference between **FAIR** and **Open Data** to researchers.
- Be able to develop **training courses** using **open learning resources**.
- Understand the range of **skills and knowledge** associated with **data stewardship**.
- Be able to identify areas where collaboration on **service provision** is most beneficial.

All five modules are **introductory** and you can choose to follow them all or pick the ones that are most relevant to you. **Each module** takes on average **one hour** to complete, requiring around **five to six** hours to complete the **entire course**. Next to recorded video presentations, you will have full scripts, PowerPoint presentations and a

…fast track



escience vidensportal

Video 2019

eScience | Få styr på data | Supercomputing | Træningskurser | Om os | Podcasts

eLearning course about the importance of good research data management (RDM)

Within the framework of the Danish National Forum for Data Management, the Danish Universities have developed the eLearning course "Research Data Management".

90% of the world's data was created within the last two years

Take the course

Module 1: Introduction

Research Data Management 0:00 / 17:20

Reference: Vlachos, E., Larsen, A.V., Zurcher, S., Hansen, A.F. (2019). 'Introduction'. In: Holmstrand, K.F., den Boer, S.P.A., Vlachos, E., Martínez-Lavanchy, P.M., Hansen, K.K. (Eds.), *Research Data Management* (eLearning course). doi: 10.11581/dtu:00000048

Module 2: FAIR principles

Research Data Management 0:00 / 21:43

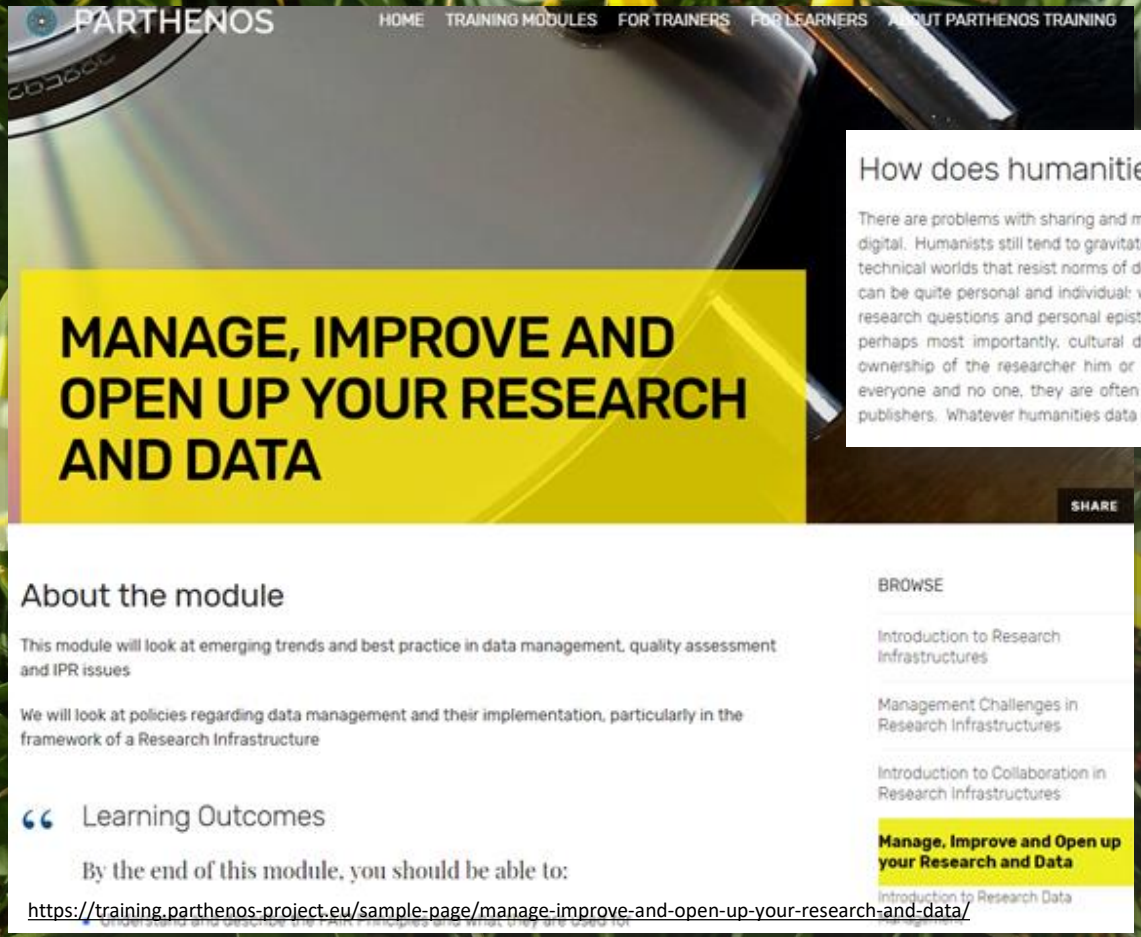
Reference: Martínez-Lavanchy, P.M., Hüser, F.J., Buss, M.C.H., Andersen, J.J., Begtrup, J.W. (2019). 'FAIR Principles'. In: Holmstrand, K.F., den Boer, S.P.A., Vlachos, E., Martínez-Lavanchy, P.M., Hansen, K.K. (Eds.), Research Data Management (eLearning course). doi:

Module 3: Data Management Plans

Research Data Management 0:00 / 21:27

Reference: den Boer, S.P.A., Buss, M.C.H, Hüser, F.J., Smed, U. (2019). 'Data Management Plans'. In: Holmstrand, K.F., den Boer, S.P.A., Vlachos, E., Martínez-Lavanchy, P.M., Hansen, K.K. (Eds.), Research Data Management (eLearning course). doi: 10.11581/dtu:00000050

...and the humanities?



https://training.parthenos-project.eu/sample-page/manage-improve-and-open-up-your-research-and-data/

How does humanities data tend to be different?

There are problems with sharing and managing the humanistic data, however. First of all, much of it is not digital. Humanists still tend to gravitate toward multimodal knowledge creation systems, hybrid digital and technical worlds that resist norms of deposit and reuse. Second, the semiotic systems of humanities data can be quite personal and individual: we prepare our sources to be useful for us, and what works for our research questions and personal epistemic instruments may not work at all for anyone else. Finally, and perhaps most importantly, cultural data is seldom if ever 'raw,' and seldom, if ever, under the sole ownership of the researcher him or herself. The records of human activity and creativity belong to everyone and no one, they are often preserved and curated by dedicated public institutions or private publishers. Whatever humanities data is, it is not simple!

… and the humanities?

DARIAH-CAMPUS Resources Topics Sources Course Registry About Contact

May 2019

DARIAH Pathfinder to Data Management Best Practices in the Humanities

Written by Erzsébet Tóth-Czifra
May, 03 2019

Source: DARIAH Pathfinders, DARIAH
Topics: Data management

TABLE OF CONTENTS

1. Why research data management?
2. Data in the Humanities
3. The devil is in the conte… a processural view on data curation
4. Sharing your data
   4.1. Cite to be cited!
   4.2. Be aware of your licensing options
   4.3. A case study: different levels of being a open scholar
5. A recipe for your researc… project: the Data Management Plan
6. Data in publications and… data as publications

1. Why research data management?

Systematically planning how you will collect, document, organize, manage, share and preserve your data has many benefits. It helps to build a common framework of understanding with your

10. THE RISK OF LOSING THE *THICK DESCRIPTION*: DATA MANAGEMENT CHALLENGES FACED BY THE ARTS AND HUMANITIES IN THE EVOLVING FAIR DATA ECOSYSTEM 235

Erzsébet Tóth-Czifra

Edmond, 2020

TABLE OF CONTENTS

OPENMETHODS

HIGHLIGHTING DIGITAL HUMANITIES METHODS AND TOOLS

HOME ABOUT WHO WE ARE JOIN US SUBMIT A CONTENT RSS FEEDS LOG IN

OpenMethods

ANALYSIS

The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models

APRIL 29, 2021 · BY ERZSEBET TÓTH-CZIFRA

ANALYSIS

Cultural Ontologies: the ArCo Knowledge Graph.

MARCH 11, 2021 · BY MARINELLA TESTORI

Introduction: Standing for 'Architecture of Knowledge', ArCo is an open set of resources developed and managed by some Italian institutions, like the MiBAC (Minister

Search …

INTERESTED IN BLOGGING ABOUT YOUR RESEARCH? THE DIGITAL HUMANITIES TOOLS AND METHODS BLOG IS FOR YOU!

hypotheses

IN COOPERATION WITH

DARIAH-EU

Data Management - caveat

[DMP]

Therefore, it doesn't necessarily matter if you plan to share your data with other scholars, what matters is considering this prospect as you work out how you are going to go about your research. It will help you to understand what it is you are doing more clearly and give you the basis to share that data later on if you so wish.

PORT DMP

PORT postgraduate online research training | SCHOOL OF ADVANCED STUDY UNIVERSITY OF LONDON

IT DOES NOT MATTER IF IN THE END YOU WILL OPENLY SHARE YOUR DATA OR NOT.
HERE YOU SEE HOW YOU ARE GOING TO GO WITH YOUR RESEARCH

An overview

Research Data Management*

**Standard/best practices for accurate data/code collection, processing, documentation, analysis, storage & preservation as a prerequisite for open science (FAIR ≠ Open).*

- **What decisions do researchers make to achieve 'FAIR' data management?**

Decision making procedures in data management and data stewardship for Open Science 2022

Connie Clare, PhD

A GOOD DATA MANAGEMENT IS ONLY A PREREQUISITE FOR FAIR/OPEN
…WITH HUGE BENEFITS

The importance & benefits of RDM

- Efficiency & avoids duplication of efforts (saves time, money & resources)
- Transparency with internal/external collaborators
- Easier data publication for long-term preservation *'prepare to share'*
- Reproducibility & verification → reusability
- Accountability for data quality → integrity & confidence
- Increased impact, greater visibility & citations
- Compliance with legislation (GDPR, legal & ethical)
- Compliance with policy (Institutional, Journal & Funder)

FAIR data management

[DMP]

DATA MANAGEMENT SHOULD BE «FAIR BY DESIGN»

RDA

Decision making procedures in data management and data stewardship for Open Science

2022

Connie Clare, PhD

RDA

What decisions do researchers make to achieve 'FAIR' data management?

Who is responsible for data management?

How will new data be collected or produced and/or how will existing data be reused?

What data types, formats, and volumes will be collected or produced?
Structured data, formats: JSON, XML, CSV

What metadata and documentation will accompany data?

Are there disciplinary standards and vocabularies that should be used?

What resources will be dedicated to data management and ensuring that data will be FAIR?

How, when and which will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Policies

DMP

What data quality control measures will be used?

If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?

What methods or software tools will be needed to access and use the data?

Will the application of a unique persistent identifier (e.g., DOI) be assigned to the dataset?

Is an ethical review (HREC, ERB) required?

Is informed consent required?

Does data need to be anonymised or pseudonymised?

Which license will be suitable to specify data modification, redistribution and reuse?

How will data for preservation be selected? Where will data be preserved long-term?

How will data and metadata be stored and backed up during the research process?

How will other legal issues, such as IPR and ownership, be managed? What legislation is applicable?

How will possible ethical issues be taken into account, and codes of conduct followed?

Data management ABC – File naming

WOULD YOU BE ABLE IN ONE YEAR TO SAY WHAT THE CONTENT IS?

EXERCISE ONE

FILE NAMING

1. Read through the following file names.
2. If you returned to this data folder in a year's time do you think you would be able to recognise what each of these files contains?
3. What information do you think you need in a file name in order to identify what is in the file's contents?

Doc. 1	My data
IMPORTANT	My Passwords
Thesis Final final	Thesis version 12
My study	Data chart for interviews
Interview with Jane	Int 1 (2)

Data management ABC – File naming

[DMP]

CHOOSE A SCHEMA AND BE CONSISTENT!
[ALL THE MORE SO IF YOU HAVE PARTNERS]

File naming conventions

The conventions comprise the following 13 rules. Follow the links for examples and explanations of the rules.

1. Keep file names short, but meaningful
2. Avoid unnecessary repetition and redundancy in file names and file paths.
3. Use capital letters to delimit words, not spaces or underscores
4. When including a number in a file name always give it as a two-digit number, i.e. 01-99, unless it is a year or another number with more than two digits.
5. If using a date in the file name always state the date 'back to front', and use four digit years, two digit months and two digit days: YYYYMMDD or YYYYMM or YYYY or YYYY-YYYY.
6. When including a personal name in a file name give the family name first followed by the initials.
7. Avoid using common words such as 'draft' or 'letter' at the start of file names, unless doing so will make it easier to retrieve the record.
8. Order the elements in a file name in the most appropriate way to retrieve the record.
9. The file names of records relating to recurring events should include the date and a description of the event, except where the inclusion of any of either of these elements would be incompatible with rule 2.
10. The file names of correspondence should include the name of the correspondent, an indication of the subject, the date of the correspondence and whether it is incoming or outgoing correspondence, except where the inclusion of any of these elements would be incompatible with rule 2.
11. The file name of an email attachment should include the name of the correspondent, an indication of the subject, the date of the correspondence, 'attch', and an indication of the number of attachments sent with the covering email, except where the inclusion of any of these elements would be incompatible with rule 2.
12. The version number of a record should be indicated in its file name by the inclusion of 'V' followed by the version number and, where applicable, 'Draft'.
13. Avoid using non-alphanumeric characters in file names.

File naming

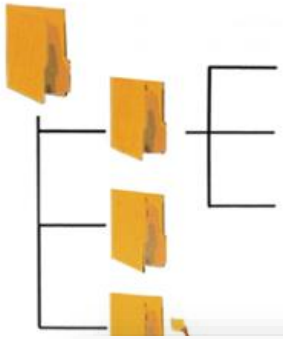
[DMP]

Data management ABC – File naming

Folder structure

CESSDA training

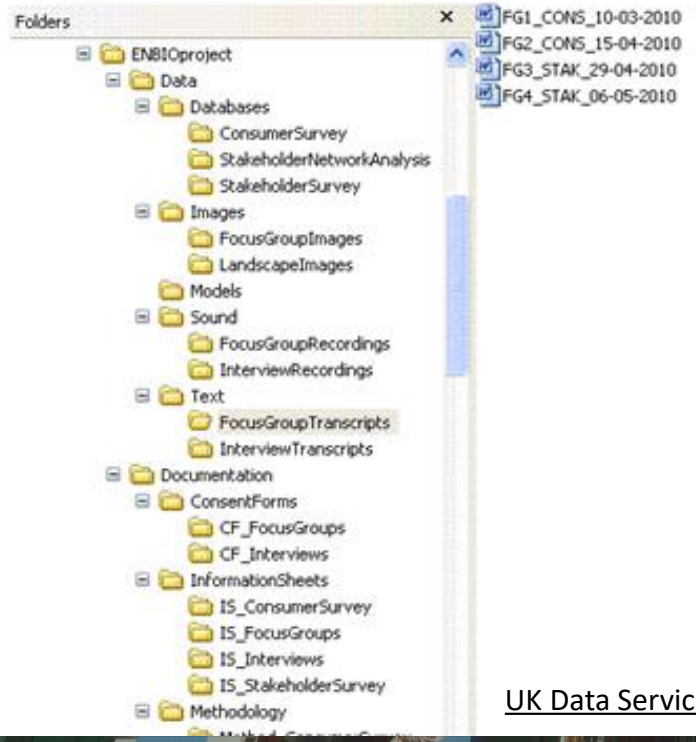
Structuring your data files in folders is important for making it easier to locate and organise files and versions. A proper folder structure is especially needed when collaborating with others.



to organise your data
plan and organisation of
l relevant to the data
to the data folders,
formation on the data
rocessing procedures.

rarchy of your files and
p or shallow hierarchy is
e several independent
advisable to create a separate data folder
ok at the examples in the accordion below

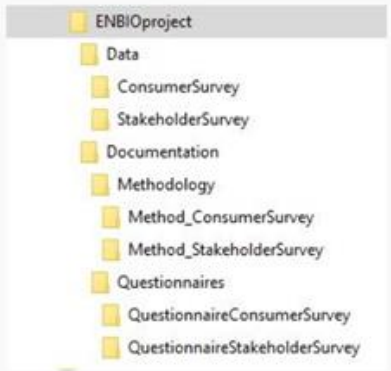
It helps to restrict the level of folders to three or four deep and not to have more than ten items in each list.



UK Data Service

Survey data

For this survey, data and documentation files are held in separate folders. Data files are to data type and then according to research activity. Documentation files are organised documentation file and research activity. It helps to restrict the level of folders to three more than ten items on each list.



[DMP]

[README file]

Data management ABC – Readme file

[DMP]

README FILE GUIDE+ TEMPLATE

Cornell University

Home About Services Data Management Planning Best Practices

RESEARCH DATA MANAGEMENT SERVICE GROUP

Comprehensive Data Management Planning & Services

Guide to writing "readme" style metadata Cornell

A readme file provides information about a data file and is intended to help ensure that the data can be correctly interpreted, by yourself at a later date or by others when sharing or publish
appropriate standard exists, for internal use, writin

Want a template? **Download one** and adapt it f

- Best practices
- Recommended content
  - General information
  - Data and file overview
  - Sharing and access information
  - Methodological information
  - Data-specific information

Best practices

Create readme files for logical "clusters" of data. In many cases it wi
multiple, related, similarly formatted files, or files that are logically gr
Sometimes it may make sense to create a readme for a single data fil

Name the readme so that it is easily associated with the data file(s)

Write your readme document as a plain text file, avoiding proprietary
document so it is easy to understand (e.g. separate important pieces of information with blank lines, rather than having all the information in one long paragraph).

Format multiple readme files identically. Present the information in the same order, using the same terminology.

Use standardized date formats. Suggested format: W3C/ISO 8601 date standard, which specifies the international standard notation of YYYY-MM-DD or YYYY-MM-DDThh:mm:ss.

Follow the scientific conventions for your discipline for taxonomic, geospatial and geologic names and keywords. Whenever possible, use terms from standardized taxonomies and vocabularies, a few of which are listed below.

Source	Content	URL
Getty Research Institute Vocabularies	geographic names, art & architecture, cultural objects, artist names	http://www.getty.edu/research/tools/vocabularies/
Integrated Taxonomic	taxonomic information on plants, animals, fungi, microbes	http://www.itis.gov/

Cornell AUTHOR_DATASET_ReadmeTemplate.txt

DATA & FILE OVERVIEW

1. File List:
<list all files (or folders, as appropriate for dataset organization) contained in the dataset, with a brief description>

2. Relationship between files, if important:

3. Additional related data collected that was not included in the current data package:

4. Are there multiple versions of the dataset? yes/no
   A. If yes, name of file(s) that was updated:
      i. Why was the file updated?
      ii. When was the file updated?

METHODOLOGICAL INFORMATION

1. Description of methods used for collection/generation of data:
<Include links or references to publications or other documentation containing experimental design or protocols used in data collection>

2. Methods for processing the data:
<describe how the submitted data were generated from the raw or collected data>

Data mai
Readme

README FILE GUIDE + TEMPLATE


Cornell University

Home About Services Data Management Planning Best Practices

RESEARCH DATA MANAGEMENT SERVICE GROUP

Comprehensive Data Management Planning & Services

Guide to writing "readme" style metadata Cornell

[DMP]

General information

1. **Provide a title for the dataset**
2. **Name/institution/address/email information for**
   - **Principal investigator (or person responsible for collecting the data)**
   - Associate or co-investigators
   - Contact person for questions
3. **Date of data collection (can be a single date, or a range)**
4. **Information about geographic location of data collection**
5. Keywords used to describe the data topic
6. Language information
7. Information about funding sources that supported the collection of the data

Data and file overview

1. **For each filename, a short description of what data it contains**
2. Format of the file if not obvious from the file name
3. If the data set includes multiple files that relate to one another, the
   structure that holds them (possible terminology might include "dat
4. **Date that the file was created**
5. Date(s) that the file(s) was updated (versioned) and the nature of t
6. Information about related data collected but that is not in the desc

Sharing and access information

1. **Licenses or restrictions placed on the data**
2. Links to publications that cite or use the data
3. Links to other publicly accessible locations of the data (see best practices for sharing data for more information about identifyi
   repositories)
4. Recommended citation for the data (see best practices for data citation)

Methodological information

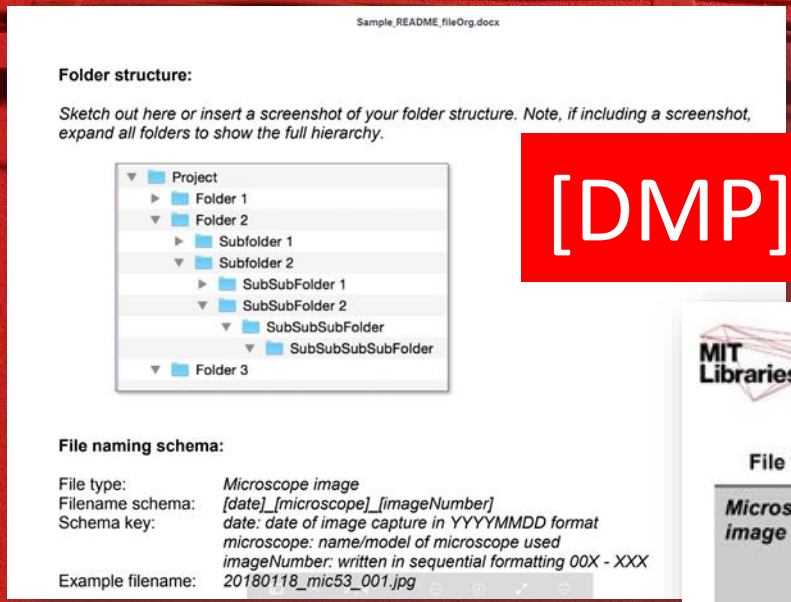
1. **Description of methods for data collection or generation** (include links or references to publications or other documentation containing experimental design or protocols used)
2. **Description of methods used for data processing (describe how the data were generated from the raw or collected data)**
3. Any software or instrument-specific information needed to understand or interpret the data, including software and hardware version numbers
4. Standards and calibration information, if appropriate
5. Describe any quality-assurance procedures performed on the data
6. Definitions of codes or symbols used to note or characterize low quality/questionable/outliers that people should be aware of
7. People involved with sample collection, processing, analysis and/or submission

Data-specific information

Repeat this section as needed for each dataset (or file, as appropriate)

1. Count of number of variables, and number of cases or rows

Data management ABC – Readme file



Sample_README_fileOrg.docx

Folder structure:

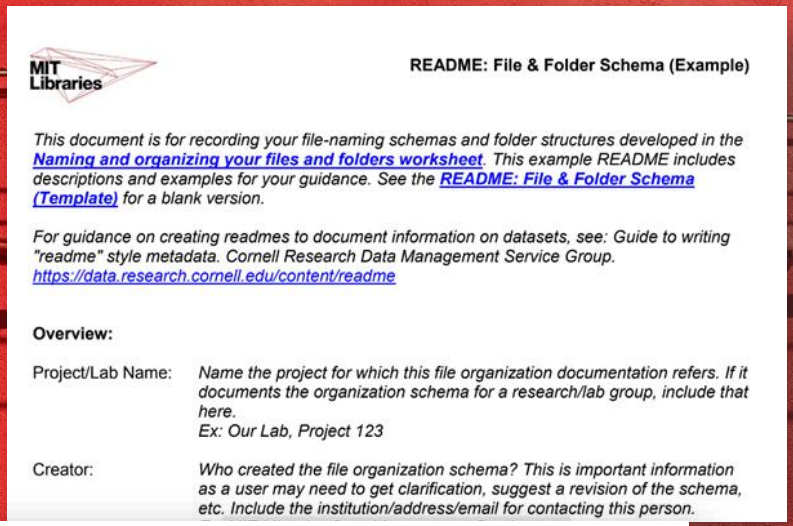
Sketch out here or insert a screenshot of your folder structure. Note, if including a screenshot, expand all folders to show the full hierarchy.

- Project
  - Folder 1
  - Folder 2
    - Subfolder 1
    - Subfolder 2
      - SubSubFolder 1
      - SubSubFolder 2
        - SubSubSubFolder
          - SubSubSubSubFolder
  - Folder 3

File naming schema:

File type: *Microscope image*
Filename schema: *[date]_[microscope]_[imageNumber]*
Schema key: *date: date of image capture in YYYYMMDD format*
microscope: name/model of microscope used
imageNumber: written in sequential formatting 00X - XXX
Example filename: *20180118_mic53_001.jpg*

[DMP]



MIT Libraries

README: File & Folder Schema (Example)

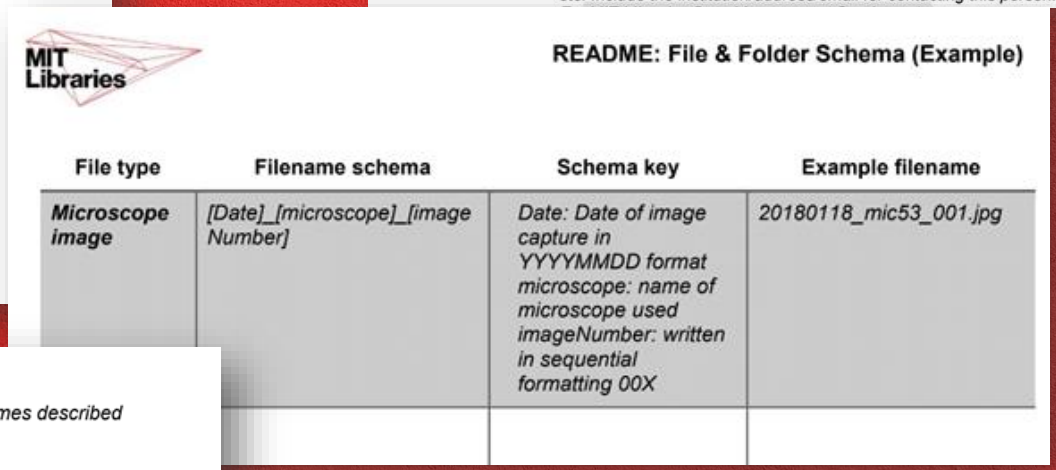
*This document is for recording your file-naming schemas and folder structures developed in the **Naming and organizing your files and folders worksheet**. This example README includes descriptions and examples for your guidance. See the **README: File & Folder Schema (Template)** for a blank version.*

For guidance on creating readmes to document information on datasets, see: Guide to writing "readme" style metadata. Cornell Research Data Management Service Group. https://data.research.cornell.edu/content/readme

Overview:

Project/Lab Name: *Name the project for which this file organization documentation refers. If it documents the organization schema for a research/lab group, include that here.*
Ex: Our Lab, Project 123

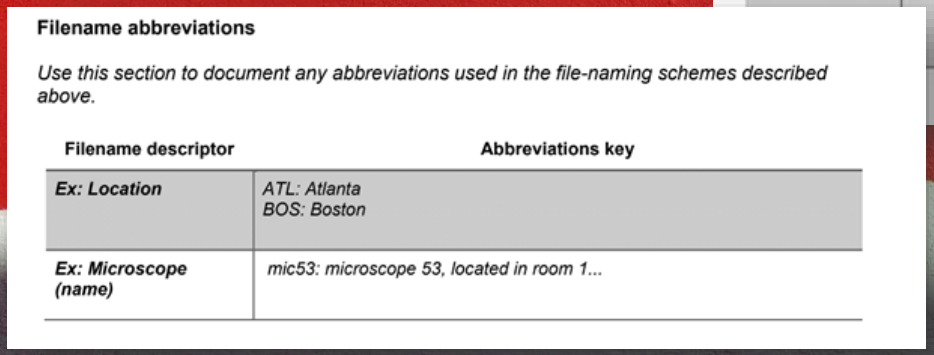
Creator: *Who created the file organization schema? This is important information as a user may need to get clarification, suggest a revision of the schema, etc. Include the institution/address/email for contacting this person.*



MIT Libraries

README: File & Folder Schema (Example)

File type	Filename schema	Schema key	Example filename
Microscope image	*[Date]_[microscope]_[image Number]*	*Date: Date of image capture in YYYYMMDD format* *microscope: name of microscope used* *imageNumber: written in sequential formatting 00X*	*20180118_mic53_001.jpg*



Filename abbreviations

Use this section to document any abbreviations used in the file-naming schemes described above.

Filename descriptor	Abbreviations key
Ex: Location	*ATL: Atlanta* *BOS: Boston*
Ex: Microscope (name)	*mic53: microscope 53, located in room 1...*

MIT data management

Data management A
Readme file

[DMP]

1. *Introductory information*
   - **Title of the dataset**
   - **For each file or group of similar files, a short description of what data it contains**
   - Explain the file naming convention, if applicable
   - Format of the file if not obvious from the file name
   - If the data set includes multiple files that relate to each other, the relationship between the files or a description of the file structure that holds them
   - Contact information; in case users have questions regarding the data files

2. *Methodological information*
   - **Method description for collecting or generating the data, as well as the methods for processing data, if data other than raw data are being contributed**
   - Any instrument-specific information needed to understand or interpret the data
   - Software (including version number) used to produce, prepare, rende compress, analyze and/or needed to read the dataset, if applicable
   - Standards and calibration information, if appropriate

4TU. CENTRE FOR RESEARCH DATA

Guidelines for README

Guidelines for creating a README file

A readme file provides information about a dataset and is intended to help ensure that the data can be correctly interpreted, by yourself at a later date or by others when sharing or publishing data.
A readme file must be submitted along with the dataset file(s).
The outline below should be completed with information relevant to the submitted dataset.

Best practices

- **Create one readme file for each dataset**
- **Name the file README;** not readme, read_me, ABOUT, etc.
- **Write your readme document as a plain text file;** save as README.txt or README.md when writing in Markdown. Or use README.pdf when text formatting is important for your file.

3. *Data specific information*
   - **Full names and definitions (spell out abbreviated words) of column headings for tabular data**
   - **Units of measurement**
   - **Definitions for codes or symbols used to record missing data**
   - **Specialized formats or abbreviations used**

4. *Sharing and Access information*
   - Licenses or restrictions placed on the data;
     Licenses allow you to specify the 'terms-of-use' for your data. The archive provides a license that is explained in its terms of use and applies this license as default selection. You can use this licensing wizard to help you to pick a more appropriate license for the use of your data. This license will then be displayed in the metadata.

[examples of documentat

Structured tabular data should have as documentation (where applicable):

- variable names, labels and descriptions (maximum 80 characters)
- units of measurement for variables
- reference to the question number of a survey or questionnaire

Example: variable 'q11hexw' with label 'Q11: hours spent taking physical exercise in a typical week' —— the label gives the unit of measurement and a reference to the question number (Q11)

- value code labels

Example: variable 'p1sex' = 'sex of respondent' with codes '1=female', '2=male', '8=don't know', '9=not answered'

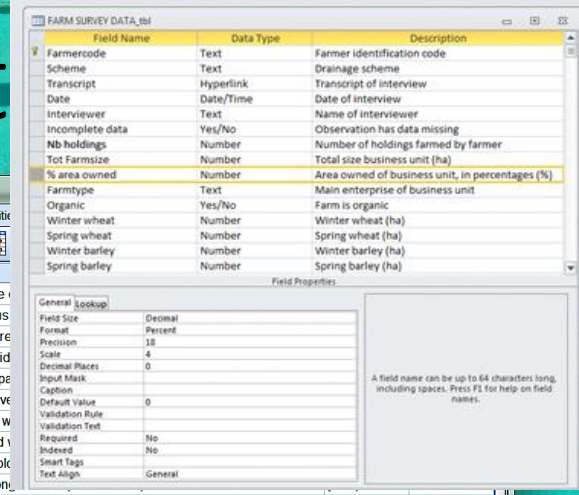
- coding and classification schemes explained, with a bibliographic and dated reference (some standards change over time)

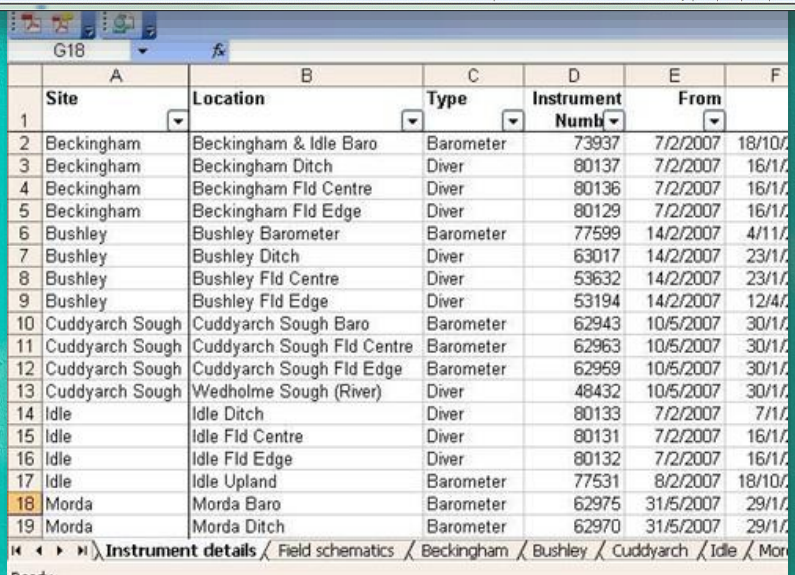
Examples: Standard Occupational Classification, 2000 —— a series of codes to classify respondents' jobs; ISO 3166 alpha-2 country codes —— an international standard of 2-letter country codes

- codes for missing data, with reason data are missing (blanks, system-missing or '0' values are best avoided)

Example: '99=not recorded', '98=not provided (no answer)', '97=not applicable', '96=not known', '95=error'

- deviating universe information for variables in case of skipped cases or questions
- derived or constructed variables created after collection, giving code, algorithm or command files used to create them —— simple derivations, such as grouping age data into age intervals, can be explained in the variable and value labels; complex derivations can be described by providing the algorithms, logical statements or functions used to create derived variables, such as the SPSS or Stata command





[examples of documentation]

UK Data Service

Variable Information Log

UK data service - data documentation

Introduction

For datasets being deposited that include secondary data resources, researchers are advised to prepare a descriptive Variable Information Log describing these resources. The Variable Information Log should include the variable name, its source, how it was collected, a brief description, and any restrictions noted on its further use. (See the notes below)

Notes

These fields should be completed for the original data sources for each variable:

Variable name:	Provide a list of all the variables (name/number) used in the dataset.
Variable label:	A brief description necessary to identify the variable.
Source:	Source of the dataset/data owner or producer (e.g. World Bank data, IMF data, Penn World Tables data).
Dataset version:	Datasets keep evolving, so best practice is to indicate which version has been used.
URL/DOI:	Provide a persistent identifier or link of the source dataset used. Alternatively, if the data are not available online, provide a brief description of how they were obtained.
License information:	Please indicate the licensing information (type of data), as it is important to ensure that the researchers have permission from the data owners. For example, Open data, Data owned by the researcher (you), Data owned by another researcher or Third party licensed data.
Unit of analysis	Indicate the unit of analysis used in the primary dataset (individuals, cases, addresses).
Date data downloaded/obtained	It is important to state the date when the dataset was downloaded or obtained and used for analysis. The data source may have been updated since that time.
Brief description of the data:	Provide a brief description of the dataset, including what was the aim of the study. If a codebook is publicly available for the data used, provide a link.
Data collection method:	Where the data collection procedure for the dataset is well documented, provide a link to that information. If there is little information available, provide a brief description on how data were gathered.

Data management ABC – Versioning

[DMP]

Data versioning

ands AUSTRALIAN NATIONAL DATA SERVICE

About us

Working with data

SHARE

What do we mean by the term ‘data versioning’?

A version is “a particular form of something differing in certain respects from an earlier form or other forms of the same type of thing □”. In the research environment, we often think of versions as they pertain to resources such as manuscripts, software or data. We may regard a new version to be created when there is a change in the structure, contents, or condition of the resource.

In the case of research data, a new version of a dataset may be created when an existing dataset is reprocessed, corrected or appended with additional data. Versioning is one means by which to track changes associated with ‘dynamic’ data that is not static over time.

Why is data versioning important?

Increasingly, researchers are required to cite and iden
to support research reproducibility and trustworthines
accurately indicate exactly which version of a dataset
particularly challenging where the data to be cited are
accessed via a web service.

Unlike the software domain, the data community doesn't yet have a standard numbering system. Three representative data version numbering patterns in use include:

- Numbering system 1
- Numbering system 2
- Numbering system 3

What tools are available for data versioning?

There is no one-size-fit-all solution for data versioning and tracking changes. Data come in different forms and are managed by different tools and methods. In principle, data managers should take advantage of data management tools that support versioning and track changes.

Example approaches include:

Git (and Github) for Data □ (with size <10Mb or 100k rows) which allows:

- effective distributed collaboration – you can take my dataset, make changes, and share those back with me (and different people can do this at once)
- provenance tracking (i.e. what changes came from where)
- sharing of updates and synchronizing datasets in a simple, effective, way.

Data versioning at ArcGIS □

- Users of ArcGIS can create a geodatabase version, derived from an existing version. When you create a version, you specify its name, an optional description, and the level of access other users have to the version. As the owner of the version, you can change these properties or delete a version at any time

Numbering system 1

Data versioning follows a similar path to software versioning, usually applying a two-part numbering rule: Major.Minor (e.g. V2.1). Major data revision indicates a change in the formation and/or content of the dataset that may bring changes in scope, context or intended use. For example, a major revision may increase or decrease the statistical power of a collection, require change of data access interfaces, or enable or disable answering of more or less research questions. A Major revision may incorporate:

- substantial new data items added to /deleted from a collection
- data values changed because temporal and/or spatial baseline changes
- additional data attributes introduced
- changes in a data generation model
- format of data items a changed
- major changes in upstream datasets.

Minor revisions often involve quality improvement over existing data items. These changes may not affect the scope or intended use of initial collection. A Minor revision may include:

- renaming of data attribute
- correction of errors in existing data
- re-running a data generation model with adjustment of some parameters
- minor changes in upstream datasets.

Data management ABC – Versioning

[DMP]

University of Leicester Version chart

Good Practice and Guidance – Document Version Control Chart (Draft)

1. Create Document/File

- Save the document according to file naming guidance/good practice.

2. Document Identification

- Identify on the document e.g. in header or footer, the author, filename, page number and date the document is created/revised.

3. Version Control Table

- Versions and changes documented with Version Control Table where significant/formal/project based.

4. Version Number

- Current version number identified on the first page and where appropriate, incorporated into the header or footer of the document.
- Version number is included as part of the file name.

5. First Draft Version

- Named as version "0-1" (no full stops in electronoic file names).
- Subsequent draft versions 0-2, 0-3, 0-4 ...

6. First Final/Approved Version

- When document is final/approved it becomes version 1-0.

7. Changes to Final Version

- Changed/revised final version becomes x-1.
- Subsequent drafts to Final version become e.g. 1-1, 1-2, 1-3 etc.

8. Further Final/Approved Documents

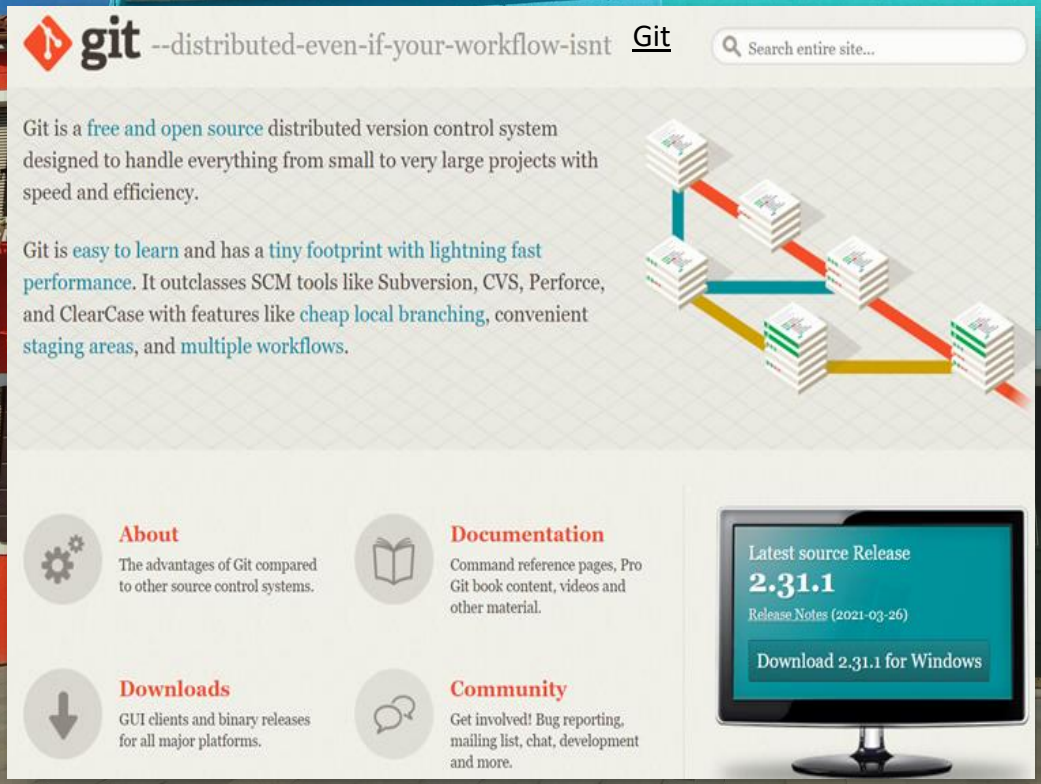
- Version number increased by "1-0" e.g. 1-0, 2-0, 3-0 etc.
- e.g. Amendments to Final 1-0 are 1-1, 1-2, 1-3 and as approved becomes 2-0.

Example version control table: UK Data Service

Title:		Vision screening tests in Essex nurseries	
File Name:		VisionScreenResults_00_05	
Description:		Results data of 120 Vision Screen Tests carried out in 5 nurseries in Essex during June 2007	
Created By:		Chris Wilkinson	
Maintained By:		Sally Watsley	
Created:		04/07/2007	
Last Modified:		25/11/2007	
Based on:		VisionScreenDatabaseDesign_02_00	
Version	Responsible	Notes	Last amended
00_05	Sally Watsley	Version 00_03 and 00_04 compared and merged by SW	25/11/2007
00_04	Vani Yussu	Entries checked by VY, independent from SK	17/10/2007
00_03	Steve Knight	Entries checked by SK	29/07/2007
00_02	Karin Mills	Test results 81-120 entered	05/07/2007
00_01	Karin Mills	Test results 1-80 entered	04/07/2007

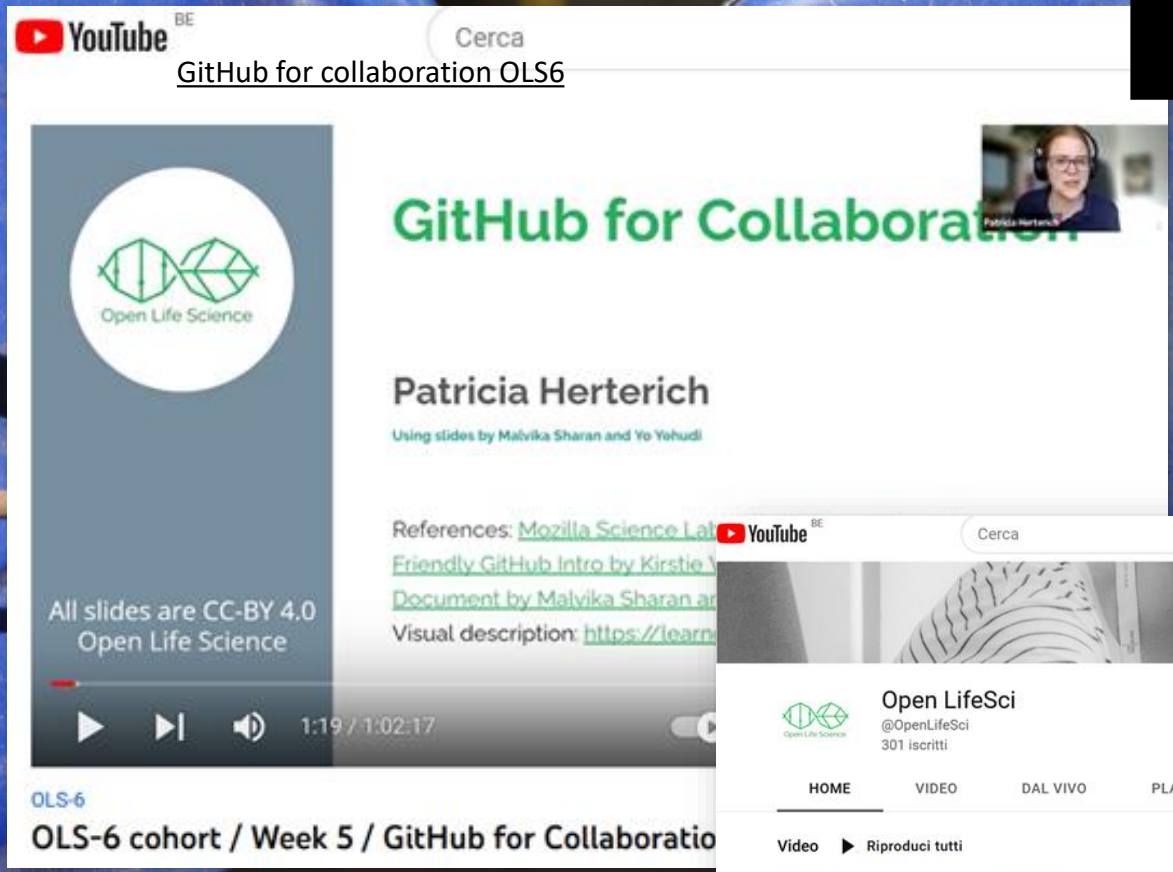
Data management ABC – Versioning

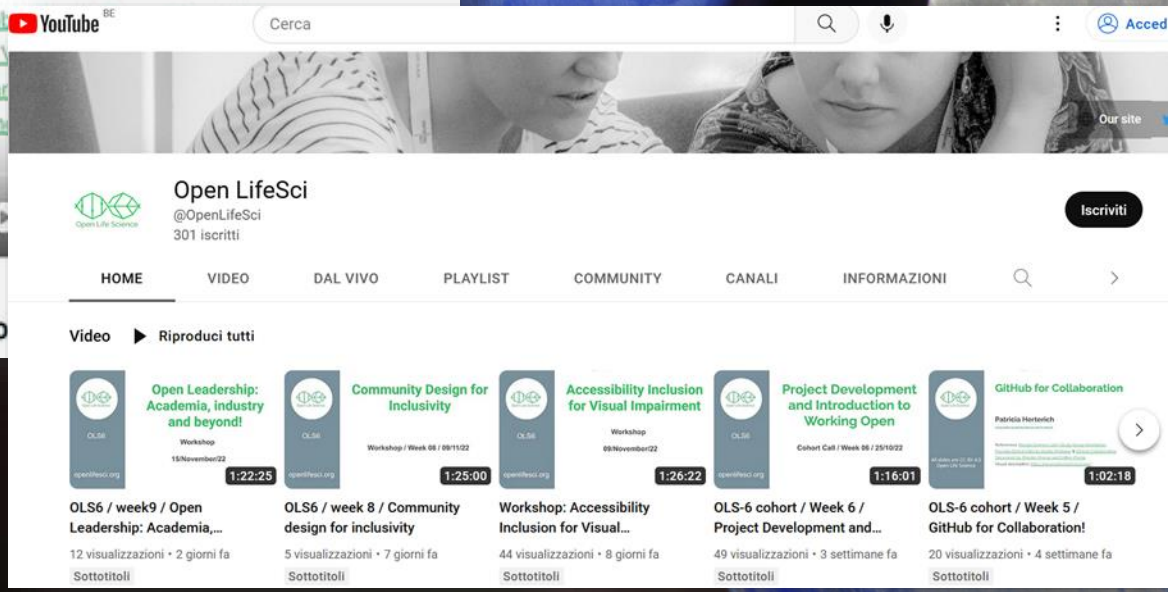
[DMP]



An easy recipe

USE GITHUB (CHECKLIST FOR TASK MANAGEMENT, DATA, TESXTS, INTERACTIONS, VERSION TRACKING…)





https://openrefine.org/

OpenRefine Download Documentation Community Blog Donate

OpenRefine

OpenRefine is a powerful free, open source tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

Download

TO CLEAN DATA

Main features

Faceting

Drill through large datasets using facets and apply operations on filtered views of your dataset.

Clustering

Fix inconsistencies by merging similar values thanks to powerful heuristics.

Reconciliation

Match your dataset to external databases via reconciliation services.

Infinite undo/redo

Rewind to any previous state of your dataset and replay your operation history on a new version of it.

Privacy

Your data is cleaned on your machine, not in some dubious data laundering cloud.

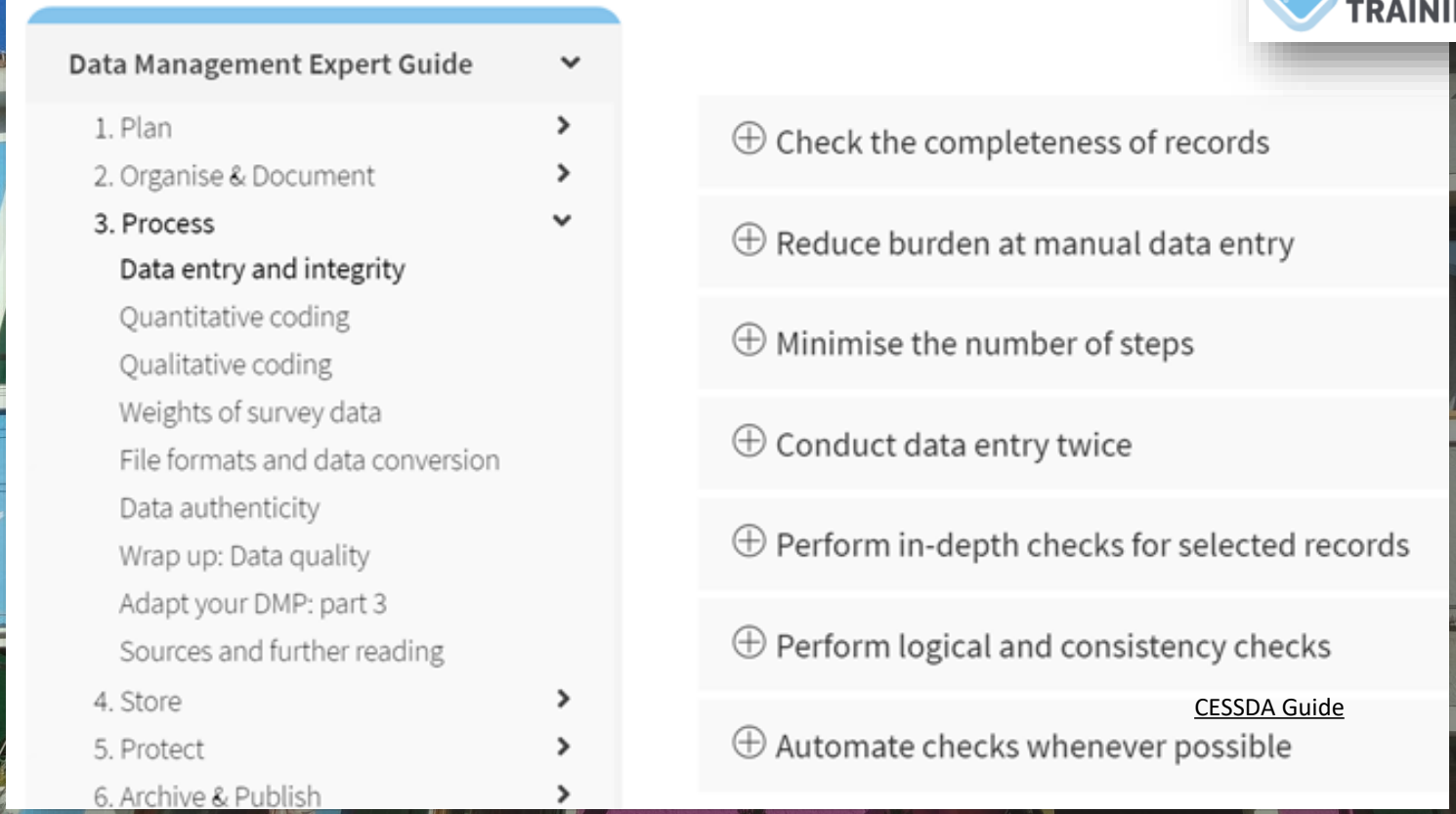
Wikibase

Contribute to Wikidata, the free knowledge base anyone can edit, and other Wikibase instances.

Data management ABC – Data entry

[DMP]

cessda TRAINING

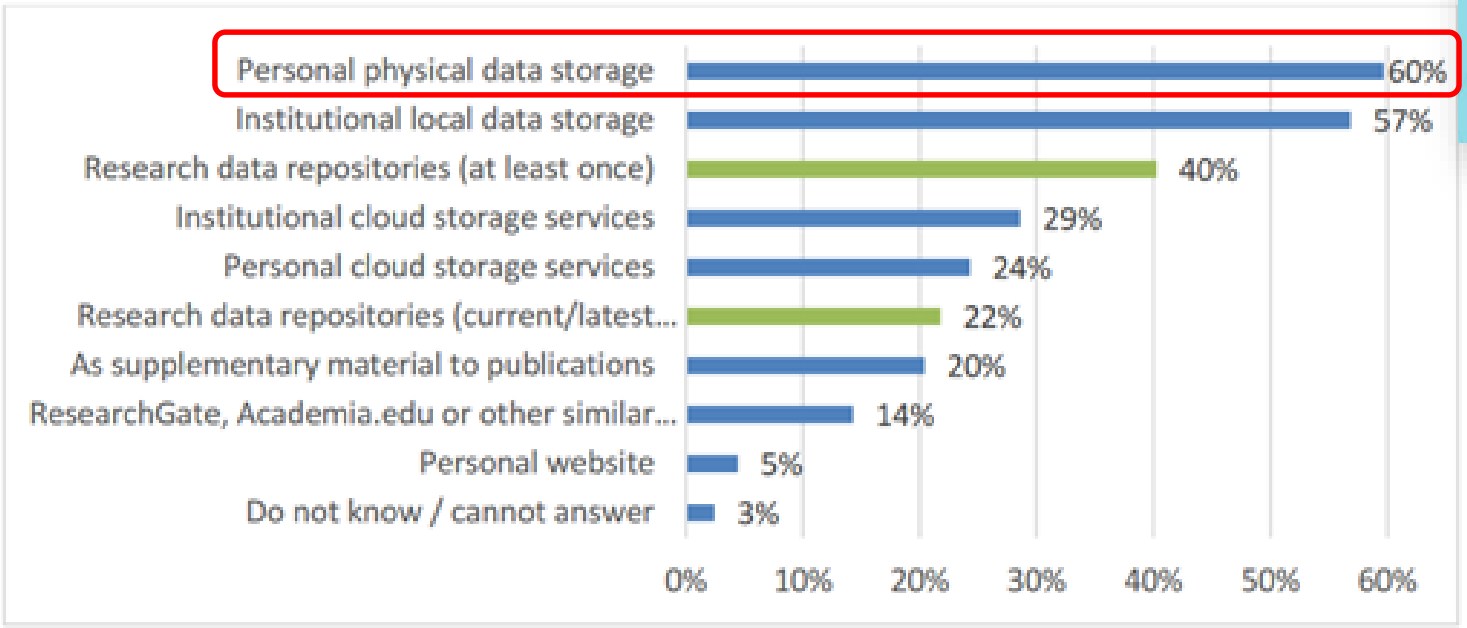


CESSDA Guide

Data management ABC / storage

WHERE DO YOU STORE YOUR DATA?

Figure 12. Locations in which respondents or their research teams stored usable data during their current/most recent research activity



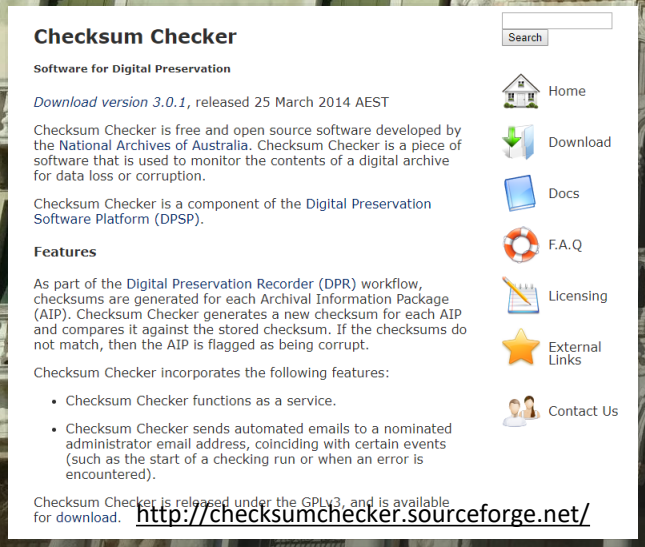
Note: Multiple answers could be selected by a single respondent. Results from the question 'Have you ever stored your research data in a research data repository?' have also been integrated into the figure. Only researchers who did not select research data repositories in the question 'Where have you or your research

Data Management ABC - preservation

[DMP]

LONG TERM OR SHORT TERM?

Storage Solutions	Advantages	Disadvantages	Suitable for
Personal Computer & Laptop	*Always available* *Portable*	*Drive may fail* *Laptop may be stolen*	*Temporary storage*
Networked drives File servers managed by your university, research group or facilities like a NAS-server	*Regularly backed up* *Stored securely in a single place*	*Costs*	*Master copy of your data* *(if enough storage space is provided ..)*
External storage devices USB flash drive, DVD/CD, external hard drive	*Low cost* *Portability*	*Easily damaged or lost*	*Temporary storage*
Cloud services	*Automatic synchronization between folders and files* *Easy to access and use*	*It's not sure whether data security is taken care of* *You don't have direct influence on how often backups take place and by whom*	*Data sharing*



http://checksumchecker.sourceforge.net/

1 2 3 4 5 **6**

Organize and document research data. Make digital versions of paper data documentation in a PDF/A format (suitable for long-term storage).

Data Management ABC- backup and storage

[DMP]

Portable devices | Cloud storage | Local storage | Networked drives

Laptops, tablets, external hard-drives, flash drives and Compact Discs

Advantages	Disadvantages/Risks	Precautions for (sensitive) personal data
• Allow easy transport of data and files without transmitting them over the Internet. This can be especially helpful when working in the field. • Low-cost solution.	• Easily lost, damaged, or stolen and may, therefore, offer an unnecessary security risk. • Not robust for long-term storage or master copies of your data and files. • Possible quality control issues due to version confusion.	Use in encry passw

Advantages	Disadvantages/Risks	Precautions for (sensitive) personal data
• Automatic backups. • Often automatic version control.	• Not all cloud services are secure. May not be suitable for sensitive data containing personal information about EU citizens. • Insufficient control over where the data is stored and how often it is backed up. • Free services by commercial providers (e.g. Google Drive, Dropbox) may claim rights to use content you manage and share them for their own purposes. • Data can be lost if your account is suspended or accidentally deleted, or if the provider goes out of business.	• Encrypt all (sensitive) personal data before uploading it to the cloud. This is particularly important to avoid conflict with European data protection regulations if you do not know in which countries servers used for storage and backup are located (see 'Security' for more information on encryption; also see 'Protecting data').

Recommendations

- Do: use cloud services for granting shared, remote and easy access to data and other files to all involved in the project.
- Do: Read the terms of service. Especially focus on rights to use content given to the service provider.
- Do: Opt for European, national, or institutional cloud services which store data in Europe if possible.
  - B2drop (EUdat, n.d.) is an example of a European cloud storage solution.
  - SWITCHdrive (SWITCH, 2017) is a Swiss solution.
  - DataverseNL (Data Archiving and Networked Services, 2017) is an example of a service for Dutch researchers that allows the storage and sharing of data both during and after the research period.
- Don't: make this your only storage and backup solution.
- Don't: use for unencrypted (sensitive) personal data.

CESSDA Guide

DIFFERENT TOOLS FOR DIFFERENT STEPS OF THE RESEARCH CYCLE.
DURING THE EXPERIMENT YOU ALSO NEED TO COLLABORATE WITH THE TEAM

What to preserve?

DCC because good research needs good data

Home | Digital curation | About us | News | Events | Resources | Training | Projects

Home > Resources > How Guides > How Develop Rdm Services

How to develop

n this section

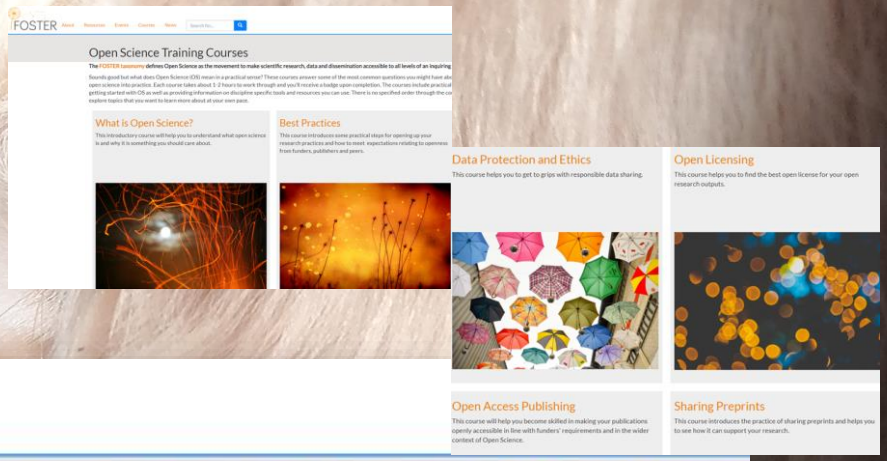
How to Develop RDM Services - a guide for HEIs

Establishing criteria for selection decisions

You should establish criteria to guide selection decisions. The DCC's How to Select and Appraise Research Data for Curation[56] proposes seven criteria as outlined below:

1. **Relevance to mission**: the resource content fits any priorities stated in the institution's mission, or funding body policy including any legal requirement to retain the data beyond its immediate use.
2. **Scientific or historical value**: is the data scientifically, socially, or culturally significant? Assessing this involves inferring anticipated future use, from evidence of current research and educational value.
3. **Uniqueness**: the extent to which the resource is the only or most complete source of the information that can be derived from it, and whether it is at risk of loss if not accepted, or may be preserved elsewhere.
4. **Potential for redistribution**: the reliability, integrity, and usability of the data files may be determined; these are received in formats that meet designated technical criteria; and Intellectual Property or human subjects issues are addressed.
5. **Non-replicability**: It would not be feasible to replicate the data/resource or doing so would not be financially viable.
6. **Economic case**: costs may be estimated for managing and preserving the resource, and are justifiable when assessed against evidence of potential future benefits; funding has been secured where appropriate.
7. **Full documentation**: the information necessary to facilitate future discovery, access, and reuse is comprehensive and correct; including metadata on the resource's provenance and the context of its creation

Learn to manage



FOSTER About Resources Events Courses News Search for...

Managing and Sharing Research Data

Data-driven research is becoming increasingly common in a wide range of academic disciplines, from Archaeology to Zoology, and spanning Arts and Science subject areas alike. To support good research, we need to ensure that researchers have access to good data. Upon completing this course, you will:

- understand which data you can make open and which need to be protected
- know how to go about writing a data management plan
- understand the FAIR principles
- be able to select which data to keep and find an appropriate repository for them
- learn tips on how to get maximum impact from your research data

Start the Free Course

Full details

Level of knowledge: Introductory: no previous knowledge is required

Topics

https://www.fosteropenscience.eu/node/2328

Learn to protect

What are personal data?

Click the plus sign to expand the text box

+ What are personal data?
+ Protecting personal data
+ Legal requirements - EU General Data Protection Regulation (GDPR)
+ Legal requirements - GDPR research exemptions

FOSTER About Resources Events Courses News

Data Protection and Ethics

This course covers data protection in particular and ethics more generally. It will help you understand the basic principles of data protection and introduces techniques for implementing data protection in your research processes. Upon completing this course, you will know:

- what personal data are and how you can protect them
- what to consider when developing consent forms
- how to store your data securely
- how to anonymise your data

Start the Free Course

Full details

Level of knowledge: Introductory: no previous knowle
is required

Topics

[personal data]

GDPR: YOU NEED A VALID LEGAL BASIS TO PROCESS PERSONAL DATA

Legal Basis

Personal data can only be processed when there is a valid legal basis to do so. The GDPR recognises six bases (grounds):

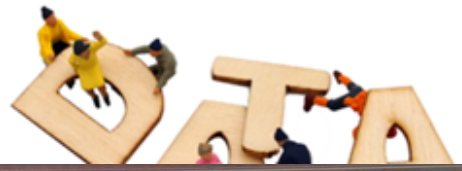
- consent of the data subject
- necessary for the performance of a contract
- legal obligation placed upon the data controller
- necessary to protect the vital interests of the data subject
- carried out in the public interest or in the exercise of official authority (public task)
- legitimate interest pursued by the data controller

The research exemption

The GDPR contains an exemption which entails that some of the principles above are slightly different when you collect and process personal data for research purposes. This is called the 'research exemption'.

> *Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subjected to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner | General Data Protection Regulation, Article 89.*

In practice, this means that Principle II. and V. are less strict. Further processing of personal data for the purposes of archiving, scientific or historical research purposes and statistical purposes is not



cessda TRAINING

CESSDA guide
Data Management Expert Guide

[personal data]

CESSDA guide
Data Management Expert Guide

I. Process lawfully, fair and transparent

The participant is informed of what will be done with the data and data processing should be done accordingly.

II. Keep to the original purpose

Data should be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes.

III. Minimise data size

Personal data that are collected should be adequate, relevant and limited to what is necessary.

IV. Uphold accuracy

Personal data should be accurate and, where necessary kept up to date. Every reasonable step must be taken to ensure that personal data that are inaccurate are erased or rectified without delay.

V. Remove data which are not used

Personal data should be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed.

VI. Ensure data integrity and confidentiality

Personal data are processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss,

[applicable laws]

Science Europe 2018

Privacy

- **Personal Data Protection Acts** are present in all European countries and concern general laws regulating the protection of personal data. They are based on European Directive 95/46/EC.[9] This Directive will be replaced in the near future by the General Data Protection Regulation (GDPR),[10] which all EU Member States will have to implement in their national legislation by May 2018.
- **Obligations to Report Data Leakage Acts** are additions to the Personal Data Protection Acts. They deal with the publication of personal data and contain sanctions in the form of penalties.
- **Medical Treatment Agreement Acts** regulate the use and preservation of personal (patient) data in and for medical research.
- **Scientific Medical Research with Humans Acts** regulate scientific research in the medical field, in particular how to handle personal health-related data. These make ethical reviews compulsory for all medical research projects.

Intellectual Property Rights

- **Copyright Acts** regulate the rights of the creator of a work. One distinguishes between exploitation rights and personal intellectual rights ('moral rights').
- The **Database Rights Act** recognises the investments made in creating and/or compiling a database. It is based on European Directive 96/9/EC.[11]
- **Related Rights Acts** or **Neighbouring Rights Acts** mostly refer to the rights of performers, phonogram producers, and broadcasting organisations.
- **Patent Acts** are for the protection of patents. Publication of research results (including data) is restricted during the application stage of a patent.

Public data

- **Public Records Acts** (Public Archives Acts) oblige all public administration offices and services to preserve their documents and transfer these, after appraisal and selection, to public archives.
- **Public Sector Information Acts** (concerning re-usability of public data) are based on European Directive 2013/37/EU[12] that focuses on the economic aspects of the re-use of public information. It encourages Member States to make as much of this information as possible available for re-use. This also covers content held by museums, libraries, and archives, but does not apply to the educational, scientific, and broadcasting sectors.

- **Freedom of information Acts** regulate and enable citizen access to documents held by public authorities or companies carrying out work for a public authority. They do not specifically deal with access to research data.
- **Heritage Acts** are relevant for archaeological research data in so far as that they regulate ownership of documentation (data) from archaeological excavations.
- **Statistical Information Acts** regulate the competencies of the statistics authorities in data gathering as well in access to data.
- **Land Registry Acts** (cadastral information) regulate the competencies of the national land registries and access to their data, with special provisions concerning personal data contained in their various databases.

Codes of Conduct/Ethical Issues

- **Codes of Conduct**, where these exist on a national level or in an institution, should be taken into account in DMPs. They contain the general principles of good academic teaching and research.
- **Codes of Practice** for the use of personal data in scientific and scholarly research are based on the Personal Data Protection Acts[13] and prescribe how to handle personal data in research practice.
- **Codes of Conduct** for Medical Research regulate how researchers should handle medical personal data. They may be based on Medical Treatment Agreement Acts.

GDPR and research

Research Data Management

Search

HOME | PLANNING RESEARCH | COLLECTING DATA | PROCESSING DATA | ARCHIVING DATA | GDPR IN RESEARCH | SUPPORT & TRAINING

Research Data Management › GDPR in research

GDPR in research

As of May 25 2018, the GDPR (General Data Protection Regulation), or AVG (Algemene Verordening Gegevensbescherming) in Dutch, will apply to the entire European Union. The GDPR has its implications for research. Anyone who collects personal data within Radboud University during their research, must follow 8 guidelines following the Privacy by design principle.

The guidelines are only applicable for research with **personal data**. Personal is any data that can lead to the identification of an individual. For example name, birth date, email-address and IP address are direct personal data. But also a combination of data can lead to the identification of an individual and should therefore be treated as personal data. If you **don't process personal data** in your research, then the GDPR is not applicable. This is for instance the case when your research only includes anonymised data (but be aware that pseudonymised data is personal data).

GDPR in research

Introduction

The GDPR in research, a.o. special categories of personal data, processing in/outside the European Economic Area (EEA), and privacy by design/default.

- **GDPR in research: introduction**
- **FAQ GDPR in research**

Data minimisation

The data minimisation principle comprises that data has to be adequate, relevant and limited to what is necessary for the purposes for which they are processed.

- **GDPR in research: data minimisation**
- **FAQ data minimisation**

Data quality

The data quality principle comprises that data has to be of good quality, i.e. the data has to be accurate and up-to-date.

- **GDPR in research: data quality**
- **FAQ data quality**

Goal setting

In the goal setting, you describe what personal data you process, with which legitimate purpose and for how long.

- **GDPR in research: goal setting**
- **FAQ goal setting**

Minimisation of use

Minimise the processing of and access to personal data, for a pre-defined purpose and period of time, and only by authorised persons.

- **GDPR in research: minimisation of use**
- **FAQ minimisation of use**

Security measures

Make sure that the personal data you collect is well secured. When working with personal data, make use of privacy protection techniques.

- **GDPR in research: security measures**
- **FAQ security measures**

Transparency

The GDPR requires the controller to be transparent to data subjects about the processing of their personal data.

- **GDPR in research: transparency**
- **FAQ transparency**

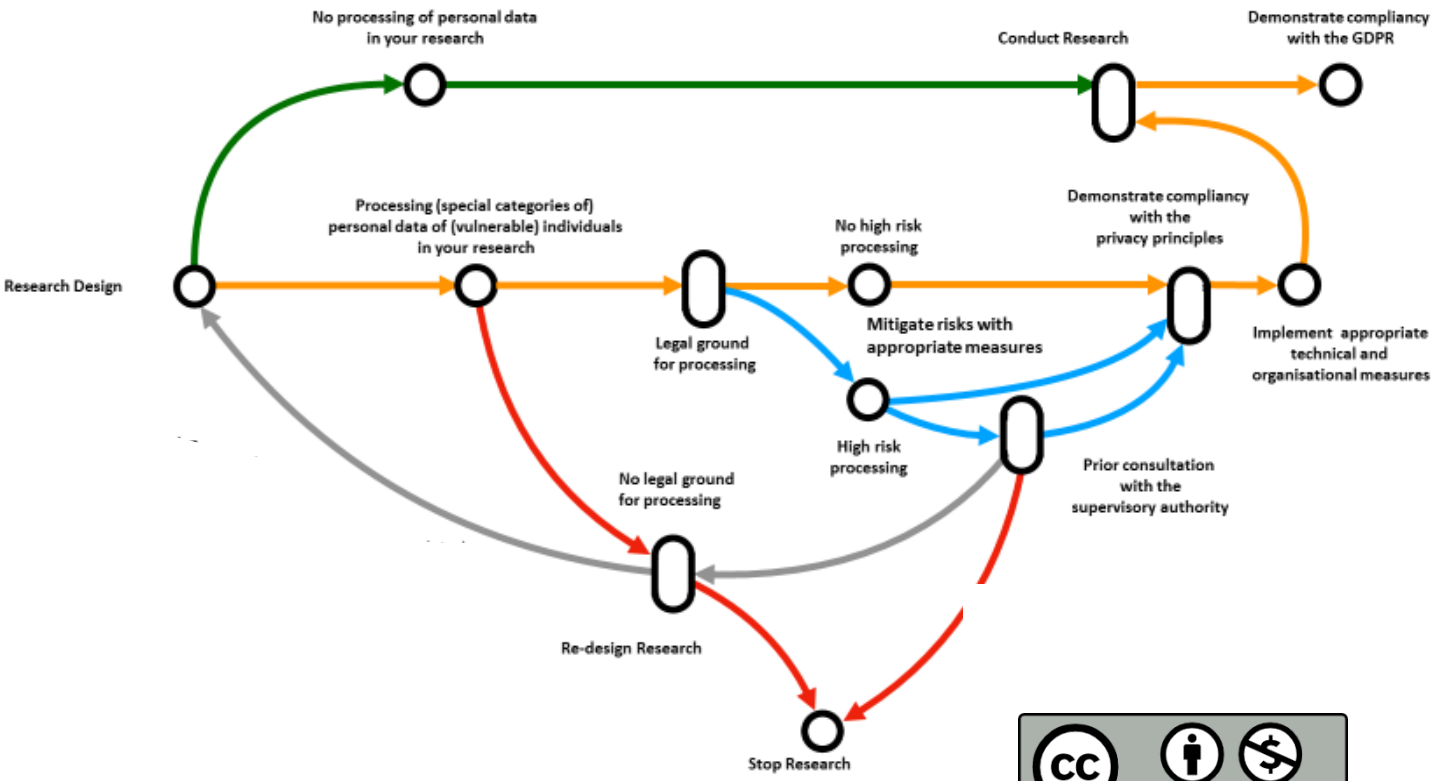
Rights of data subjects

Fundamental of the GDPR are the right of data subjects concerning the processing of their personal data.

- **GDPR in research: rights of data subjects**
- **FAQ rights of data subjects**

[Data and GDPR]

The Privacy Impact Assessment (PIA) Route Planner for Academic Research Inspired by Harry Beck's London Metro Map



Erasmus University Rotterdam
marlon.domingus@eur.nl
February 2018

https://surfdrive.surf.nl/files/index.php/s/BPRxchnZ44NZAgW

The Logic of a Privacy Impact Assessment (PIA) for Academic Research

Q1. Do you process (special categories of) personal data of (vulnerable) individuals in your research?

YES → **Q2. What is the legal ground for this processing?**

NO
Proceed - no measures required for safeguardingprivacy.

Lawfulness of Processing (GDPR*, Article 6, 89):

1. The individuals participating in your research have freely given their explicit consent for one or more specific purposes.
2. Your research contributes to a legitimate interest, yet results in no high risks for the individuals participating in the research.
3. Your research has a scientific, historical or statistical purpose, yet results in no high risks for the individuals participating in the research.

YES → **Q3. Is this processing a high risk processing?**

NO
Stop research or redefine research.

Criteria for high risk processing (WP29 - DPIA Guideline):**

1. Evaluation or scoring
2. Automated-decision making with legal or similar significant effect
3. Systematic monitoring
4. Sensitive data or data of a highly personal nature
5. Data processed on a large scale
6. Matching or combining datasets
7. Data concerning vulnerable data subjects
8. Innovative use or applying new technological or organisational solutions
9. When the processing itself prevents data subjects from exercising a right or using a service or a contract

YES

NO
Proceed - measures required for safe-guarding privacy.

"Personal Data" (GDPR*, Article 4):
Any information relating to an identified or identifiable natural person: a name, an identification number, location data, an online identifier, one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

"Special Categories of Personal Data (Sensitive Data)" (GDPR, Article 9):
Data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation.

Action

Records of processing activities (GDPR*, Article 30):
The university shall maintain a digital record of the processing activities in your research to demonstrate compliancy to the GDPR.
This register contains:

1. The name and contact details of the researcher, the research partners and service providers;
2. The purposes of the processing;
3. A description of the categories of data subjects and of the categories of personal data;
4. The categories of recipients to whom the personal data have been or will be disclosed.

Action

Data protection by design and by default (GDPR*, Article 25):
Implement appropriate technical and organisational measures:

1. **Individual participating in your research (data subject).** Is the participant well informed, aware of possible risks for her/him and aware of the purpose of the research?
2. **Data.** Is the data de-identified and encrypted?
3. **Access Management.** How is access managed and controlled for the PI / team (expanded) / public?
4. **Software / Platform.** Are the *Terms of Service* for used software / platform checked (where is the data and who has access and has which usage rights)?
5. **Devices.** Are devices used safe? Encrypted drive, encrypted communication, strong password / two factor authentication.
6. **Partners.** Are the research partners / service partners trusted and are appropriate legal agreements made, with regards to roles, rights and responsibilities?
7. **Safe and secure collaboration.** Is the ((cross border) communication to, in and from the) collaboration platform end to end encrypted, are roles and permissions defined and implemented, is logging and monitoring implemented?
8. **Risk definition and mitigation**. Are risks defined and mitigated? Is a risk audit procedure started?

Action

Prior consultation (GDPR*, Article 36):

1. The Data Protection Officer shall, on behalf of the researcher, consult the supervisory authority, prior to the processing (the research) when the processing would result in a high risk *in the absence of measures* to mitigate the risk.

Action

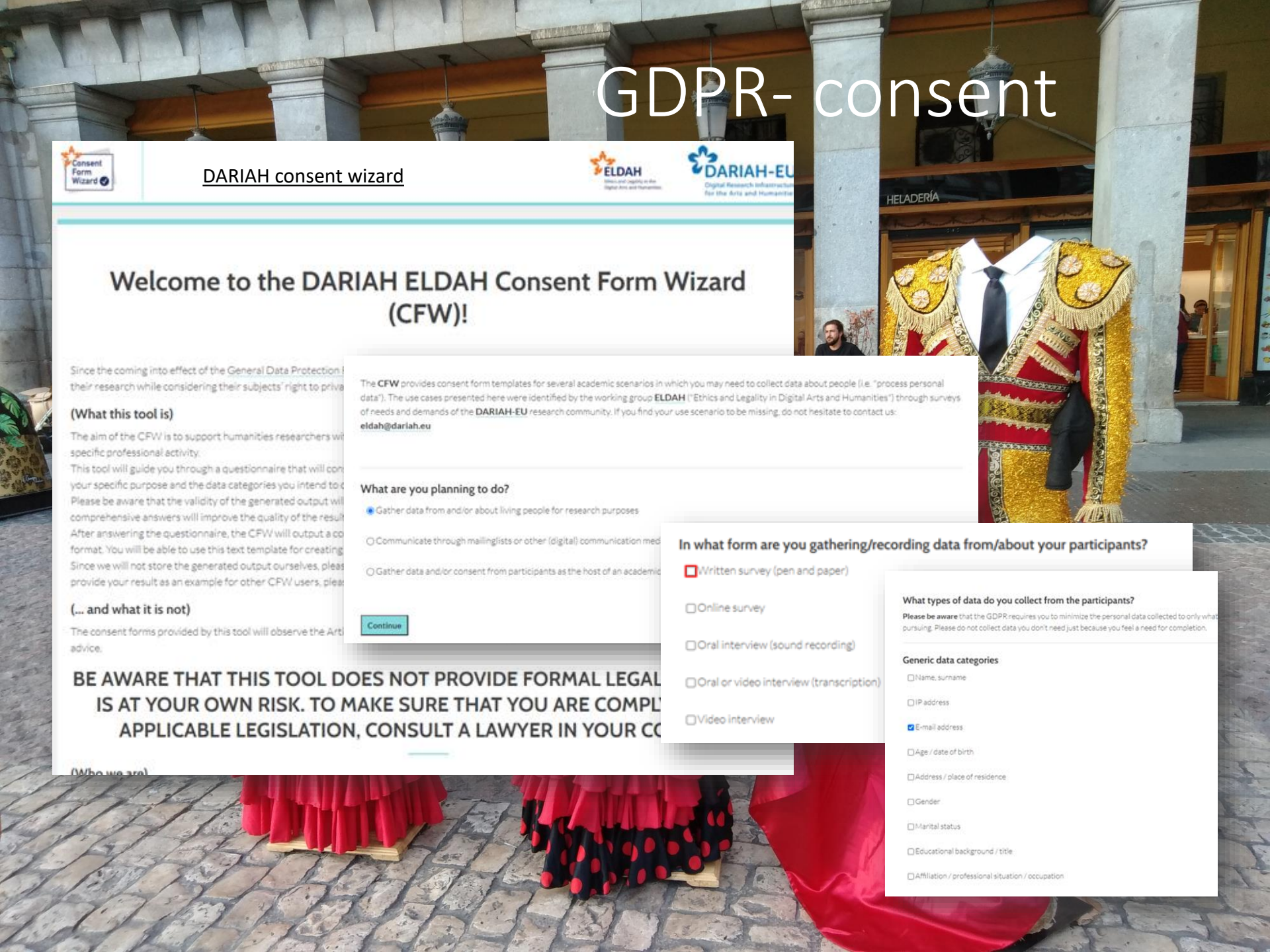
Principles relating to processing of personal data (GDPR*, Article 5):
Demonstrate compliancy with the principles: lawfulness, fairness, transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity, confidentiality and accountability.

* Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Online available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

** Article 29 Data Protection Working Party: *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679*. Adopted on 4 April 2017. As last Revised and Adopted on 4 October 2017. Online available at: https://ec.europa.eu/newsroom/document.cfm?doc_id=47711

GDPR- consent

DARIAH consent wizard



[anonimyzing]

Anonymisation

UK Data service

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.

A person's identity can be disclosed from:

- **Direct identifiers** such as names, postcode information or pictures
- **Indirect identifiers** which, when linked with other available information, could ide someone, for example information on workplace, occupation, salary or age

You decide which information to keep for data to be useful and which to change. Rem key variables, applying pseudonyms, generalising and removing contextual informatic from textual files, and blurring image or video data could result in important details be missed or incorrect inferences being made. See example 1 and example 2 for balanci anonymisation with keeping data useful for qualitative and quantitative data.

Anonymising research data is best planned early in the research to help reduce anonymisation costs, and should be considered alongside obtaining informed consen data sharing or imposing access restrictions. Personal data should never be disclose research information, unless a participant has given consent to do so, ideally in writin

Quantitative data | Qualitative data | Step-by-step

Anonymising **quantitative data** may involve removing or aggregating variables or reducing the precision or detailed textual meaning of a variable.

Primary anonymisation techniques

- **Remove direct identifiers** from a dataset. Such identifiers are often not necessary for secondary research.

  Example: Remove respondents' names or replace with a code; remove addresses, postcode information, institution and telephone numbers.

- **Aggregate or reduce the precision** of a variable such as age or place of residence. As a general rule, report the lowest level of geo-referencing that will not potentially breach respondent confidentiality. The exact scal of data collected, but very detailed geo-references like full p small towns or villages are likely to be problematic. Coded c which may be potentially revealing can be aggregated into l aggregation of a disclosive variable is not possible, conside removed from the dataset.

  Example: Record the year of birth rather than the day record postcode sectors (first 3 or 4 digits) rather tha aggregate detailed 'unit group' standard occupational employment codes up to 'minor group' codes by rem

- **Generalise the meaning** of a detailed text variable by repla disclosive free-text responses with more general text.

  Example: Detailed areas of medical expertise could i doctor. The expertise variable could be replaced by m coded into generic responses such as 'one area of m more areas of medical speciality', etc.

- **Anonymise relational data** where relations between variables in related or linked datasets or in combination with other publicly available outputs may disclose identities.

  Example: In confidential interviews on farms the names of farmers have been replaced with codes and other confidential information on the nature of the farm businesses and their locations have been disguised to anonymise the data.

  However, if related biodiversity data collected on the same farms, using the same farmer codes, contain detailed locations for biodiversity data alone the location would not be confidential. Farmers could be identified by combining the two datasets.

  The link between farmer codes and biodiversity location data should be removed, for example by using separate codes for farmer interviews and for farm locations.

- **Anonymise geo-referenced data** by replacing point coordinates with non-disclosing features or variables; or, preferably, keep geo-references intact and impose access restrictions on the data instead.

  Point data may fix the position of individuals, organisations or businesses studied, which could disclose their identity. Point coordinates may be replaced by larger, non-disclosing geographical areas such as polygon features (km^2 grid, postcode district, county), or linear features (random line, road, river). Point data can also be replaced by meaningful alternative variables that typify the geographical position and represent the reason why the locality was selected for the research, such as poverty index, population density, altitude, vegetation type. In this way, the value of data is maintained

[anonimyzing]

Amnesia OpenAIRE

High accuracy Data Anonymization.

Perform research and share your results that satisfy GDPR guidelines by using data anonymization algorithms.

GET STARTED

Unlock sensitive data analysis

Use Amnesia to transform personal data to anonymous data that can be used for statistical analysis. Data anonymized with Amnesia are *statistically guaranteed* that they cannot be linked to the original data.

- Guarantees no links to the original data
- Offers k-anonymity & km-anonymity
- Allows minimal reduction of information quality

Become GDPR compliant

Create anonymous datasets from personal data that are treated as statistics by GDPR. Anonymous data can be used without the need for consent or other GDPR restrictions, greatly reducing the effort needed to extract value from them.

- Guarantees anonymity
- Goes beyond pseudo-anonymization
- Anonymized data are not constrained by GDPR

High Usability & Flexibility

Anonymization tailored to user needs through a graphical interface. Guide the algorithm and decide trade-offs with simple visual choices. Developers can incorporate Amnesia anonymization engine to their project through a ReST API.

- Easy usage interface
- Adjustable settings
- Visualization of anonymization choices

How it works

Get anonymous data in 3 steps

1 **Insert your data**

Amnesia accepts complex object relational data in delimited text files.

2 **Select and Preview the data to anonymize**

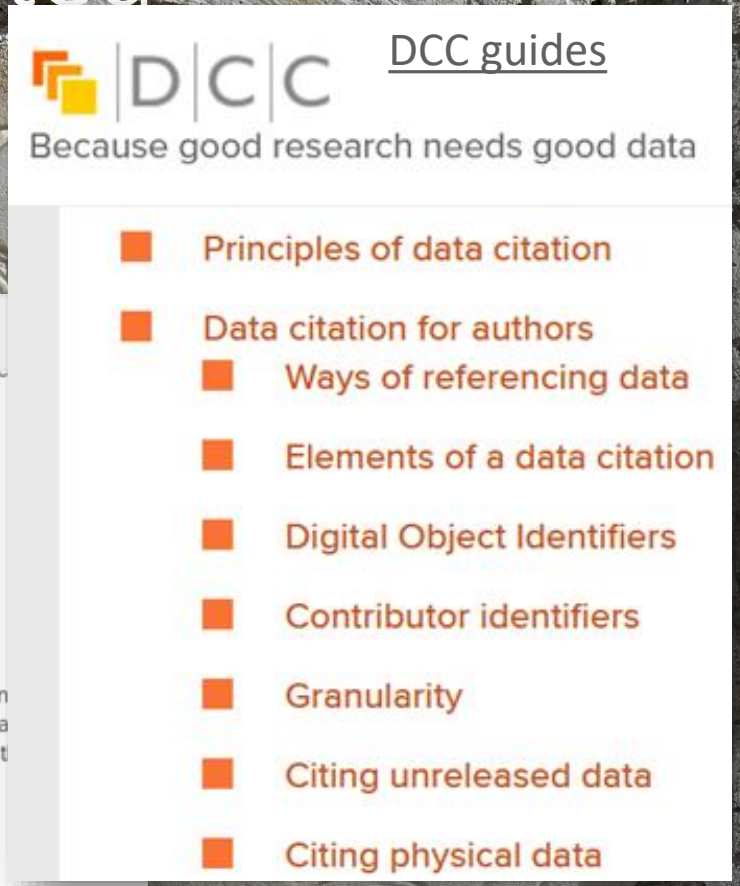
Visual representations of anonymization parameters and results allow non-expert users to tailor the anonymization process to their needs.

3 **Download your data anonymized**

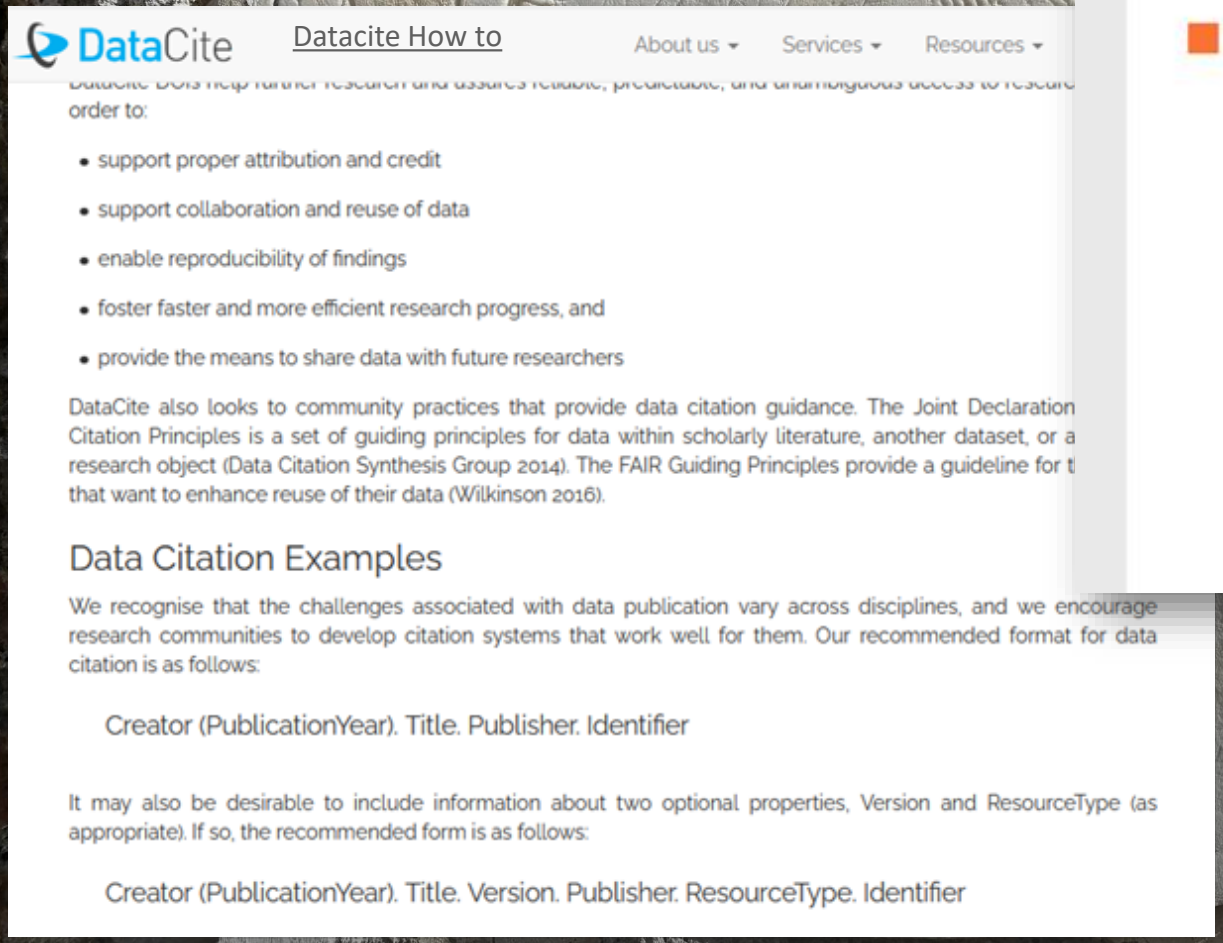
The process is completed without any sensitive data leaving your premises!

…data need to be cited

DCC guides



Datacite How to



…wrapping up

10 Simple Rules for the Care and Feeding of Scientific Data

https://arxiv.org/pdf/1401.2134v1.pdf

Alyssa Goodman[1], Alberto Pepe[1,*], Alexander W. Blocker[4], Christine L. Borgman[2], Kyle Cranmer[3], Merce Crosas[4], Rosanne Di Stefano[1], Yolanda Gil[5], Paul Groth[6], Margaret Hedstrom[7], David W. Hogg[3], Vinay Kashyap[1], Ashish Mahabal[8], Aneta Siemiginowska[1], Aleksandra Slavkovic[9]

Rule 1. Love your data, and help others love it too.

Data management is a repeat-play game. If you take care to make your data easily available to others, others are more likely to do the same–eventually. While we wait for this new sharing-equilibrium to be reached, you can take two important actions. First, cherish, document, and **publish your data**, preferably using the robust methods described in Rule 2. Get started now, as: better tools and resources for data management are becoming more numerous; universities and research communities are moving toward bigger investments in data repositories (Rule 8); and more librarians and scientists are learning data management skills (Rule 10). At the very least, loving your own data available will serve *you*: you'll be able to find and reuse your own data if you treat them well. Second, enable and **encourage others to cherish, document, and publish their data**. If you are a research scientist, chances are that not only are you an author, but also a reviewer for a specialized journal or conference venue. As a reviewer, **request that the authors of papers you review provide documentation and access to their data** according to the rules set out in the remainder of this article. While institutional approaches are clearly essential (Rules 8 and 10), changing minds one scientist at a time is effective as well.

2. DATA FAIR BY DESIGN

FAIR train – GoFAIR video

…FAIR means [for machines]


SCIENTIFIC DATA

We'd like to understand how you use our websites in order to imp

Open Access | Published

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...] Barend Mons

FAIR guide, Nature, March 2016

FINDABLE

- IDENTIFIERS
- METADATA

ACCESSIBLE

- WHERE TO FIND THE DATA AND UNDER WHAT ACCESS CONDITIONS
- **NOT «OPEN»**
- OPEN FORMATS

INTEROPERABLE

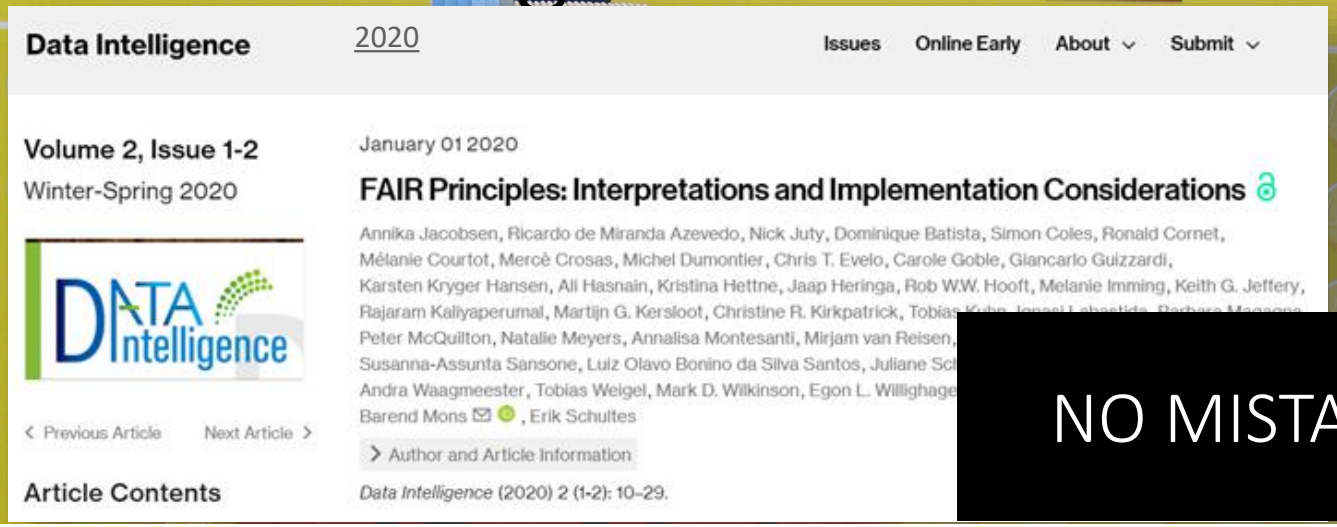
- STANDARDS
- ONTOLOGIES

REUSABLE

- LICENSES
- DOCUMENTATION

MACHINE-READABLE

…before starting for FAIR



Data Intelligence 2020 Issues Online Early About Submit

Volume 2, Issue 1-2
Winter-Spring 2020

DATA Intelligence

< Previous Article Next Article >

Article Contents

January 01 2020

FAIR Principles: Interpretations and Implementation Considerations

Annika Jacobsen, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, Mercè Crosas, Michel Dumontier, Chris T. Evelo, Carole Goble, Giancarlo Guizzardi, Karsten Kryger Hansen, Ali Hasnain, Kristina Hettne, Jaap Heringa, Rob W.W. Hooft, Melanie Imming, Keith G. Jeffery, Rajaram Kaliyaperumal, Martijn G. Kersloot, Christine R. Kirkpatrick, Tobias Kuhn, Ignasi Labastida, Barbara Magagna, Peter McQuilton, Natalie Meyers, Annalisa Montesanti, Mirjam van Reisen, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Juliane Sch… Andra Waagmeester, Tobias Weigel, Mark D. Wilkinson, Egon L. Willighage… Barend Mons, Erik Schultes

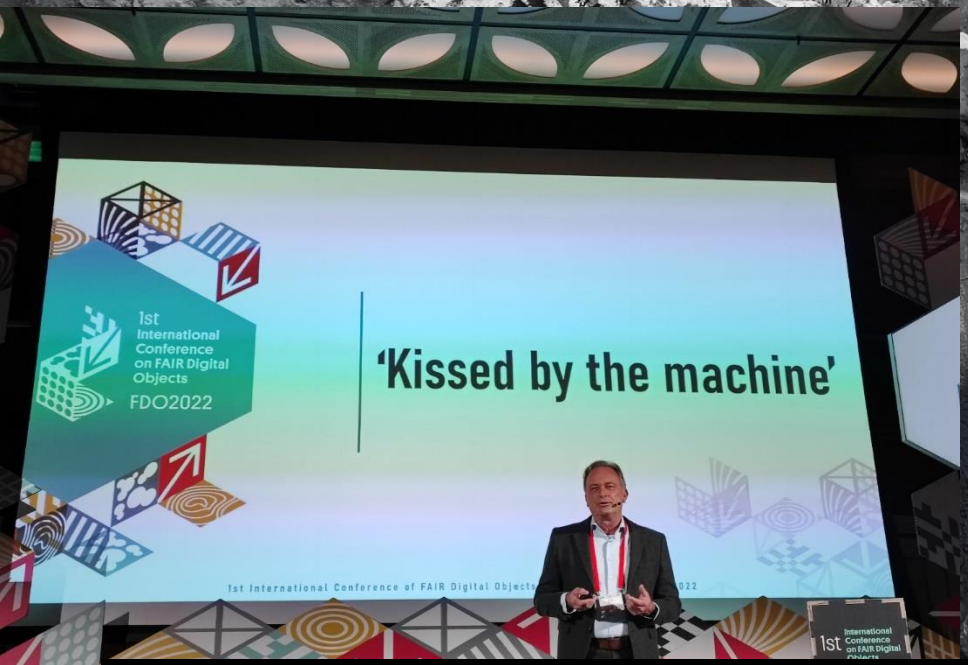
> Author and Article Information

Data Intelligence (2020) 2 (1-2): 10–29.

NO MISTAKES!

- Findability: Digital resources should be easy to find for both humans and computers. Extensive machine-actionable metadata are essential for automatic discovery of relevant datasets and services, and are therefore an essential component of the FAIRification process [14].
- Accessibility: Protocols for retrieving digital resources should be made explicit, for both humans and machines, including well-defined mechanisms to obtain authorization for access to protected data.
- Interoperability: When two or more digital resources are related to the same topic or entity, it should be possible for machines to merge the information into a richer, unified view of that entity. Similarly, when a digital entity is capable of being processed by an online service, a machine should be capable of automatically detecting this compliance and facilitating the interaction between the data and that tool. This requires that the meaning (semantics) of each participating resource – be they data and/or services service – is clear.
- Reusability: Digital resources are sufficiently well described for both humans and computers, such that a machine is capable of deciding: if a digital resource *should* be reused (i.e., is it relevant to the task at-hand?); if a digital resource *can* be reused, and under what conditions (i.e., do I fulfill the conditions of reuse?); and *who to credit* if it is reused.

Kissed or missed?



FAIR PRINCIPLES ARE
«MACHINE ACTIONABLE»
(MORE THAN READABLE)
FAIR = FULLY AI READY
IF NOT… YOU'LL BE MISSED (INSTEAD OF KISSED) BY THE MACHINE

RDA

Decision making procedures in data management and data stewardship for Open Science

Connie Clare, PhD

RDA

Data-centric AI

Automated decision making using data.

Data is fundamental for training and deploying AI models.

Data management and/or curation is a crucial step to feed into AI model.

'Machine learning models are only as good as the data they're trained on' - https://fairmlbook.org/datasets.html (Chapter 8)

Clearbox AI

Clearbox

We are on a mission to harness powerful AI technologies t improve businesses and society in a trustworthy and human-centered way.

s flexible product / Rea

clearbox AI

Your

Synthetic Data

provider

RDA

Data stewardship challenges & AI ethics

Black box AI - Model inputs and operations remain a mystery. Unknown input data provenance and quality. Automated data retrieval lead to inconsistent results.

AI bias due to generalisation (insufficient representative input data), or unsuitable data collection, processing (cleaning), quality, mislabelling and model design. Synthetic (output) data generated inherits and propagates bias affecting scientific validity.

Data misuse - Using data as input for an AI model that causes harm.

Lack of standards, tools and mechanisms to evaluate data quality ar whether datasets are fit for purpose.

ARTIFICIAL INTELLIGENCE

- WORKS IF DATA ARE GOOD
- THERE ARE ETHICAL ISSUES

Scenario

Volume 2, Issue 1-2
Winter-Spring 2020
2020

DATA Intelligence

January 01 2020

The Need of Industry to Go FAIR

Herman van Vlijmen , Albert Mons , Arne Waalkens , Wo
Christine Kirkpatrick , Luiz Olavo Bonino da Silva Santos , Ber
Sebastiaan Knijnenburg , Scott Lusher , Rudi Verbeeck , Jean

Author and Article Information

Data Intelligence (2020) 2 (1-2): 276–284.

https://doi.org/10.1162/dint_a_00050

Previous Article Next Article

2. THE VALUE OF FAIR DATA

Research data is one of the most valuable resources we have in the world, as it is the key ingredient to innovation, ultimately leading to societal benefits, like alternative energy options, or treatments of diseases. Every element of data could potentially contain a clue that can lead to an important discovery. However, in industry, much like in academia, research data is rarely leveraged beyond its original intended purpose [2]. This is not only based on deliberate data protection, but also on a lack of findability. That means that making data FAIR in industry, and ensuring interoperability and reusability presents a huge opportunity for industry, but ultimately also for society as a whole.

- RESEARCH DATA IS RARELY LEVERAGED BEYOND ITS ORIGINAL PURPOSE
- OFTEN FOR A LACK OF FINDABILITY
- EVERY ELEMENT HAS POTENTIAL INNOVATIVE CLUES

FAIR principles

FAIR principles

«ACCESSIBLE» DOES NOT MEAN «OPEN».

DATA CAN BE CLOSED, PROVIDED YOU – AND MACHINES - KNOW WHERE TO FIND THEM AND UNDER WHAT ACCESS CONDITIONS

RECOMMENDATIONS

» Clarify all legal issues at the beginning of your research project and include the findings of this process in the data management plan.

» Use checklists adequate to your research topic/discipline.

» Check the resources indicated by DARIAH, CLARIN. (see further reading).

» In the case of personal data ensure that only relevant people can access the data and that these are clearly identified (see GDPR).

» Ask for consent to share anonymised data and establish transparent and well-documented anonymisation routines that consider not just direct identifiers, but also how a combination of indirect identifiers could reveal identities. (See for example the guide on informed consent in the CESSDA data management expert guide).

» Avoid collection of (sensitive and non-sensitive) personal data when possible.

» Get legal support (IPR, copyright, patents, trademarks etc.) from your home institution. If there is no dedicated office for this purpose, try to get information from your university library, as its staff are often confronted with such issues.

» If you need permission from the copyright holder in order to use sources like images for your publication, try to get one that covers both printed and digital copies.

» Finally, check the recommendations in the section on Licences, that are closely related to this section.

RECOMMENDATIONS

» To ensure the best possible stewardship of your data, choose to deposit it in a digital repository that is certified by a recognised standard such as the CoreTrustSeal. The Registry of Research Data Repositories (re3data) provides a good starting point, noting disciplines, standards, content types, certification status and more. FAIRsharing (manually curated information on standards, databases, policies and collections) allows you to search databases by subject, and includes entries tagged 'Humanities and Social Sciences'.

» Use disciplinary repositories where they exist, as they are more likely to be developed around domain expertise, disciplinary practices and community-based standards, which will promote the findability, accessibility, interoperability and ultimately the reuse and value of your data. The level of curation available in a repository is key to data quality and reusability.

» Datasets should be assigned persistent identifiers (PID). Most repositories that are designed for long-term preservation will automatically assign or 'mint' persistent identifiers for your datasets, so choosing a quality repository will automate this step. Consider as well signing up for ORCID, a free service that assigns persistent identifiers to individuals/authors.

» To facilitate findability of all research outputs, bidirectional links should be created between publications related outputs, such as data (using PIDs).

» Include the richest metadata possible with your deposited data so that others can find it, understand the parameters under which it was created, and understand the conditions under which they can access and/or reuse it. See recommendations in this report in the sections on Licences and Metadata for more information.

RECOMMENDATIONS

» If applicable, determine if the body funding your research has particular requirements for a DMP, or offers a template for framing your plan. If there is no required template, choose an existing appropriate one (e.g. via DMPOnline).

» Devise a DMP prior to collecting data. Define and plan for your data: all research projects deal with data. If your project includes the analysis of text corpora, for example, then the corpora themselves are data, and you should make sure they are clearly described, documented, and managed according to the FAIR principles so your research is reusable by others.

» Plan documentation of metadata: In order for your data to be comprehensible in the future and/or reusable by others, they will need descriptive metadata created according to a common schema to understand the context/purpose of the research. The richer the metadata, the more intelligible and useful the dataset (see section on Metadata).

» Use standardised terminology to increase interoperability. Consider employing vocabularies or ontologies that follow FAIR principles to increase interoperability and findability (e.g. see FAIRsharing.org).

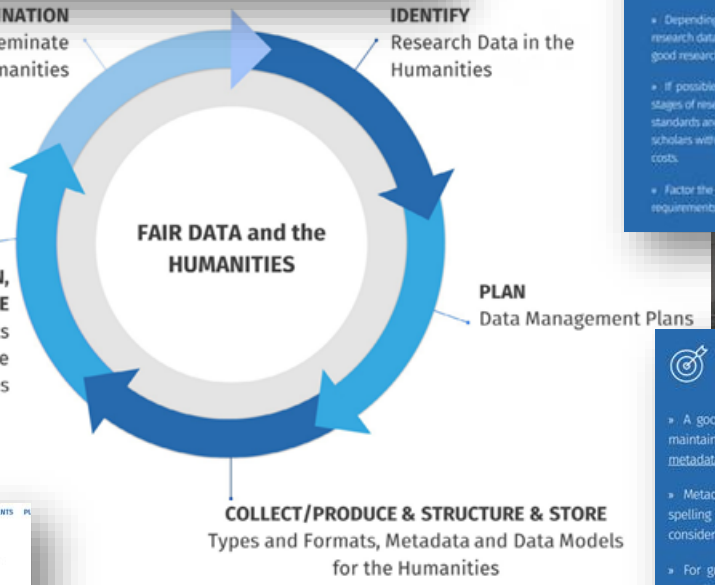
» Consider the right questions to be answered in your DMP that can account for discipline-specific requirements. The DMP templates suggested by funders are quite high level and provide generic guidance for file naming or versioning conventions, database structuring, and can be a good start. Tools like the dmponline.dcc.ac.uk provide discipline specific examples that can be of further reference.

» DMP as living documents: Update your data management plan regularly in order to take into account any potential relevant changes such as using new data types and/or models, technology, new institutional data management policies, reassessing legal aspects or licences for legal compliance etc.

» Depending on the size of the organisation: think of providing institutional support for research data management (RDM); organise information sessions to raise awareness about good research data management, and the risks of not managing it early.

» If possible, consider involving library and/or repository support staff from the initial stages of research data management planning to discuss the best solutions, specifications, standards and protocols along which the repository operates. Repository staff can also assist scholars with understanding any specific data management requirements and associated costs.

» Factor the cost of research data management (time or human resources) into budgetary requirements at the point of application.



RECOMMENDATIONS

» Data models go FAIR: the FAIR Guiding Principles, correctly applied, ensure data are findable, accessible, interoperable and reusable. Data modelling should take this into account by using formal, easily accessible languages for knowledge representation, providing persistent identifiers, open standards, well documented Application Programming Interfaces (API), generic user interfaces and rich metadata. The FAIRification process developed by the GO FAIR initiative offers a system on how to shape the data modelling.

» Use open standards, and whenever possible, standardised technologies and procedures should be used. The World Wide Web Consortium W3C maintains several standards relevant for data models like XML and RDF. Within XML the Text or Music Encoding Initiative TEI/MEI or specific expressions of them have become standards for text or music editions. The query language SPARQL and the representation tool for linked data JSON-LD are common standards for RDF (refers to FAIR principle I1).

» Prefer human and machine-readable systems: coding of data models and of the actual data that is both human and machine-readable in a unified way provides better sustainability and long-term accessibility than machine-readable only code (binary codes), that may use different formats for data model description and the actual data. For both, hierarchical data models and graph-based data, various serialisations (file formats) are available that fulfil this condition (XML, TEI/XML, Turtle, N3, RDF/XML), whereas SQL based technologies need bigger efforts.

» Normalise as much as possible: to avoid redundant information, the content of databases should be normalised as far as possible, using for example authority files like VIAF and identifiers like DOI, ARK, ISNI, GND and the like. To foster the exchange of data, standardised vocabularies and ontologies are needed as well, but an overall ontology for the humanities has not yet been established. The ontology CIDOC-CRM and especially some extensions are well on their way to become a reference model for cultural heritage information.

» Data models follow the data management plan (DMP): when establishing a data model, researchers should keep the whole lifecycle of their data in mind, as it should be outlined in a DMP. Therefore, an extensive documentation of the data model, its software and tools are highly relevant and facilitates the transfer of data in a secure and trusted repository in order to keep them accessible. The same is true here: the more you use open standards for your data model, the easier this task becomes.

RECOMMENDATIONS

» A good starting point is to consult the Metadata Standards Directory, a community-maintained directory hosted by the Research Data Alliance: https://rd-alliance.github.io/metadata-directory/

» Metadata works best when terminology is consistent, e.g. naming conventions are followed, spelling is normalised, and so on. Depending on the complexity and size of your metadata, consider using a tool such as Open Refine to 'clean' your metadata.

» For greater searchability and interoperability, researchers should also consider using controlled vocabularies to identify common terminology when populating metadata fields. ...rary of Congress maintains a controlled vocabulary for subject headings: ...thorities/subjects.html

...e metadata should include a clear and explicit reference to the dataset ...h the inclusion of a PID in the metadata (see section on Trustworthy Data ...rsistent Identifiers).

...data as rich as possible in order to better contextualise your data and ...sider more detailed descriptions, and fuller provenance information, as ...ectrum of available metadata fields.

...be machine-readable.

Sustainable and FAIR Data Sharing in the Humanities

ALLEA Report | **February 2020**

February 2020

FAIR research software

16 Mar 2022

FAIR Principles for Research Software (FAIR4RS Principles)

By Neil Chue Hong

2022

FAIR for Research Software (FAIR4RS) WG

Group co-chairs: Michelle Barker, Paula Andrea Martinez, Leyla Garcia, Daniel S. Katz, Neil Chue Hong, Jennifer Harrow, Fotis Psomopoulos, Carlos Martinez-Ortiz, Morane Gruenpeter

The FAIR4RS Principles are:

F: Software, and its associated metadata, is easy for both humans and machines to find.
F1. Software is assigned a globally unique and persistent identifier. • F1.1. Components of the software representing levels of granularity are assigned distinct identifiers. • F1.2. Different versions of the software are assigned distinct identifiers. F2. Software is described with rich metadata. F3. Metadata clearly and explicitly include the identifier of the software they describe. F4. Metadata are FAIR, searchable and indexable.
A: Software, and its metadata, is retrievable via standardized protocols.
A1. Software is retrievable by its identifier using a standardized communications protocol. • A1.1. The protocol is open, free, and universally implementable. • A1.2. The protocol allows for an authentication and authorization procedure, where necessary. A2. Metadata are accessible, even when the software is no longer available.
I: Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.
I1. Software reads, writes and exchanges data in a way that meets domain-relevant community standards. I2. Software includes qualified references to other objects.
R: Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software).
R1. Software is described with a plurality of accurate and relevant attributes. • R1.1. Software is given a clear and accessible license. • R1.2. Software is associated with detailed provenance. R2. Software includes qualified references to other software. R3. Software meets domain-relevant community standards.

Table 1: The FAIR Principles for Research Software

FAIR RESEARCH SOFTWARE

FAIR software

Practical guide to Software Management Plans 2022

Core requirement (Section 5.1)	Example SMP question(s) (Section 6.1)
Purpose	Please provide a brief description of your software, stating its purpose and intended audience.
Version control	How will you manage versioning of your software?
Repository	How will you make your software publicly available? If you do not plan to make it publicly available, you should provide a justification.
User documentation	How will your software be documented for users? Please provide a link to the documentation if available. How will you document your software's contribution guidelines and governance structure?
Software licencing and compatibility	What type of licence will your software have? How will you check that it respects the licences of libraries and dependencies it uses?
Deployment documentation	How will the installation requirements of your software be documented? Please provide a link to the installation documentation if available. This documentation should include a complete and unambiguous description of dependencies to other software, datasets, and hardware.
Citation	How will users of your software be able to cite your software? Please provide a link to your software citation file (CFF) if available.
Developer documentation	How will your software be documented for future developers?
Testing	How will your software be tested? Please provide

Core requirement (Section 5.1)	Example SMP question(s) (Section 6.1)
Purpose	Please provide a brief description of your software, stating its purpose and intended audience.
Version control	How will you manage versioning of your software?
Repository	How will you make your software publicly available? If you do not plan to make it publicly available, you should provide a justification.
User documentation	How will your software be documented for users? Please provide a link to the documentation if available. How will you document your software's contribution guidelines and governance structure?
Software licencing and compatibility	What type of licence will your software have? How will you check that it respects the licences of libraries and dependencies it uses?
Deployment documentation	How will the installation requirements of your software be documented? Please provide a link to the installation documentation if available. This documentation should include a complete and unambiguous description of dependencies to other software, datasets, and hardware.
Citation	How will users of your software be able to cite your software? Please provide a link to your software citation file (CFF) if available.
Developer documentation	How will your software be documented for future developers?
Testing	How will your software be tested? Please provide a link to the (automated) testing results.
Software Engineering quality	Do you follow specific software quality standards? If yes, which ones?
Packaging	How will your software be packaged and distributed? Please provide a link to available packaging information (e.g. entry in a package registry, if available)
Maintenance	How do you plan to procure long term maintenance of your software?

Core requirement (Section 5.1)	Example SMP question(s) (Section 6.1)
Purpose	Please provide a brief description of your software, stating its purpose and intended audience.
Version control	How will you manage versioning of your software?
Repository	How will you make your software publicly available? If you do not plan to make it publicly available, you should provide a justification.
User documentation	How will your software be documented for users? Please provide a link to the documentation if available. How will you document your software's contribution guidelines and governance structure?
Software licencing and compatibility	What licence will you give your software? How will you check that it respects the licences and dependencies it uses?
Deployment documentation	How will the installation requirements of your software be documented? Please provide a link to the installation documentation if available.
Citation	How will users of your software be able to cite your software? Please provide a link to your software citation file (CFF) if available.
Developer documentation	How will your software be documented for developers?
Testing	How will your software be tested? Please provide a link to the (automated) testing results.
Software Engineering quality	Do you follow specific software quality standards? If yes, which ones?
Packaging	How will your software be packaged and distributed? Please provide a link to available packaging information (e.g. entry in a package registry, if available)
Maintenance	How do you plan to procure long term maintenance of your software?

Core requirement (Section 5.1)	Example SMP question(s) (Section 6.1)
Purpose	Please provide a brief description of your software, stating its purpose and intended audience.
Version control	How will you manage versioning of your software?
User documentation	How will your software be documented for users? Please provide a link to the documentation if available. How will you document your software's contribution guidelines and governance structure?
Deployment documentation	How will the installation requirements of your software be documented? Please provide a link to the installation documentation if available. This documentation should include a complete and unambiguous description of dependencies to other software, datasets, and hardware.
Software licencing and compatibility	What licence will your software have? How will you check that it respects the licences of libraries and dependencies it uses?

Table 1. Core requirements of an SMP and example SMP questions for a low level of software management.

…YOU NEED SOFTWARE MANAGEMENT PLANS!

Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud

2017

Article type: Research Article

Authors: Mons, Barend[a; b; c; *] | Neylon, Cameron[d] | Velterop, Jan[e] | Dumontier, Michel[f] | da Silva Santos, Luiz Olavo Bonino[b; g] | Wilkinson, Mark D.[h]

4. ...and what FAIR is not

FAIR is not a standard: The FAIR guiding principles are sometimes incorrectly referred to as a 'standard', even though the original publication explicitly states they are not [25]. The guiding principles allow many different approaches to rendering data and services Findable, Accessible, Interoperable, to serve the ultimate goal: the reuse of valuable research objects. Standards are prescriptive, while guidelines are permissive. We suggest that a variety of valuable standards can and should be developed, each of which is guided by the FAIR Principles. FAIR simply describes the qualities or behaviours required of data resourc achieve – possibly incrementally – their optimal discovery and scholarly reuse.

FAIR is not equal to RDF, Linked Data, or the Semantic Web The reference artic Scientific Data [25] emphasises the machine-actionability of data and metadata. This implies (in fact, requires) that resources that wish to maximally fulfil the FAIR guidel must utilise a widely-accepted machine-readable framework for data and knowledge

FAIR is not just about humans being able to find, access, reformat and finally reuse data: The official press release fo authors' position clearly: "The re data publication autonomously, Principles. Computers are now a recent surveys, the time reported dealing with discovering and reu been pegged at 80% [19]. Were th deal with FAIR data and services, is today. The avoidance of time-v stewardship. To serve this potent and services should be actionable wherever possible.

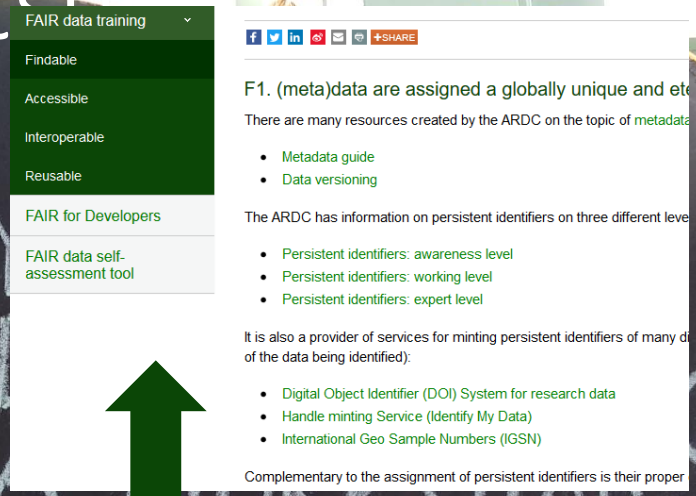
FAIR is not equal to Open: The 'A' in FAIR stands for 'Accessible under well defined conditions'. There may be legitimate reasons to shield data and services generated with public funding from public access. These include personal privacy, national security, and competitiveness. The FAIR principles, although inspired by Open Science, explicitly and

- PRINCIPLES, NOT STANDARD [IMPLEMENTATION NEEDED]
- NOT JUST FOR HUMANS
- NOT EQUAL TO LINKED DATA, RDF...
- NOT EQUAL TO «OPEN»

3. What FAIR is...

FAIR refers to a set of principles, focused on ensuring that research objects are reusable, and actually will be reused, and so become as valuable as is possible. They deliberately do not specify technical requirements, but are a set of guiding principles that provide for a continuum of increasing reusability, via many different implementations. They describe characteristics and aspirations for systems and services to support the creation of valuable research outputs that could then be rigorously evaluated and extensively reused, with appropriate credit, to the benefit of both creator and user.

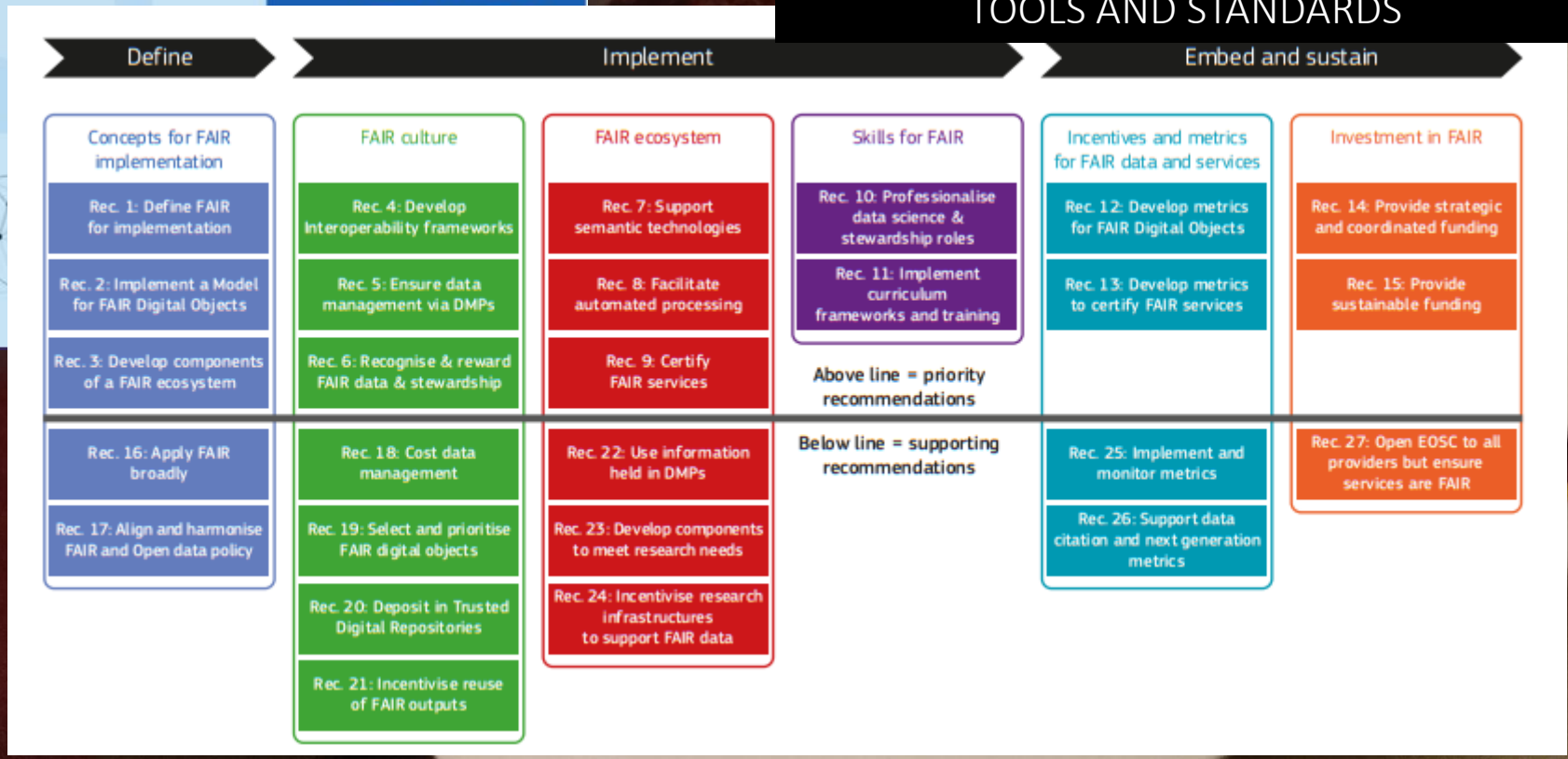
FAIR in a nut



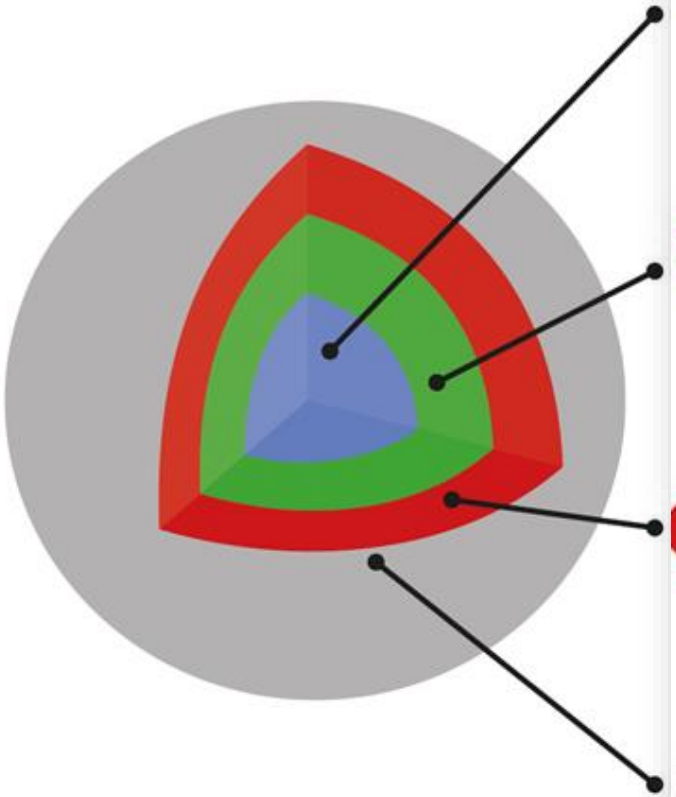
https://www.ands.org.au/working-with-data/fairdata/training

Nov. 20, 2018

«FAIR» ARE PRINCIPLES AND NOT STANDARDS.
COMMUNITIES AND DISCIPLINES SHOULD DEFINE «FAIR» FOR IMPLEMENTATION ACCORDING TO THEIR SPECIFIC DATA AND TOOLS AND STANDARDS



FAIR object



DIGITAL OBJECT

Data, code and other research outputs

At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS

Persistent and unique (PIDs)

Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and support citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

Open, documented formats

Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

METADATA

Contextual documentation

In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.

ECD, 2015

FAIR: technology VS domain

Technical infrastructure (generic operations)
Data/metadata (domain-specific content)

FAIR GENERIC VS
DOMAIN SPECIFIC
STRICTLY INTERLINKED

Box 2 | The FAIR Guiding Principles https://www.nature.com/articles/sdata201618

To be 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 it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

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

To be Reusable:
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

E.Schultes, 2019

FAIR for dummies

RESEARCHERS' RESPONSIBILITY

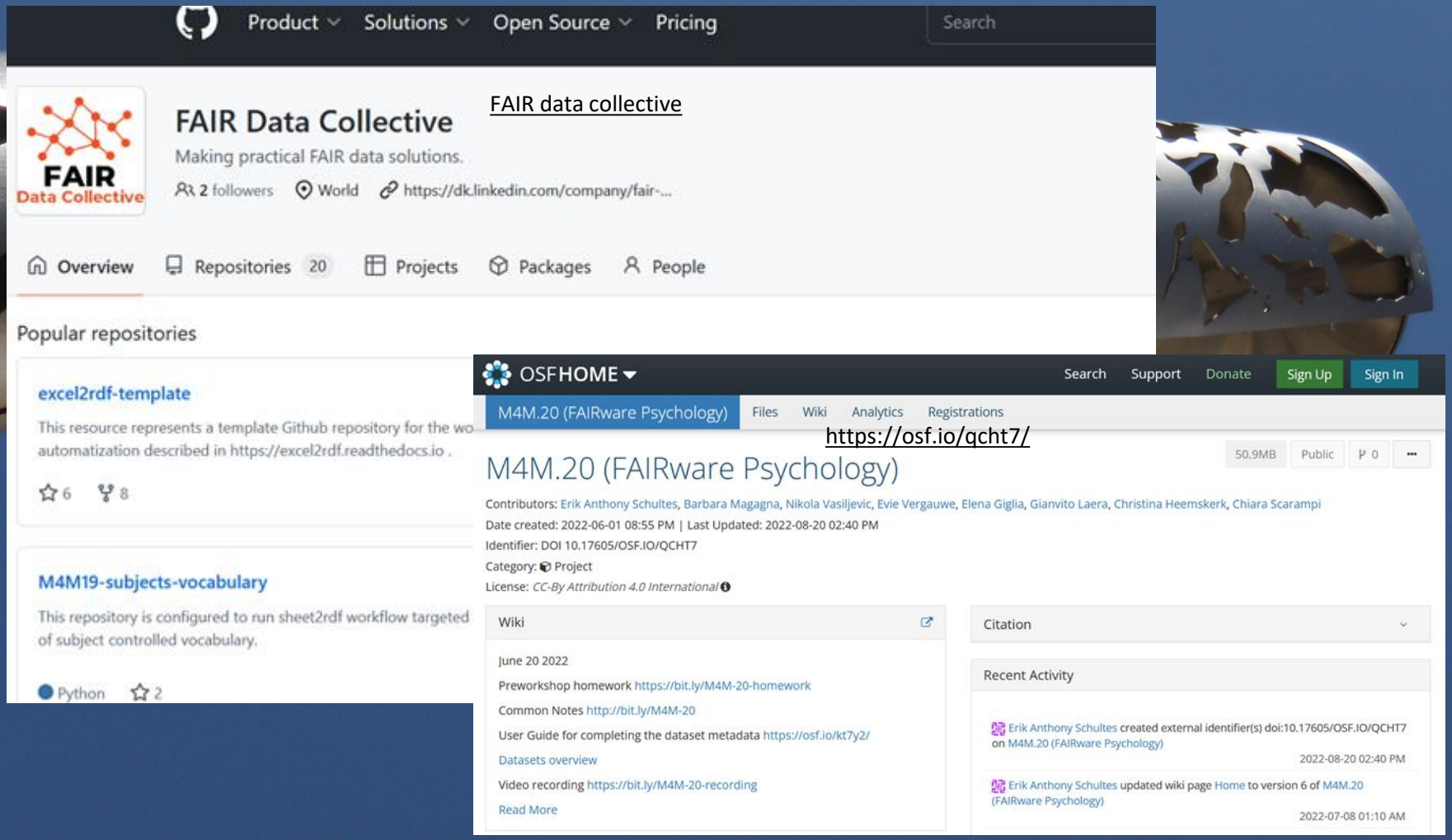
REPOSITORY TAKES CARE OF…

Explanation of the FAIR data principles 2019

Wilkinson et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* 3, doi:10.1038/sdata.2016.18

Principle		In other words	Researcher's responsibility	Requirements to be fulfilled by the repository
To be findable: Data and metadata should be easy to find by both, humans and computer systems. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.	F1. (meta)data are assigned a globally unique and persistent identifier	Each data set is assigned a globally unique and persistent identifier (PID), for example a DOI, ARK, RRID… These identifiers allow to find, cite and track (meta)data.	Ensure that each data set is assigned a globally unique and persistent identifier. Certain repositories automatically assign identifiers to data sets as a service. If not, researchers must obtain a PID via a PID registration service.	A repository needs to have a predictable way to assign a PID to each component of a dataset (e.g. each file or nanopublication), in order to be able to include these identifiers into the corresponding metadata before the submission.
	F2. data are described with rich metadata (defined by R1 below)	Each data set is thoroughly (see below, in R1) described: these metadata document how the data was generated, under what term (license) and how it can be (re)used, and provide the necessary context for proper interpretation. This information needs to be machine-readable.	Fully document each data set in the metadata, which may include descriptive information about the context, quality and condition, or characteristics of the data. Another researcher in any field, or their computer, should be able to properly understand the nature of your dataset. Be as generous as possible with your metadata (see R1).	Allow researchers to upload metadata for each data set.
	F3. metadata clearly and explicitly include the identifier of the data it describes	The metadata and the data set they describe are separate files. The association between a metadata file and the data set is obvious thanks to the mention of the data set's PID in the metadata.	Make sure that the metadata contains the data set's PID.	Allow researchers to upload metadata for each data set.
	F4. (meta)data are registered or indexed in a searchable resource	Metadata are used to build easily searchable indexes of data sets. These resources will allow to search for existing data sets similarly to searching for a book in a library.	Provide detailed and complete metadata for each data set (see F2).	Request and store part of the metadata in a structured way, for example by providing a form with specific fields to be completed or by providing an XML schema to be used by the researchers. For example the storing of PID's, author names, disciplines, etc. will facilitate the creation of indexes. However, it must remain possible to provide arbitrary metadata in addition.

FAIR – how to

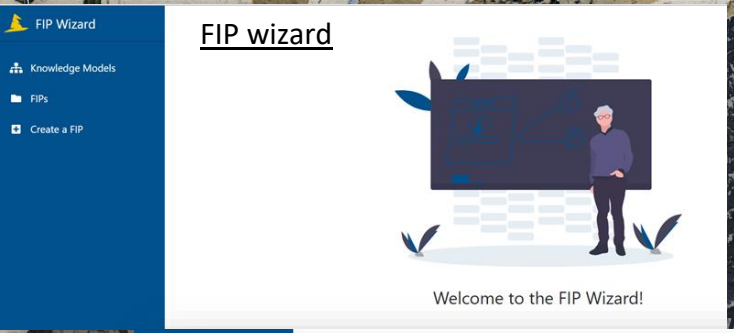


FAIR data collective

https://osf.io/qcht7/

FAIR Implementation profiles

FIP wizard



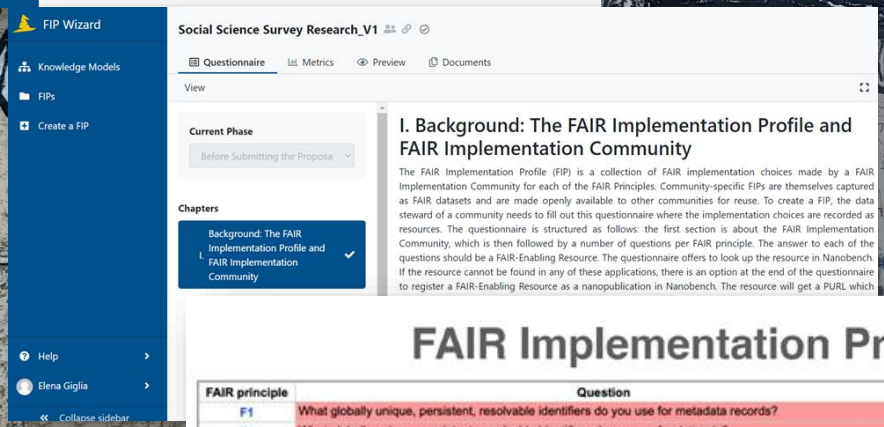
2020

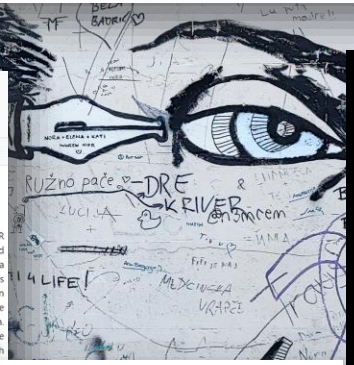
International Conference on Conceptual Modeling
ER 2020: Advances in Conceptual Modeling pp 138-147 | Cite as

Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence

Authors | Authors and affiliations

Erik Schultes, Barbara Magagna, Kristina Maria Hettne, Robert Pergl, Marek Suchánek, Tobias Kuhn





CREATE FAIR IMPLEMENTATION PROFILES REUSBALE BY YOUR COMMUNITY - KEYWORD: CONVERGENCE

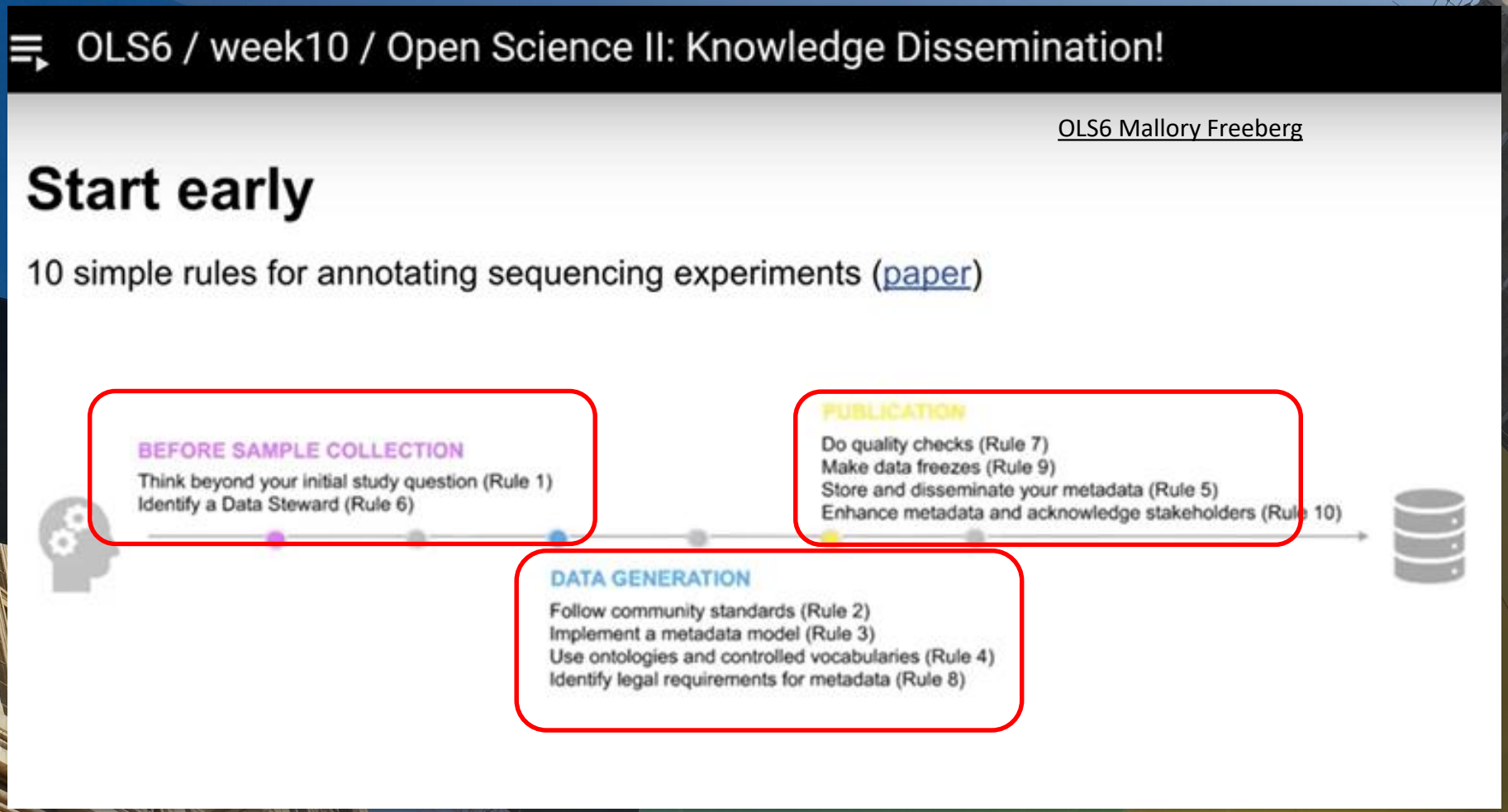
FAIR Implementation Profile

FAIR principle	Question	FAIR enabling resource types
F1	What globally unique, persistent, resolvable identifiers do you use for metadata records?	Identifier type
F1	What globally unique, persistent, resolvable identifiers do you use for datasets?	Identifier type
F2	Which metadata schemas do you use for findability?	Metadata schema
F3	What is the technology that links the persistent identifiers of your data to the metadata description?	Metadata-Data linking mechanism
F4	In which search engines are your metadata records indexed?	Search engines
F4	In which search engines are your datasets indexed?	Search engines
A1.1	Which standardized communication protocol do you use for metadata records?	Communication protocol
A1.1	Which standardized communication protocol do you use for datasets?	Communication protocol
A1.2	Which authentication & authorisation technique do you use for metadata records?	Authentication & authorisation technique
A1.2	Which authentication & authorisation technique do you use for datasets?	Authentication & authorisation technique
A2	Which metadata longevity plan do you use?	Metadata longevity
I1	Which knowledge representation languages (allowing machine interoperation) do you use for metadata records?	Knowledge representation language
I1	Which knowledge representation languages (allowing machine interoperation) do you use for datasets?	Knowledge representation language
I2	Which structured vocabularies do you use to annotate your metadata records?	Structured vocabularies
I2	Which structured vocabularies do you use to encode your datasets?	Structured vocabularies
I3	Which models, schema(s) do you use for your metadata records?	Metadata schema
I3	Which models, schema(s) do you use for your datasets?	Data schema
R1.1	Which usage license do you use for your metadata records?	Data usage license
R1.1	Which usage license do you use for your datasets?	Data usage license
R1.2	Which metadata schemas do you use for describing the provenance of your metadata records?	Provenance model
R1.2	Which metadata schemas do you use for describing the provenance of your datasets?	Provenance model

How-to

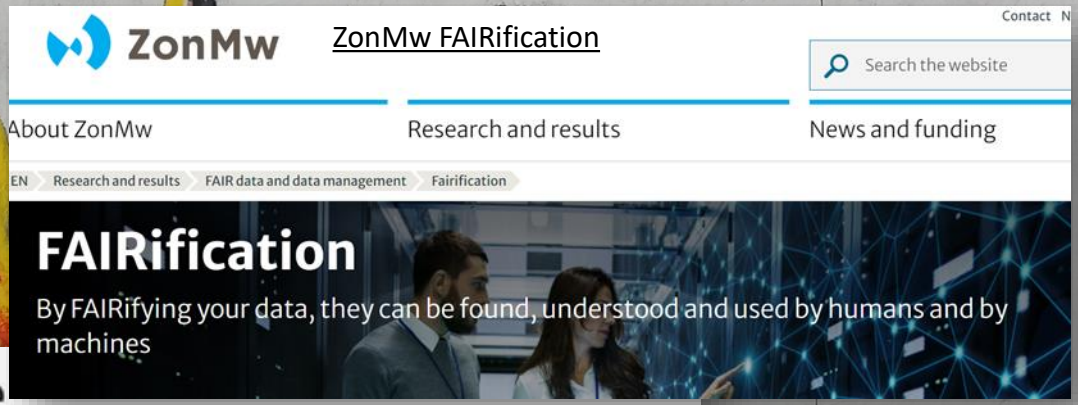
THIS IMAGE CREATED BY
Noa, Anna, Lilian e Charlotte
PERFECTLY SHOWS WHAT
«MAKING DATA FAIR» MEANS

FAIRifying


OLS6 / week10 / Open Science II: Knowledge Dissemination!
OLS6 Mallory Freeberg
Start early
10 simple rules for annotating sequencing experiments (paper)
BEFORE SAMPLE COLLECTION
Think beyond your initial study question (Rule 1)
Identify a Data Steward (Rule 6)
DATA GENERATION
Follow community standards (Rule 2)
Implement a metadata model (Rule 3)
Use ontologies and controlled vocabularies (Rule 4)
Identify legal requirements for metadata (Rule 8)
PUBLICATION
Do quality checks (Rule 7)
Make data freezes (Rule 9)
Store and disseminate your metadata (Rule 5)
Enhance metadata and acknowledge stakeholders (Rule 10)

FAIRification

ZonMw FAIRification



FAIRification in practice

The purpose of this section is to provide background information for researchers and data stewards who are active in FAIRifying their data. With the term FAIRification we stress that the creation of FAIR data is a process, in which data gradually become more FAIR. At the end, data are optimally reusable, both by humans and -where possible- by machines, with full compliance to privacy protection regulations (if relevant). FAIRification is important for all types of data, whether they are generated through research, innovation processes, or societal activities.

- Read more about the FAIR guiding principles
- FAIR is not an 'all or nothing' state
- Data and 'other things' to FAIRify
- Some important aspects of FAIR data that we have to keep in min
- As open as possible, as closed as necessary
- Data management and FAIR data stewardship are related, but not the same
- The FAIR data-ecosystem: infrastructure and services
- The FAIR data-ecosystem: data stewardship capacity
- What can we do with FAIR data?

FAIR is not an 'all or nothing' state

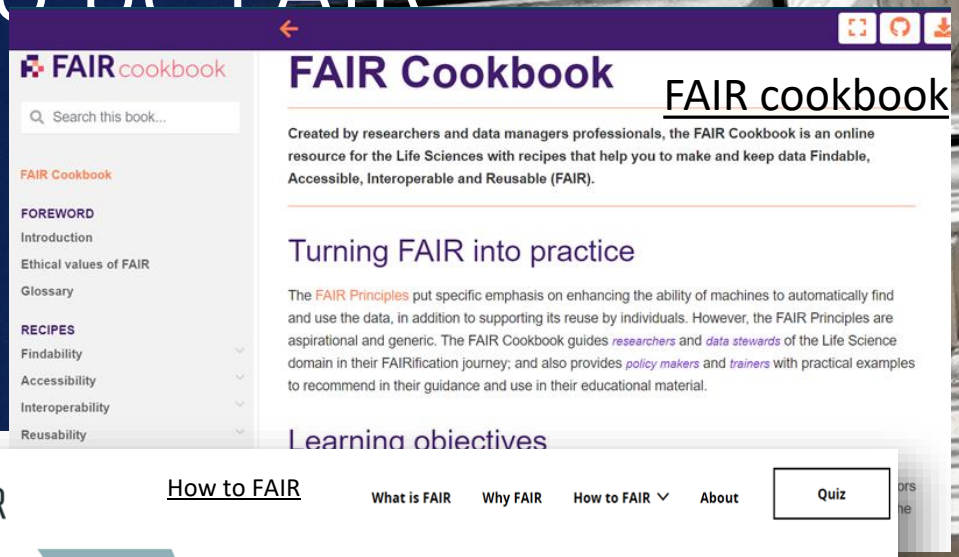
FAIR data is not a well-defined endpoint. Instead, data may gain a certain level of FAIRness through data stewardship actions, taking FAIR principles as a guidance. Depending on their goals, researchers and data stewards may decide to focus specifically on for instance findability, or interoperability (etc). Implementing all FAIR principles is very challenging though, and for most researchers and data stewards not yet possible because they lack the appropriate knowledge, tools or infrastructure. Strictly speaking, however, as long as data (or their metadata) are not machine readable, they should not be labelled as 'FAIR'.

You can read more about a step-by- step workflow for FAIRification, and take a look at some examples of tools therefore, such as the RDA FAIR Data Maturity Model, and the Data Stewardship Wizard.

ZonMw requires grant holders to take actions to make data as findable, accessible, interoperable and reusable as possible, and appropriate for the type of project. ZonMw's M4M-workshops for the COVID-19 research programme were the first step towards machine readability, and thereby achieve some 'true' FAIRness of data in projects it funds.You can read more about the concept of metadata for machines (M4M) and find out how they are produced, and can be used.

PRACTICAL AND QUICK GUIDE

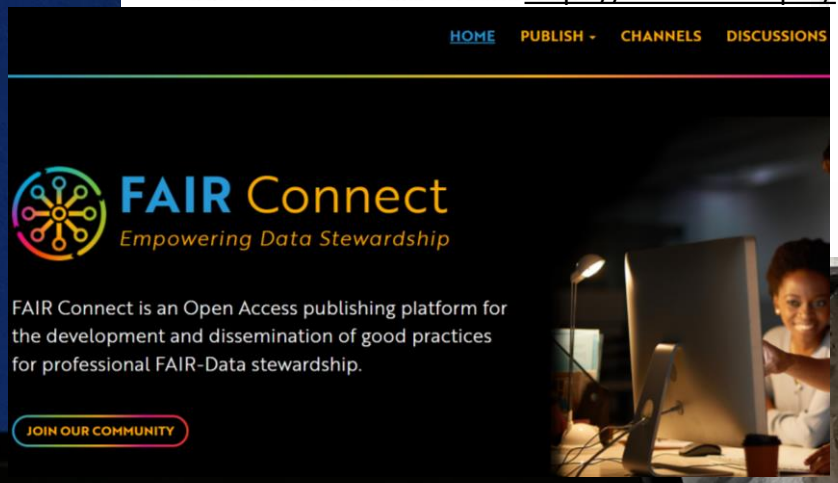
Support / How to be FAIR

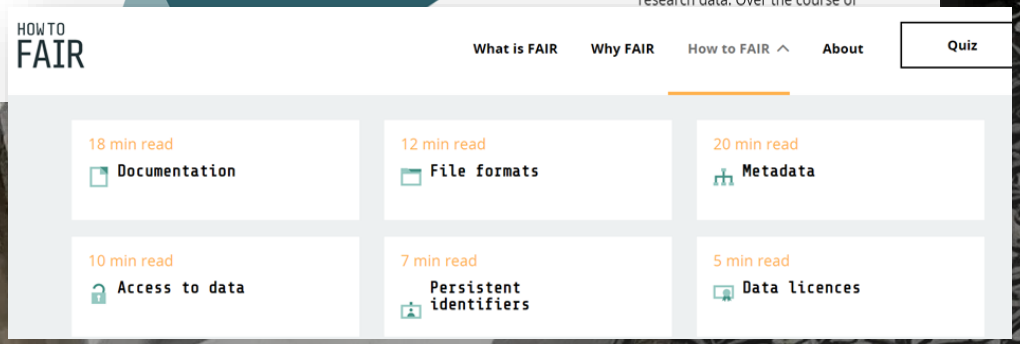


FAIR cookbook

How to FAIR

https://fairconnect.pro/





FAIRifi

Jan 2020

Volume 2, Issue 1-2
Winter-Spring 2020

DATA Intelligence

January 01 2020

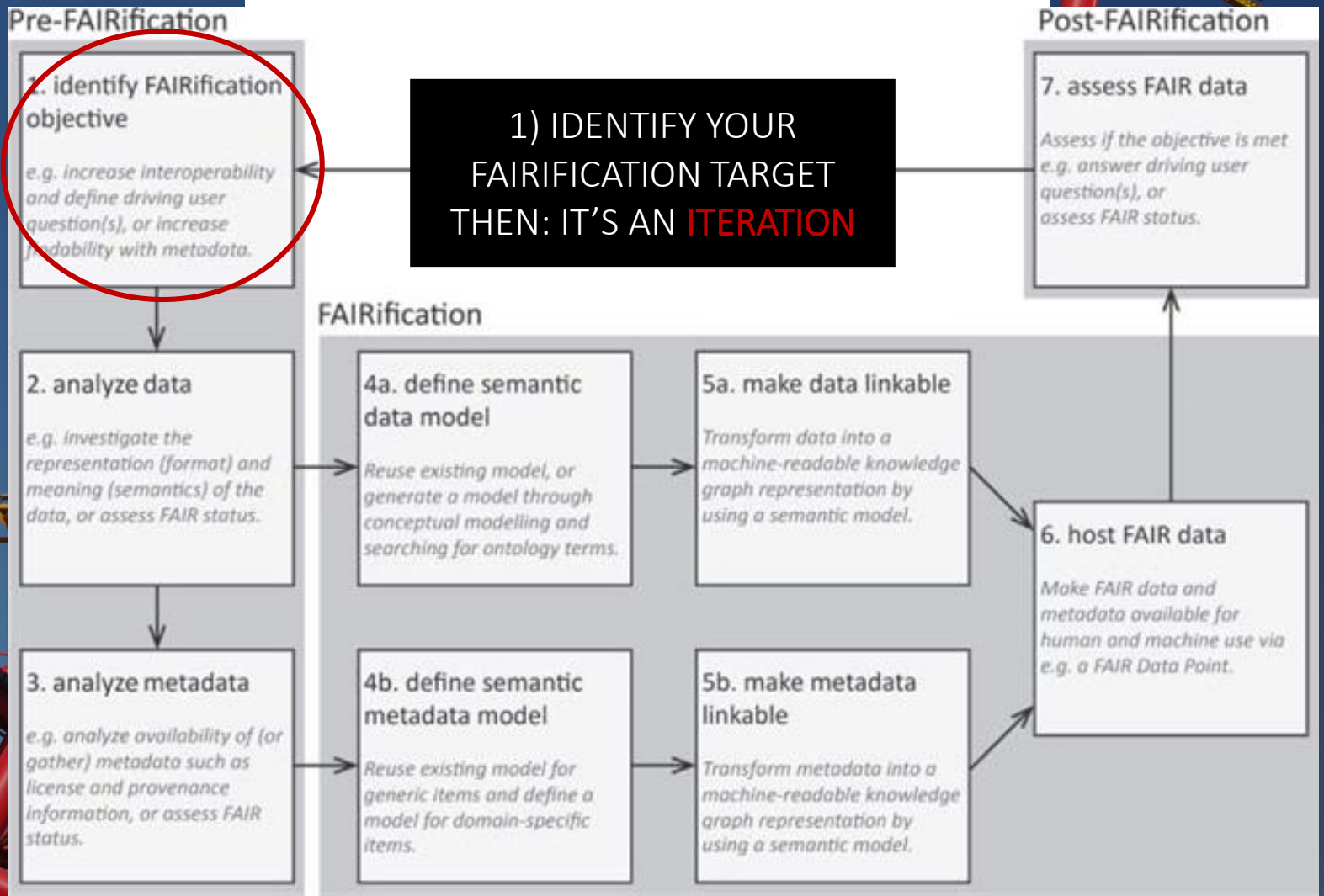
A Generic Workflow for the Data FAIRification Process

Annika Jacobsen , Rajaram Kaliyaperumal , Luiz Olavo Bonino da Silva Santos , Barend Mons , Erik Schultes , Marco Roos , Mark Thompson

Author and Article Information

Data Intelligence (2020) 2 (1-2): 56–65.

https://doi.org/10.1162/dint_a_00028



FAIR is a process

FAIR = CONTINUUM
«AS FAIR AS POSSIBLE»

Inclusiveness: consider FAIR as a process

If FAIR is not seen as a continuum, we risk losing communities who are not well advanced in sharing their data in a FAIR way, as well as advanced communities for whom the effort to attain optional indicators doesn't outweigh the effort required. In addition to avoiding "mandatory" criteria, using multi-step maturity scales to measure the FAIRness level of a resource, instead of a yes/no evaluation for each criterion, would provide an inclusive system, and a way to set up

EOSCFAIR
Executive Board Working Group

Interim recommendations on FAIR Metrics for EOSC
February 2020
Draft for consultation

Feb. 2020

To check your FAIRness

FAIRassist.org https://fairassist.org/#!/

Help you discover resources to measure and improve FAIRness.

FAIRassist is the new, under development, educational component of the well established FAIRsharing resource.

Resource	Execution Type	Key Features	Organisation	Target Objects	Reading Material
5 Star Data Rating Tool	Manual - questionnaire	Based on rating systems and maturity models	CSIRO OzNome	Datasets	
AutoFAIR	Semi-automated	A portal for automating FAIR assessments for bioinfo	Department of Computer		
Data Stewardship Wizard	Predictive; based on a manually filled questionnaire	Helps researchers to design a data stewardship proc highest reasonable FAIR data.			
F-UJI	Automated	The REST API support a programmatic assessment o objects based on a set of core metrics developed by t metrics specification is available at https://doi.org/10			
FAIR Data Self-Assessment Tool	Manual - questionnaire	Educational and informational purposes			
FAIR Evaluator	Automated	1. Core universal maturity indicators 2. Compliance tests 3. Evaluation tool			
FAIR enough	Automated	1. Core universal maturity indicators and community compliance tests 2. Stable and fast evaluations execution (less than 1min for most evaluated resources, no commercial license required) 3. Library for defining, publishing and registering new maturity indicators 4. Supports ORCID authentication for creating collections and authoring evaluations			Maastricht Uni
FAIR-Aware	Manual - questionnaire	1. Online self-assessment that helps to assess current level of awareness on making datasets FAIR before depositing them in a data repository. 2. Added guidance texts explain the what, why, and how of each FAIR practice. 3. Trainer functionality allows flexible use of the tool for your own purpose			FAIRsFAIR (D
FAIR-Checker	Automated	FAIR-Checker is a web interface to evaluate FAIR metrics (as implemented through the FAIR Evaluation Service APIs https://fairsharing.github.io/FAIR-Evaluator-FrontEnd) and to provide developers with technical FAIRification hints. It's also a Python framework aimed at easing the implementation of FAIR metrics.			IFB (ELIXIR-
FAIRdat	Manual - questionnaire	A 5-star rating of the FAIR principles			DANS
FAIRness self-assessment grids	Manual - checklist	1. Assessment grids: quick and extensive 2. Designed as a decision tree 3. Researcher focused			RDA-SHARC
FAIRshake	Manual - questionnaire,	1. FAIR metrics (questions) and rubrics (collection of metrics)			NIH Data Com

… FAIRy shades

Findable

Does the dataset have any identifiers assigned? — No identifier

Is the dataset identifier included in all metadata records/files describing the data? — No

How is the data described with metadata? — The data is not described

What type of repository or registry is the metadata record in? — The data is not described in any repository

Accessible

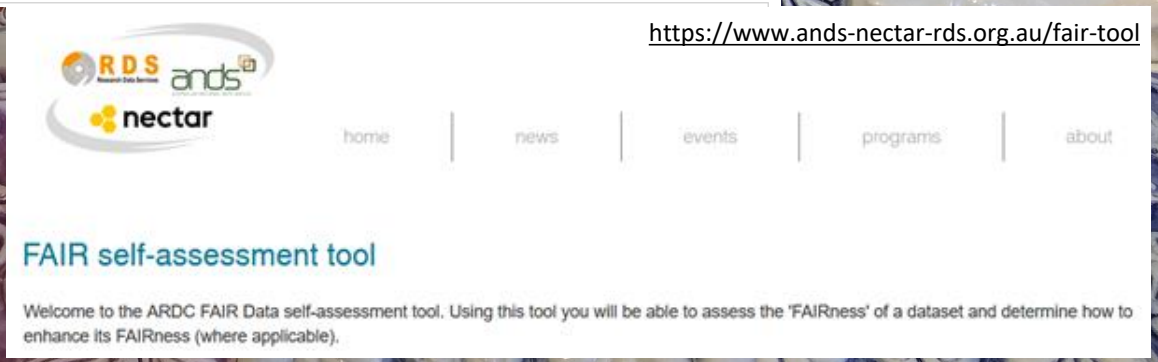
How accessible is the data? — No access to data or metadata

Is the data available online without requiring specialised protocols or tools once access has been approved? — No access to data

Will the metadata record be available even if the data is no longer available? — Unsure

VERY USEFUL TO ASK THE RIGHT QUESTIONS BUT IT'S SUBJECTVE…

https://www.ands-nectar-rds.org.au/fair-tool



FAIR aware

- QUESTIONS
- TESTS KNOWLEDGE
- TESTS WILLINGNESS
- GIVES INFO

FAIR Aware

Data Archiving and Networked Services DANS

FAIRaware

Let's assume you have research data almost ready for uploading to a repository: do you already know how you and the repository can work together to make the data as findable, accessible, interoperable and reusable (FAIR) as possible? By guiding you through the assessment process, the FAIR-Aware tool can help you to better understand the FAIR Principles and how making data FAIR can increase the potential value and impact of your data.

FAIR-Aware is an disciplinary-agnostic online tool developed by the FAIRsFAIR project. Different scientific communities can adapt it to their own use. You should, however, have a target dataset in mind to be able to answer the questions and complete the assessment.

FAIR questions

FINDABLE

1. Are you aware that a data(set) should be assigned a global persistent and resolvable identifier when deposited with a dat repository?

2. Are you aware that when you deposit a data(set) in a data you will need to provide discovery metadata in order to make data(set) findable, understandable and reusable to others?

3. Are you aware that the data repository providing access to data(set) should make the metadata describing your data(set) in a format readable by machines as well as humans?

ACCESSIBLE

4. Are you aware that access to your data(set) may need to b and that metadata should include licence information under w data(set) can be reused?

5. Are you aware that metadata should remain available over if the data(set) is no longer accessible?

unique persistent and resolvable identifier when deposited with a data repository?

What does this mean?

A **persistent identifier** is a long-lasting reference to a resource. The **data(set)** you deposit in a **data repository** should be assigned a globally unique, persistent and resolvable identifier (PID) so that both humans and machines can find it. Persistent identifiers are maintained and governed so that they remain stable and direct the users to the same relevant object consistently over time. Examples of PIDs include Digital Object Identifier (DOI), Handle, and Archival Resource Key (ARK).

Why is this important?

If your data(set) or metadata does not have a PID, you run the risk of **"link rot"** (also known as "link death"). When your data(set) or metadata is moved, updated to a new version, or deleted, older hyperlinks will no longer refer to an active page. Without a PID, others will not be able to find or reuse your data(set) or metadata in the long-term.

How to do this?

When you upload your data(set) or metadata to a data repository, the data repository (or other service providers) usually assigns a PID. Repositories ensure that the identifier continues to point to the same data or metadata, according to access terms and conditions you specified.

There are many different types of PIDs, each with their own advantages, disadvantages, and disciplines they are typically used in. Generally speaking, the data repository will have thought about these aspects before deciding which PID type to use. In case you have to choose the PID type yourself, you can visit the Knowledge Hub on the PID Forum for guidance. Some disciplines or organisations also provide tools to help you make this choice, see for example this Persistent Identifier Guide for cultural heritage researchers. Once you have chosen a PID type, you can search for data repositories providing that specific PID in registries such as Re3data or FAIRsharing (see related databases).

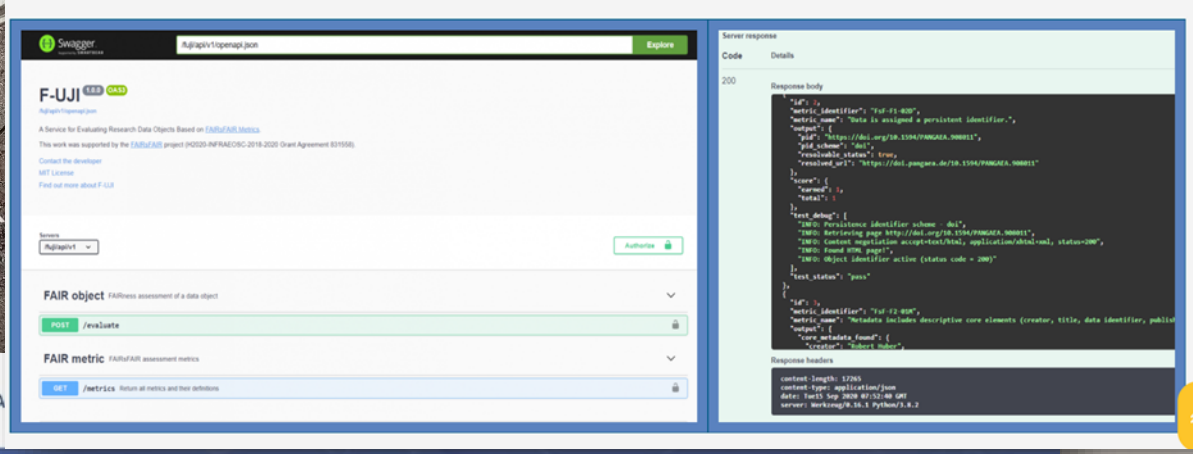
Not all data you produce during your research will need a PID. In general, those that underpin published findings or have longer term value are worth assigning a PID. If in doubt about which data should be allocated a PID, speak to your local research data management support team or the data repository.

F-UJI

FAIRSFAIR

About FAIR Support FAIR Landscape Tools & Software Competence Centre Events Project Outputs Outreach

Screenshots of the tool below



FAIRSFAIR F-UJI About FA

F-UJI Automated FAIR Data Assessment Tool

Home / F-UJI Automated FAIR Data Assessment Tool

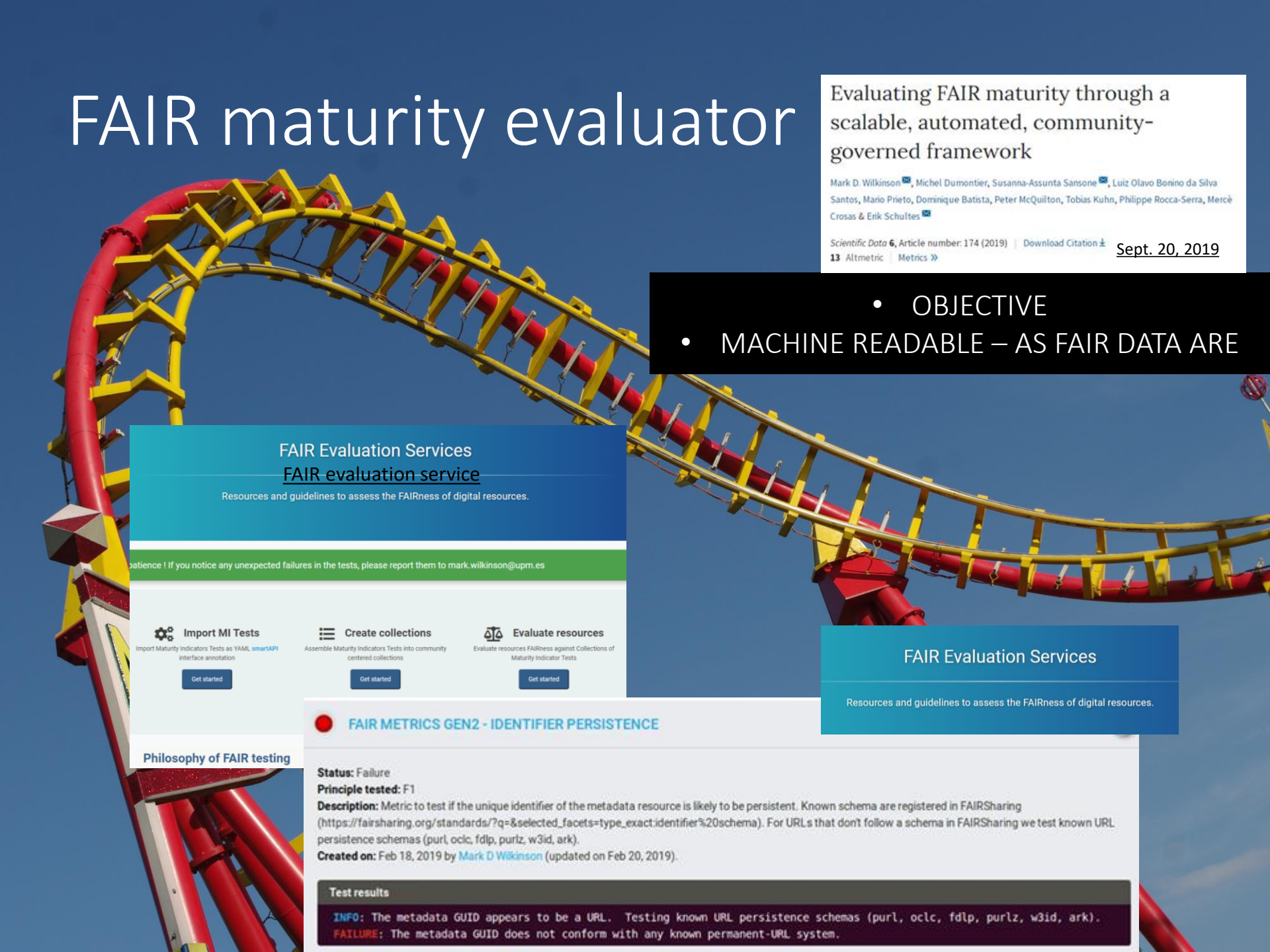
FAIRNESS EVALUATION (IN BETA)

FAIRsFAIR has developed F-UJI, a service based on REST, and is piloting a programmatic assessment of the FAIRness of research datasets in five trustworthy data repositories.

F-UJI Automated FAIR Data Assessment Tool

The F-UJI assessment is based on **16 out of 17 core FAIR object assessment metrics** developed within FAIRsFAIR and each corresponding to a part or the whole of a FAIR principle. F-UJI adheres to existing web standards and PID resolution services best practices and utilises external registries and resources such as re3data[1] and Datacite[2] APIs, SPDX License List[3], RDA Metadata Standards Catalog[4], and Linked Open Vocabularies (LOV)[5] For information on the practical tests implemented against the metrics, see Devaraju, Huber, et al., 2020.

FAIR maturity evaluator



Evaluating FAIR maturity through a scalable, automated, community-governed framework

Mark D. Wilkinson, Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Dominique Batista, Peter McQuilton, Tobias Kuhn, Philippe Rocca-Serra, Mercè Crosas & Erik Schultes

Scientific Data **6**, Article number: 174 (2019) | Download Citation
13 Altmetric | Metrics »

Sept. 20, 2019

- OBJECTIVE
- MACHINE READABLE – AS FAIR DATA ARE

FAIR Evaluation Services

FAIR evaluation service

Resources and guidelines to assess the FAIRness of digital resources.

patience ! If you notice any unexpected failures in the tests, please report them to mark.wilkinson@upm.es

Import MI Tests
Import Maturity Indicators Tests as YAML smartAPI interface annotation
Get started

Create collections
Assemble Maturity Indicators Tests into community centered collections
Get started

Evaluate resources
Evaluate resources FAIRness against Collections of Maturity Indicator Tests
Get started

Philosophy of FAIR testing

FAIR Evaluation Services

Resources and guidelines to assess the FAIRness of digital resources.

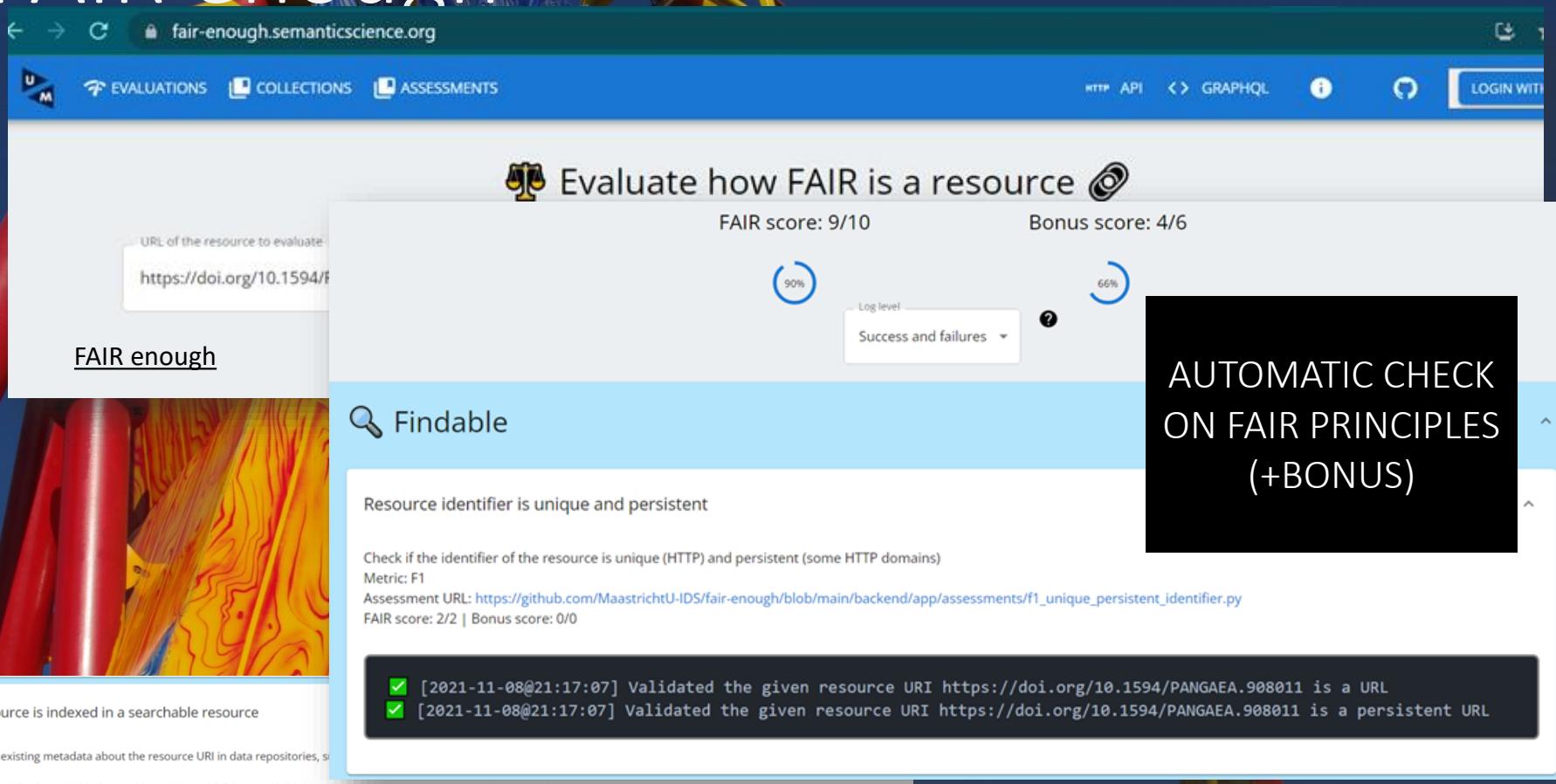
FAIR METRICS GEN2 - IDENTIFIER PERSISTENCE

Status: Failure
Principle tested: F1
Description: Metric to test if the unique identifier of the metadata resource is likely to be persistent. Known schema are registered in FAIRSharing (https://fairsharing.org/standards/?q=&selected_facets=type_exact:identifier%20schema). For URLs that don't follow a schema in FAIRSharing we test known URL persistence schemas (purl, oclc, fdlp, purlz, w3id, ark).
Created on: Feb 18, 2019 by Mark D Wilkinson (updated on Feb 20, 2019).

Test results

INFO: The metadata GUID appears to be a URL. Testing known URL persistence schemas (purl, oclc, fdlp, purlz, w3id, ark).
FAILURE: The metadata GUID does not conform with any known permanent-URL system.

FAIR enough



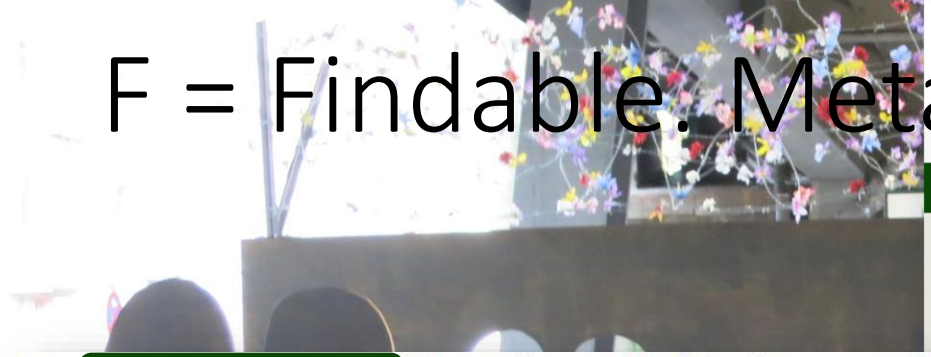
AUTOMATIC CHECK ON FAIR PRINCIPLES (+BONUS)

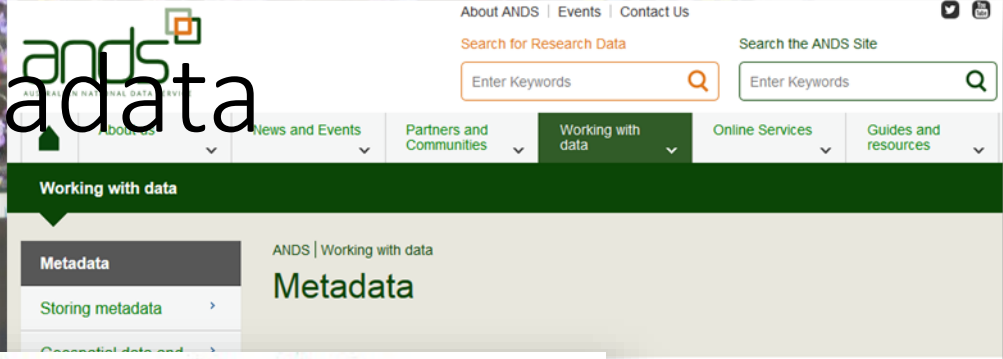
...FAIR fo

SELF-ASSESSMENT ON FAIR-ENABLING [INSTITUTIONS]

FINDABLE

F = Findable: Metadata

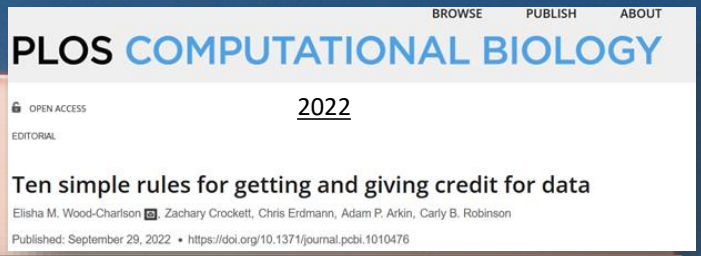




- **Descriptive metadata**: information required for discovery and assessment of the collection,
  - e.g. title, contributors, subject or keywords, study description, and location and dates of the study.
- **Provenance metadata**: this relates to the origins and processing of the data, and enables interpretation and reuse of the data. It ranges from the human to the highly technical, and usually requires some knowledge of the domain to create.
  - e.g. Where did the data come from? Why was it collected? Who collected it, when and where? What instruments/technologies were used to collect the data, and how were they set up? How has the data been processed?
- **Technical metadata**: fundamental information for a person or a computer application to read the data.
  - e.g. How is the data set up? What formats, and versions of formats, are used? How is the database configured? How does it relate to other data?
- **Rights and access metadata**: information to enable access, and licensing or usage rules.
  - e.g. How can someone access the data? Who is allowed to view or modify the data, or the metadata, and under what conditions? Who has some kind of authority over the data? Are there costs associated with access? Under what licence is the data being made available?
- **Preservation metadata**: this builds on the history from the Provenance, Rights and Technical metadata, and also includes information to allow the data to be managed for long-term accessibility.
  - e.g. Has there been any restructuring or other changes to the files, e.g. due to migration to new file formats? What software has been used to access the data?
- **Citation metadata**: information required for someone to cite the data
  - e.g. Creator(s), Publication Year, Title, Publisher, Identifier.

Metadata

F=Findable - Metadata



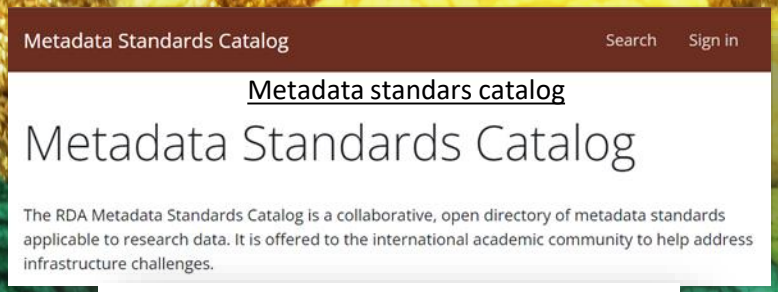
2022

Data management stage	Metadata fields	Standardize public resources
Sample	Latitude, longitude, date/time, temperature, biome/ ecosystem, depth and/or elevation of sampling site, etc.	Environmental Ontology (ENVO), Minimum Information about x Sequence (MIxS), International Geo/General Sample Number (IGSN)
Preparation	Laboratory protocol(s): DNA extraction, purification, amplification.	Protocols.io, e-laboratory notebook/management software
Data processing	Software tools for QA/QC, assembly, annotation. Include reference (if published), version, and parameters used.	Community guidelines for describing and citing software [23–25]
Feature	E.g., Annotations of sequence data, such as taxonomy or function	NCBI Taxonomy, Genome Taxonomy Database toolkit (GTDB-tk); Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), etc.
FAIR: Findable (i.e., PID metadata)	Data owner(s), organization, keywords	ORCID, Researcher Organization Registry (ROR); keyword selection [26] to enhance search engine optimization (SEO)
FAIR: Accessible	Usage license, privacy protocols, transfer protocols	Creative Commons, HTTP
FAIR: Interoperable	Type and size of data, file formats, etc.	.csv,.tsv, etc.
FAIR: Reusable	See data processing.	Workflow notebooks (e.g., [27])

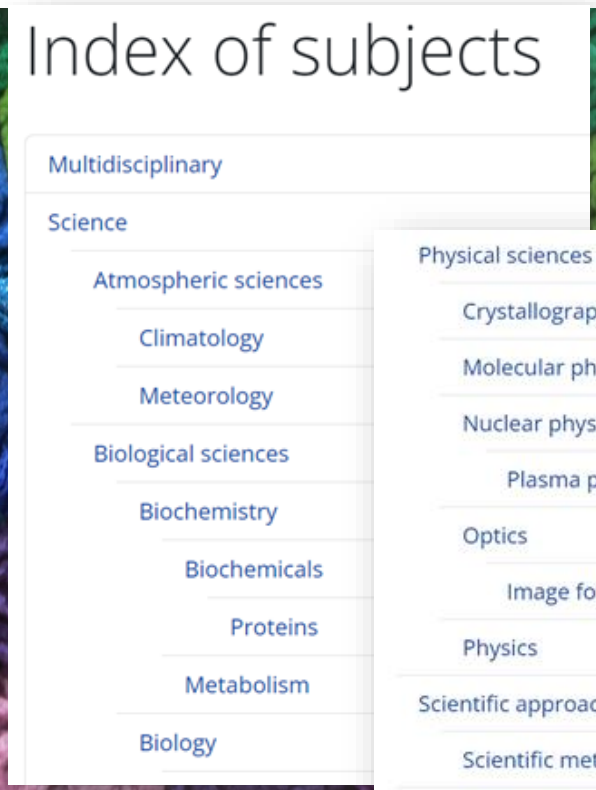
https://doi.org/10.1371/journal.pcbi.1010476.t001

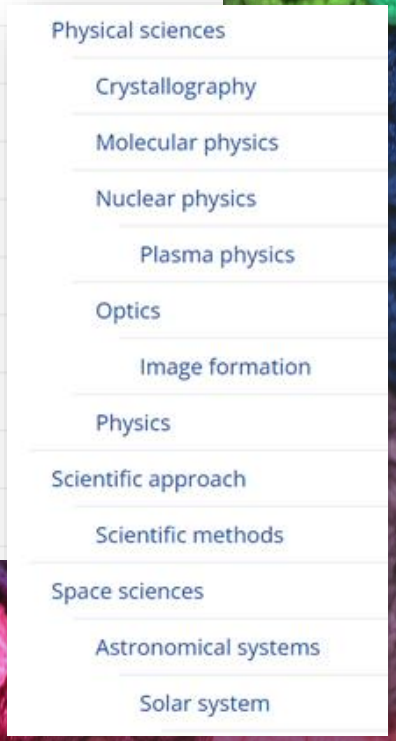
EXAMPLE OF MINIMAL METADATA SET

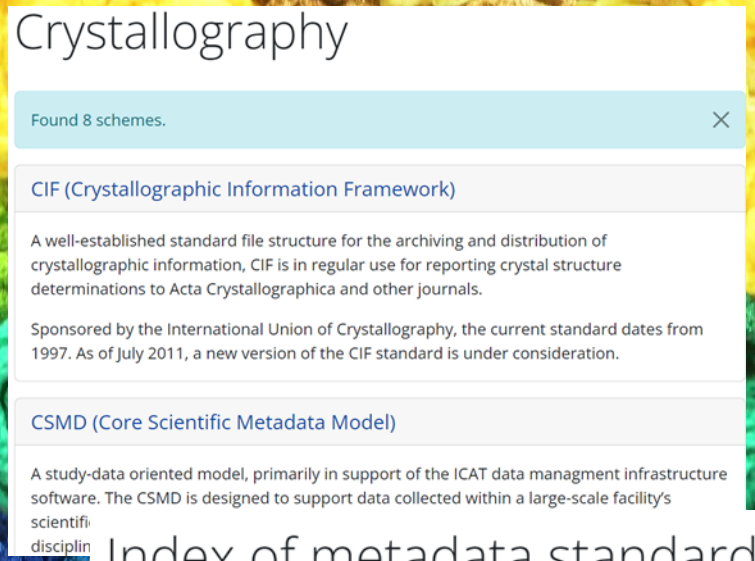
F = Findable. Metadata standards

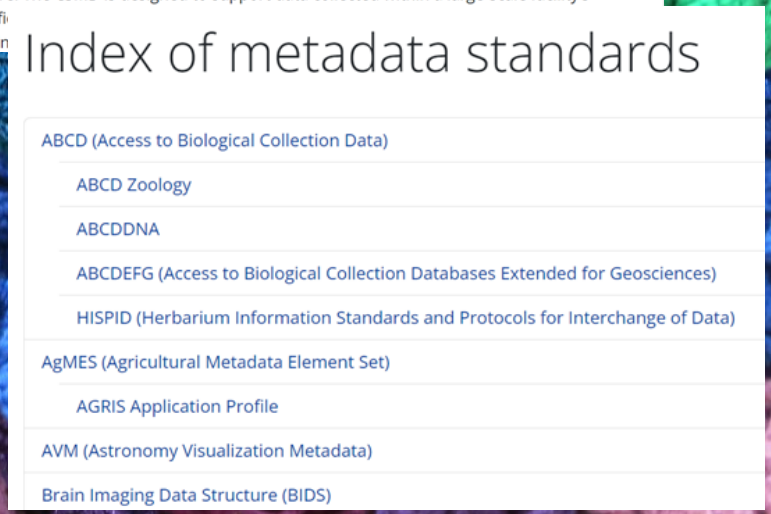


Metadata standars catalog









F= Findable - Met

OpenRefine Download Documentation Community Blog Donate

https://openrefine.org/

OpenRefine

OpenRefine is a powerful free, open source tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

Download

Faceting

Drill through large datasets using facets and apply operations on filtered views of your dataset.

Clustering

Fix inconsistencies by merging similar values thanks to powerful heuristics.

Reconciliation

Match your dataset to external databases via reconciliation services.

Infinite undo/redo

Rewind to any previous state of your dataset and replay your operation history on a new version of it.

Privacy

Your data is cleaned on your machine, not in some dubious data laundering cloud.

Wikibase

Contribute to Wikidata, the free knowledge base anyone can edit, and other Wikibase instances.

FRICTIONLESS DATA

Introduction Projects Universe Adoption People Fellows Development Worl

https://frictionlessdata.io/

Data software and standards

Frictionless is an open-source toolkit that brings simplicity to the data experience - whether you're wrangling a CSV or engineering complex pipelines.

Why Frictionless Data? Get Started

How can I use Frictionless?

You can use Frictionless to describe your data (add metadata and schemas), validate your data, and transform your data. You can also write custom data standards based on the Frictionless specifications. For example, you can use Frictionless to:

- easily add metadata to your data before you publish it.
- quickly validate your data to check the data quality before you share it.
- build a declarative pipeline to clean and process data before analyzing it.

Usually, new users start by trying out the software. The software gives you an ability to work with Frictionless u
interfaces or programming languages.

As a new user you might not need to dive too deeply into the standards as our software incapsulates its concept
other hand, once you feel comfortable with Frictionless Software you might start reading Frictionless Standards
better understanding of the things happening under the hood or to start creating your metadata descriptors mo
proficiently.

TOOL TO
- DESCRIBE DATA
- VALIDATE
- BUILD A PIPELINE

F = findable. Metadata tools

What CEDAR does

https://metadatacenter.org/

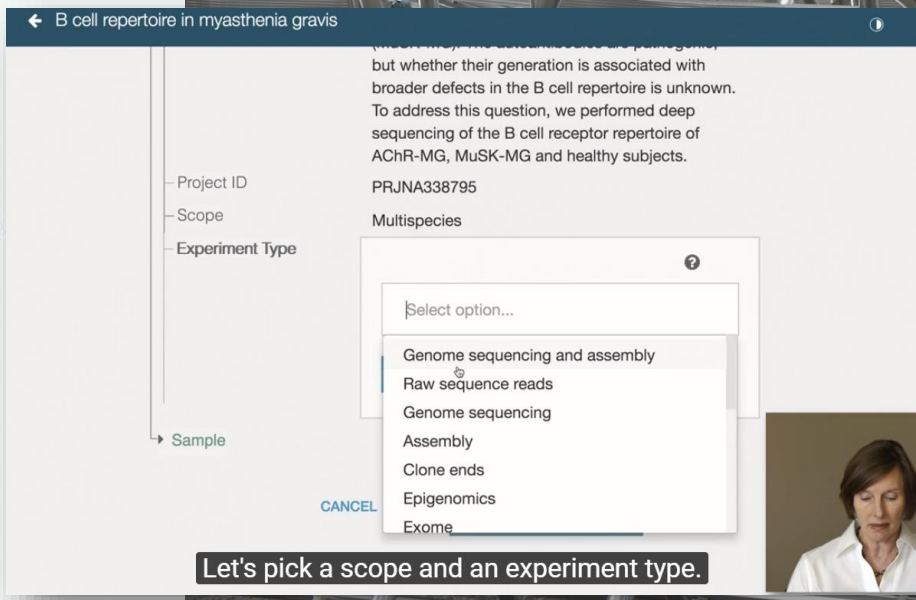
The CEDAR Workbench, as we refer to the suite of CEDAR tools, makes it easy to collect and use metadata. Eventually our tools wil
metadata record is created to its eventual processing, and even enhancement, by users and analysts. But for now, CEDAR tools hel
to users, and download the information that users have provided.

What can CEDAR do for me already?

As of its production release, in February 2017, CEDAR addresses these scenarios:

- create user-friendly, shareable forms for collecting metadata, with features like
  - nested and repeatable elements and fields
  - reusable elements
  - control over tool tips, field titles, and field descriptions
- share your forms and metadata
  - provide a link to your metadata editors, so they can enter metadata responses based on your forms
  - share your forms and other content with individuals or a group
  - create and manage groups to make permissions simpler
- associate your questions (fields) and possible answers (values) with controlled terms
  - select any term or collection of terms from the NCBO BioPortal semantic repository
  - combine different terms from different controlled vocabularies into a single set of options
  - create your own terms, or term lists ('value sets') that can be re-used
- view responses meeting your (simple) search critieria, in several forms
  - CEDAR Metadata Editor's metadata view
  - an in-line JSON-LD format, used by CEDAR for all its metadata instances
  - download of JSON-LD files via the CEDAR REST API, for offline integration with your workflow
- use the Workbench Desktop interface to manage your content
  - use My Workspace to see your items, or Shared with Me to see other items you can access
  - select an item and control-click or use the 3-dot menu in the upper right to share it, copy it, delete it, or get info on it
- enable intelligent metadata suggestions in your template by using a field's Suggestions tab
  - CEDAR keeps track of metadata entered for that field
  - users will see a drop down list of the most popular metadata entries, and can select from them
- remotely access CEDAR content and capabilities using the CEDAR REST API

With these capabilities, you can capture simple or rich metadata for your project, build a repository of project metadata, or design
particular needs. Advanced users can even submit metadata entries through CEDAR's REST API.



Findable — Metadata cre

FAIR cookbook

FAIR Cookbook

- FAIR Cookbook
- Introduction
- Assessing FAIR
- Infrastructure for FAIR
- Improving Findability
- Improving Accessibility
- Improving Interoperability
  - How to interlink data from different sources?
  - Identifier mapping with BridgeDB
  - Which vocabulary to use?
  - Requesting terms addition to terminology artefacts

Creating a Metadata Profile

Recipe metadata

identifier: RX.X version: v1.0

Difficulty level

Reading Time

20 minutes

Recipe Type

Hands-on

Executable Code

Yes

Intended Audience

Principal Investigator

Data Manager

Data Scientist

How to generate a metadata template

The following steps are intended as a starting point to guide the generation of a metadata template.

Step 1: Define competency questions

- What are the questions you would like to address with the template? Without a set of a competency questions, important variables may easily be forgotten. It is equally possible to collect too much metadata, making the resulting metadata model opaque and difficult to navigate. Competency questions serve as a guide to identify the most relevant experimental factors.

Step 2: Define a Minimal Set Of Metadata (MSOM) according to these questions

- Compile metadata from different sources
- Generate consolidated view on metadata by merging attributes as far as possible
- Differentiate metadata available for most of the studies from metadata occurring rarely (sparse matrix)
- Identify gaps in the metadata available for most of the studies comprising data that is considered important but has not been captured in the past
- Define a MSOM to be captured in the future from the metadata that is available for most of the studies and the metadata considered to be important
- Identify available community standards regarding minimal sets of metadata
- Add metadata attributes from those community standards to the MSOM, if they are not yet included
- Assign cardinality to the MSOM (identify mandatory metadata and how many times the attributes may be reported. Some metadata might not be mandatory but are still important to capture, if available)
- Identify appropriate ontologies representing your data and establish an application ontology (see recipe 4 of UC3)
- Assign, as far as possible, ontologies to the MSOM and the sparse matrix

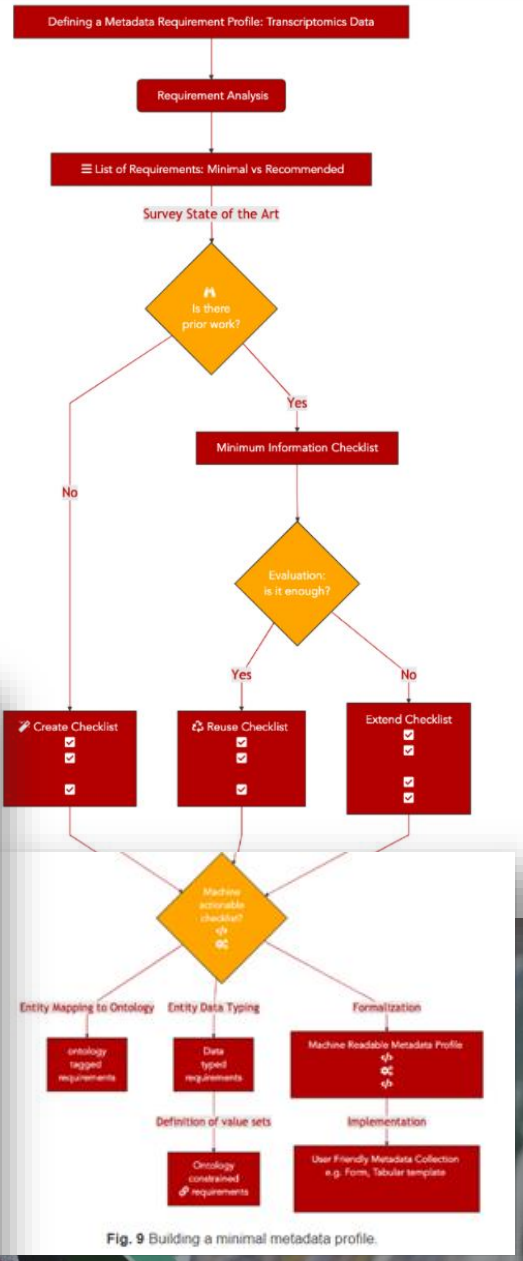
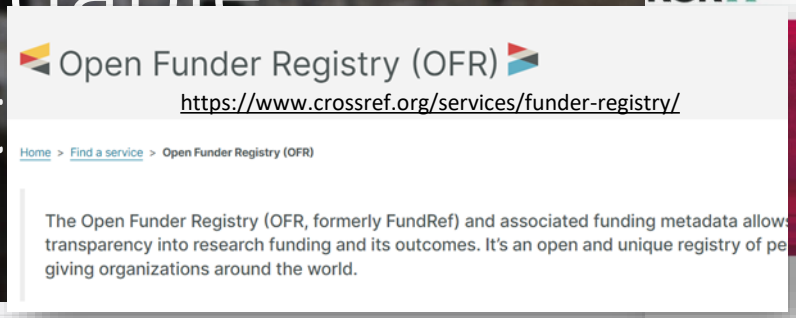
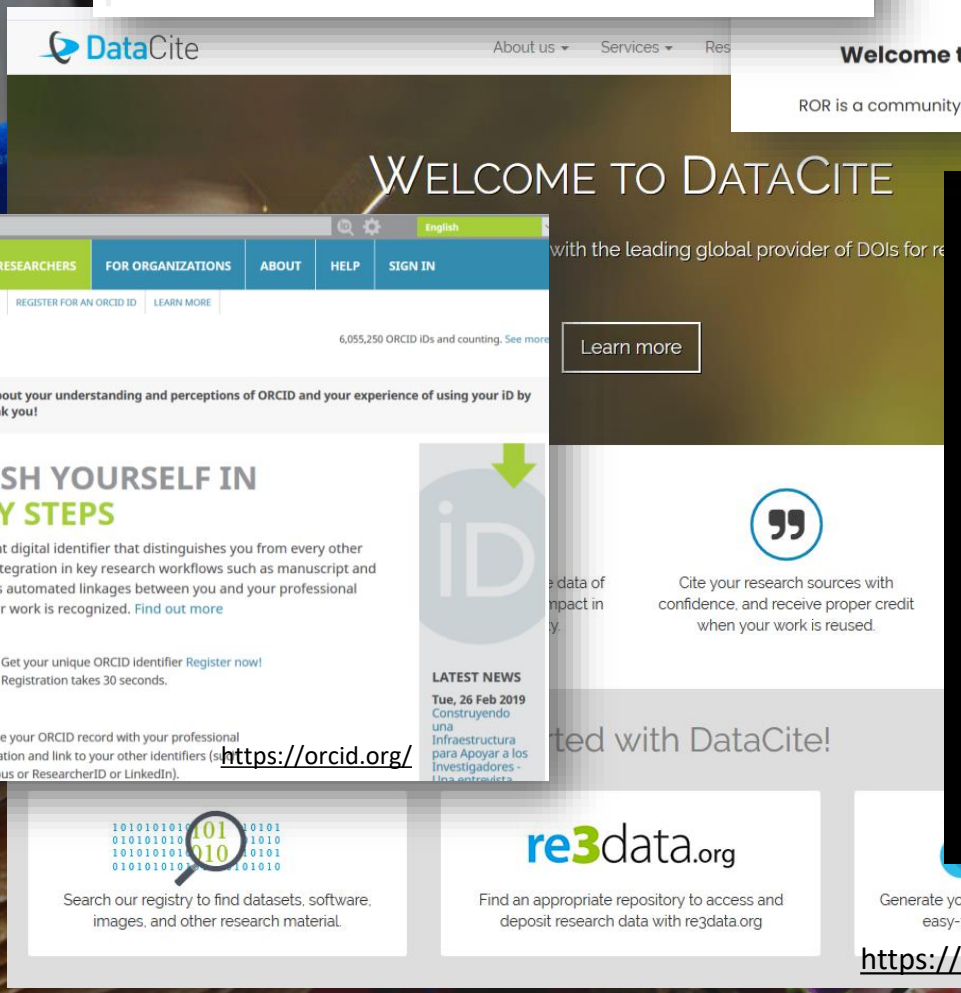


Fig. 9 Building a minimal metadata profile.

F = Findable
Persist





- THINGS: DOI DIGITAL OBJECT IDENTIFIER
- PEOPLE: ORCID
- INSTITUTIONS: ROR
- FUNDERS: OFR

ACCESSIBLE

A = Accessible

ACCESSIBLE≠OPEN
«ACCESS» CAN BE RESERVED OR RESTRICTED OR EMBARGOED

- **Open access**

  Data that can be accessed by any user whether they are registered or not. Data in this category should not contain personal information unless consent is given (see 'Informed consent').

- **Access for registered users (safeguarded)**

  Data that is accessible only to users who have registered with the archive. This data contains no direct identifiers but there may be a risk of disclosure through the linking of indirect identifiers.

- **Restricted access**

  Access is limited and can only be granted upon request. This access category is for the most sensitive data that may contain disclosive information.

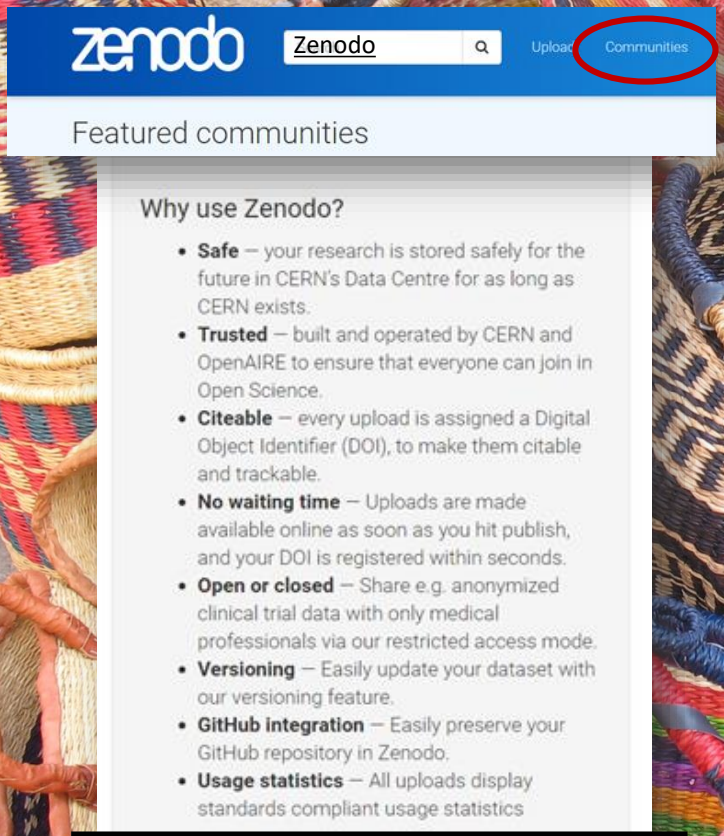
  Restricted access requires the long-term commitment of the researcher or person responsible for the data to handle the upcoming permission requests.

- **Embargo**

  Besides offering the opportunity for restricted access 'for eternity' most data repositories allow you to place a temporary embargo on your data. During the embargo period, only the description of the dataset is published. The data themselves will become available in open access after a certain period of time.

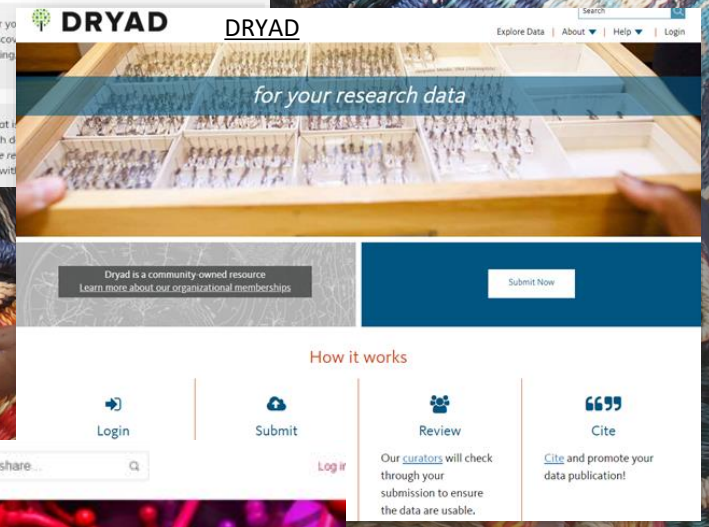
CESSDA Guide

A = Accessible – Data repositories

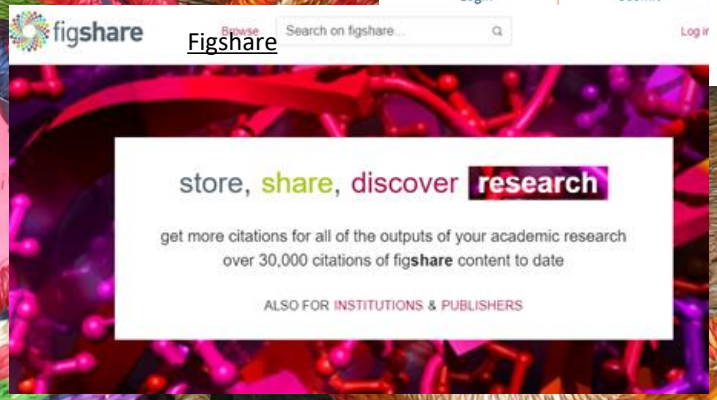


Zenodo

Dataverse



DRYAD



Figshare

YOU CAN CREATE A «COMMUNITY» [THE PROJECT?]

A = Accessible
FAIR enabling r

DataverseNL is a publicly accessible data repository platform, open to researchers of affiliated institutes and their collaborators to deposit and share research data openly with anyone. It facilitates making your research data FAIR (Findable, Accessible, Interoperable, Reusable).

DataverseNL supports the creation of custom terms of use and restrictions in order to control access to your research data. DataverseNL facilitates long-term access, persistent identifiers, and preservation by storing a backup copy for safekeeping.

Receive academic credit and recognition by making your data more discoverable to the research community online. Collaborate in teams and track changes as Dataverse provides increased user control over managing changes to a project.

What are the benefits of using DataverseNL for sharing your research data?

- Your dataset receives a DOI (a persistent identifier) and it will be easy for others to refer to your data with the provided citation information.
- You can manage access and reuse of your own data.
- DataverseNL provides safe storage for your data.
- Sharing your data increases the impact and visibility of your research.
- You can link your dataset to the related publications.
- You can meet grant requirements by depositing your research data in DataverseNL.
- You can update your stored dataset during your research, and keep track of your changes with version control.
- By using the Guestbook-feature you can check the use of your data by others. Dataverse also displays download statistics per dataset and per file.

What are the costs for depositing my data in DataverseNL?

Whether storage costs are charged to the researcher can differ per institute. You can contact your local DataverseNL contact for more information. See this list for contact information per institute.

r institutes it is possible to get a CoreTrustSeal
ification for their dataverses within DataverseNL,
Tilburg University did. A connection to the DANS
ta Vault secures Long Term Preservation of the

DATAVERSE NL ENABLES DATA FAIRIFICATION

A = Accessible – FAIR enabling repositories

PERFECT EXAMPLE OF «FAIR ENABLING» NATIONAL REPOSITORY

RAVNETRYKK 2022

UiT goes open: Et festlig skrift til Stein Høydalsvik

DataverseNO: A National, Generic Repository and its Contribution to the Increased FAIRness of Data from the Long Tail of Research

Philipp Conzett

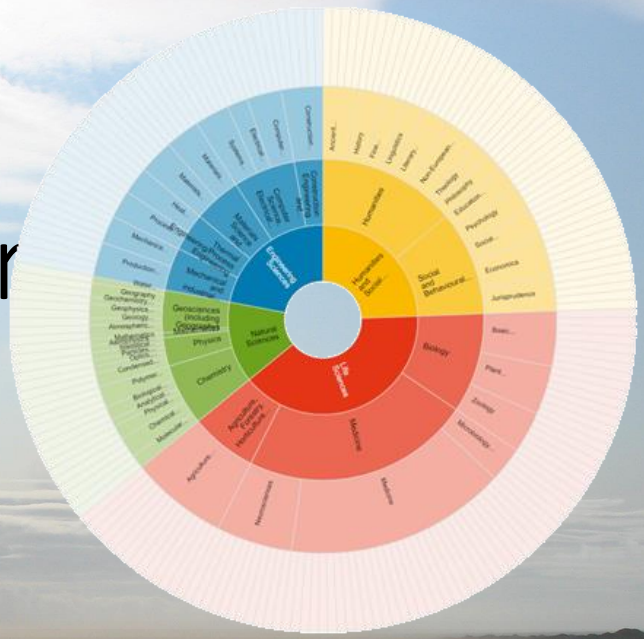
Dataverse. This article presents the organization and operation of DataverseNO, and investigates how the repository contributes to the increased FAIRness of small and medium sized research data. Sections 1 to 3 present background information about the FAIR Data Principles (section 1), how FAIR may be turned into reality (section 2), and what these principles and recommendations imply for data from the so-called long tail of research, i.e. small and medium-sized datasets that are often heterogenous in nature and hard to standardize (section 3). Section 4 gives an overview of the key organizational features of DataverseNO, followed by an evaluation of how well DataverseNO and the repository application Dataverse as such support the FAIR Data Principles (section 5). Section 6 discusses how sustainable and trustworthy the repository is. The article is rounded up in section 7 by a brief summary including a look into the future of the repository.

Principle	Implementation in Dataverse	Applied in DataverseNO
R1		
R1.1	Included in metadata: data use license/waiver; data access and use terms. But, licenses other than CC0 are not predefined and by default not machine-readable.	Yes. Almost all datasets are published under default license CC0.
	By default no support for explicit information about metadata license	Terms for reuse of metadata described on website
R1.2	Rich citation metadata including information about data authors and other contributors, providers, distributors, related data (input data)	Yes
	Versions with changes documented automatically	Yes
	W3C PROV support	No
R1.3	DDI for social science data	Partially/in some datasets
	Metadata blocks for other community standards	Partially/in some datasets
	Ongoing work on support for more domains.	No
	Custom metadata	No
	FITS for astronomy data	N/A (so far)
	File format conversion to reusable formats (tabular)	Partially/in some datasets
		Data in preferred file formats
		Datasets include ReadMe file.

Table 2: The implementation of Findability in Dataverse and its adoption in DataverseNO. Adapted from Crosas (2020).

Principle	Implementation in Dataverse	Applied in DataverseNO
F1	Support for DOI and Handle	Yes (DOI)
	Always at the dataset level	Yes
	Optionally at file level	Yes
F2	Metadata standards in human- and machine-readable formats: Dublin Core; Documentation Data Initiative (DDI); DataCite; Schema.org	Yes
	Optional custom metadata	No
F3	Dataset PID is part of metadata record presented on Dataset landing page.	Yes
	File PID is part of metadata record presented on File landing page.	Yes
	PIDs are included in exported metadata files.	Yes
F4	DataCite metadata is harvested and indexed by DataCite Search.	Yes. In addition: B2FIND and VLO.
	Schema.org metadata is indexed by Google Dataset Search.	Yes

A = Accessible.
Looking for a repositor



re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
Search... Search

https://www.re3data.org/

2,000 Data Repositories and Science Europe's Framework for Discipline-specific Research Data Management

By offering detailed information on more than 2,000 research data repositories, re3data has become the most comprehensive source of reference for research data infrastructures globally. Through the development and advocacy of a framework for discipline...

Read more

Three new DOI Fabrica features to simplify account management

Last month month we launched DOI Fabrica, the modernized version of the DataCite Metadata Store (MDS) web frontend. It is the one place for DataCite providers and their clients to create, find, connect and track every single DOI from their organization...

Read more

One step closer towards instant DOI search results

Art Art? You might be wondering, what this pink and green picture illustrates? A few months ago we couldn't show you this picture; the data that we used to created it, did not exist. And the answer to what this illustrates – this is simply a distorted...

Read more

https://www.re3data.org/

A = Accessible. Data repositories

DCC guide

Because good research needs good data

Where to keep research data

TO SHARE

TO PRESERVE

Domain/discipline-specific data repository, data centre or 'scientific database'

- Pros - most likely to offer both the specialist domain knowledge and data management expertise needed to ensure your data collection is properly kept and used
- Cons - most likely to be selective, requiring advance planning of the effort needed to meet high standards for metadata and documentation

General-purpose data repository: e.g. Dryad, Figshare, Zenodo

- Pros - most likely to offer useful search, navigation and visualisation functionality
- Cons - requires scrutiny of terms and conditions to ensure consistency with your funder, journal or institution's policies on cost recovery, copyright/IP, long-term preservation

Institutional data repository

- Pros - most likely to accept any data of value, especially if n suitable home can be found for it elsewhere, and to ensure t policy requirements for long-term access are met
- Cons - unlikely to be as well-resourced as either general-purp or domain repositories

Journal supplementary material service

- Pros - most likely to comply with the journal or publisher's requirements
- Cons - may be costly, unlikely to offer a data repository's functionality or long-term solution

Departmental, project or personal web page

- Pros - might provide functionality tailored to your data collection and/or your existing data users and peer network
- Cons - least likely to make your data collection visible to new users and contacts, or to sustain long-term access to your data collection

Institutional data archive or vault

- Pros - most likely to have considered the total costs of long-term storage, and to ensure that policy requirements for long-term access are met
- Cons - may be less likely to offer the same ease of use as third-party storage or archiving services

Safe centres or havens

- Pros - most likely to meet stringent security requirements for handling sensitive data, and to ensure that legal requirements for data protection are met
- Cons - may be less likely to offer similar levels of digital preservation as a data archiving third-party service or institutional data archive

Cloud storage third-party services

- Pros - most likely to offer easy to use file store and share functionality
- Cons - long-term reliability and costs of data retrieval may be unpredictable; terms and conditions need careful scrutiny to ensure it complies with policy requirements for long-term access and other legal requirements, e.g. a data centre location within the European Union

Data archiving third-party services

- Pros - likely to offer cost-effective long-term storage with guarantee of accessibility, including data that may not be shareable for confidentiality reasons
- Cons - less likely to offer administrative interface to manage access and preservation policies (although some services offer

A = Accessible. Data repositories

Checklist: is it the right repository for your data?

The checklist that follows addresses the five key questions posed in this guide:

1. is the repository reputable?
2. will it take the data you want to deposit?
3. will it be safe in legal terms?
4. will the repository sustain the data value?
5. will it support analysis and track data usage?

DCC checklist

CHECKLIST TO HELP YOU CHOOSING THE RIGHT REPOSITORY

Findable, accessible and interoperable

Level 1	Level 2	Level 3
Metadata publishing: Data collections are catalogued in a repository according to funder expectations so that they are discoverable by title, creator, and date of deposition ❑	Repository publishes other pertinent information as metadata fields to enhance cross-disciplinary discovery ❑	Metadata is catalogued to enhance reuse according to sector-leading standards, or to fulfil domain-specific purposes ❑
Stable identifiers: Enables a DOI or other open standard identifier to be assigned to a landing page for each ingested dataset/ collection ❑	Supports assignment of related persistent IDs per dataset/ collection ❑	Supports assignment of multiple persistent IDs at different levels of granularity within dataset/ collection ❑
Discovery metadata: Provides Datacite mandatory metadata and exposes it according to open access repository protocols ❑	Provides metadata elements to enable broader discovery (e.g. geo-spatial) to reflect best practice changes and local needs ❑	Exposes discovery metadata as Linked Open Data to optimise automatic discovery ❑
Metadata harvesting: Sufficient information can be harvested about data deposited with third-party repositories, to	Metadata can be routinely harvested with links to data producer IDs (e.g. ORCID), any grant information and	Metadata on the externally held research data is sufficiently

Legal terms and conditions

Personal data or data which may identify individuals when linked to other data should not be stored outside the European Economic Area, unless in a legal jurisdiction that ensures personal data is adequately protected ❑

By agreeing to the terms and conditions the depositor will not be breaching other **Data Protection** principles, or the terms of any confidentiality agreement with data subjects or owners (e.g. consent form, consortium agreement) ❑

By agreeing to the terms and conditions the depositor will not be in breach of **copyright**, or any contract terms covering **Intellectual Property** in the research, (e.g. the grant conditions or a consortium agreement) ❑

Anything deposited that is not publicly accessible can be retrieved by the institution in response to a valid **Freedom of Information** request ❑

Criteria for the selection of a trustworthy repository

TRUSTWORTHY REPOSITORIES

Trustworthy repositories should meet the following minimum criteria:

1. **Provision of Persistent and Unique Identifiers (PIDs)**
   a. Allow data discovery and identification
   b. Enable searching, citing, and retrieval of data
   c. Provide support for data versioning

2. **Metadata**
   a. Enable finding of data
   b. Enable referencing to related relevant information, such as other data and publications
   c. Provide information that is publicly available and maintained, even for non-published, protected, retracted, or deleted data
   d. Use metadata standards that are broadly accepted (by the scientific community)
   e. Ensure that metadata are machine-retrievable

3. **Data access and usage licences**
   a. Enable access to data under well-specified conditions
   b. Ensure data authenticity and integrity
   c. Enable retrieval of data
   d. Provide information about licensing and permissions (in ideally machine-readable form)
   e. Ensure confidentiality and respect rights of data subjects and creators

4. **Preservation**
   a. Ensure persistence of metadata and data
   b. Be transparent about mission, scope, preservation policies, and plans (including governance, financial sustainability, retention period, and continuity plan)

A = Accessible. Data journals

Data journals list

Title	URL	Charge	Notes for authors (N.B. we suggest checking in particular for policy on submission of data already published)	Publisher	Notes on Subject Area
Journal of Open Archaeology Data	http://openarchaeologydata.metajnl.com/		http://openarchaeologydata.metajnl.com/about/submissions	Ubiquity Press	Archaeology
Open Health Data	http://openhealthdata.metajnl.com/		http://openhealthdata.metajnl.com/about/submissions#authorGuidelines	Ubiquity Press	Public Health
Journal of Open Psychology Data	http://openpsychologydata.metajnl.com/		http://openpsychologydata.metajnl.com/about/submissions#onlineSubmissions	Ubiquity Press	Psychology
			nature.com/sdata/for-authors nature.com/sdata/for-authors#data-deposition	Nature	"open to submissions from a broad range of scientific disciplines, but

UCL Home » / Open@UCL Blog » / Data journals and data reports – don't miss out

Data journals and data reports – don't miss out on this useful publishing format!

Aug. 2021

By Kirsty, on 17 August 2021

Guest post by James Houghton – Research Data Support Officer

Why not publish a data report article?

Publishing with a data journal offers several benefits. First, a data report article is more formal than a publication of data files in a repository and is a peer reviewed publication which then contributes to a researcher's publication record which is important for CVs and advancement for many. Second, they allow a more detailed explanation of a dataset and any analysis or code related to it than is usually otherwise possible. Third, the appearance of an article in a recognised journal can help to drive visibility of a dataset for other researchers. In practice it my often be the case that a repository will be used to host material which is discussed at length in a paper.

elsevier.com/journals/genomics-data/2213-5960/guide-for-

Dataset Description

Object Name

- *walkers* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for records made by individual walkers during stage-one fieldwalking.
- *counts* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for pot-sherds countedduring stage-one fieldwalking.
- *pottery* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for the main pottery database, assembled various artefact specialists.
- *petrography* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for those sherds sampled for thin section petrography.
- *lithics* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for the main lithics database.
- *other* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for the main database of all non-ceramic and non-lithic finds.
- *structs* – three files providing the data, metadata and field type definitions (.csv, .txt, .csvt respectively) for the main database of all standing remains, except for terraces.
- *coast* – a vector polygon dataset (.shp and associated files) with the shape of Antkythera's coastline.
- *geology* –a vector polygon dataset (.shp and associated files) with the main bedrock units on Antkythera.
- *tracts* – a vector polygon dataset (.shp and associated files) with the main stage-one survey units.
- *grids* – a vector polygon dataset (.shp and associated files) with the main stage-two survey units.
- *terraces* – vector line dataset (.shp and associated files) with all observable agricultural terraces (i.e. the location

Data journals

Panayiota Polydoratou

Alexander Technological Educational Institute of Thessaloniki

European Commission Workshop
Alternative Open Access Publishing Models: Exploring New Territories in Communication
Brussels, 12 October 2015

Reposit
UK Ar
10.5284

Publica
05/02/2012

Language
English (a Greek language summary of the project methods and results can be found at www.ucl.ac.uk/asp/ or www.tuarc.trentu.ca/asp/).

License
Creative Commons CC-BY 3.0

Reuse Potential

Due to their unusual coverage of an entire landscape, these datasets would provided a good basis for developing a tutorial on survey, GIS and/or spatial analysis in archaeology. They also lend themselves to the comparative analysis of evidence from other intensive Mediterranean surveys that are in the public domain (e.g. http://dx.doi.org/10.5284/1000271, http://dx.doi.org/10.5284/1000208, http://dx.doi.org/10.5284/1000103 and, to a lesser extent, also http://dx.doi.org/10.5284/1000351), albeit with due attention to the fact that the intensive methods used are not identical. The ASP data is particularly reusable because artefact locations, dates and identifications are recorded individually in the database rather than in aggregate. The standing structures and terraces from Antikythera are also the kinds

A = Accessible. Formats

Data Archiving and Networked Services
DANS

HOME DEPOSIT

If your data are stored in other formats than those mentioned below, please contact DANS.

Type	Preferred format(s)	Non-preferred format(s)
Text documents	• PDF/A (.pdf) • ODT (.odt)	• Microsoft Word (.doc) • Office Open XML (.docx) • Rich Text File (.rtf) • PDF other than PDF/A (.pdf)
Plain text	• Unicode text (.txt)	• Non-Unicode text (.txt)
Markup language	• XML (.xml) • HTML (.html) • Related files: .css, .xslt, .js, .es	• SGML (.sgml) • Markdown (.md)
Programming languages	• MATLAB • NetCDF • Text-Fabric • Python	
Spreadsheets	• ODS (.ods) • CSV (.csv)	• Microsoft Excel (.xls) • Office Open XML Workbook (.xlsx) • PDF/A (.pdf)

https://dans.knaw.nl/en/file-formats/

A – Accessible - Formats

NationalArchives

NATIONAL ARCHIVES

Appendix A: Tables of File Formats

Quick Links

Computer Aided Design	Digital Audio	Digital Moving Images
Digital Cinema	Digital Video	
Digital Photographs	Scanned Tex	
Geospatial Formats	Presentation	
Structured Data Formats	Email	
Calendars	Navigational	

Symbol Key

Preferred Formats

Acceptable Formats

Geospatial Formats

Geospatial records include digital cartographic data files and aerial photography that are created and processed in Geographic Information Systems (GIS) or other soft applications for spatial analysis.

Preferred Formats

Preferred Formats	Format Versions	Format Specifications
Geospatial Tagged Image File Format	1.8.2	Geo TIFF Format Specification: (http://geotiff.maptools.org/spec/geotiffhome.html)
Geographic Markup Language	2.0 through 3.2	ISO 19136:2007 & Version 3.2, OGC document 07-036: (http://www.opengeospatial.org/standards/is)
Topologically Integrated Geographic Encoding and Referencing Files	2006 Second Edition	2006 Second Edition TIGER/Line®: (https://www.census.gov/programs-surveys/geography/technical-documentation/complete-technical-documentation.html)
Keyhole Markup Language	2.2	Open Geospatial Consortium Inc. OGC 07-147r2: (http://www.opengeospatial.org/standards/kml/)

Acceptable Formats

Acceptable Formats	Format Versions	Format Specifications
Vector Product Format		MIL-STD-2407: (http://earth-info.nga.mil/publications/specs/printed/2407/2407_VPF.pdf)
ESRI ARC/INFO Interchange File Format		Reverse engineered specification: (http://avce00.maptools.org/docs/v7_e00_cover.html)
TerraGo Geospatial PDF	GeoPDF Encoding Best PracticeVersion 2.2	Open Geospatial Consortium Inc. OGC 08-139r2: (http://www.opengeospatial.org/standards/is)
ESRI Shapefile (Compound)	1997 – current version	ESRI Shapefile Technical Description: (http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf)

Acceptable for Imminent Transfer Formats

A = Accesible - Formati

FAIRCOOKBOOK

GITHUB

ORD
ction
cing the FAIR
es
ng on the ethical
of FAIR
cing our FAIRification
ork
ing projects for
ation
g FAIR and the notion
data
tanding the relation
n FAIR and Knowledge
g for FAIRification with
synthetic biomedical
s
Awareness in Public

12. Converting from proprietary to open format

FAIR cookbook converting

Recipe Overview

Reading Time
20 minutes

Executable Code
Yes

Difficulty

Converting from proprietary to open format

Recipe Type
Hands-on

Audience
Principal Investigator, Data Manager, Data Scientist

Maturity Level & Indicator
DSM-3-R4

Cite me with FCB029

12.3. FAIRification Objectives, Inputs and Outputs

Actions.Objectives.Tasks	Input	Output
formatting	Waters MS format	mzML
text annotation	PSI-MS	annotated text

12.4. Table of Data Standards

Data Formats	Terminologies	Models
mzML	PSI-MS	

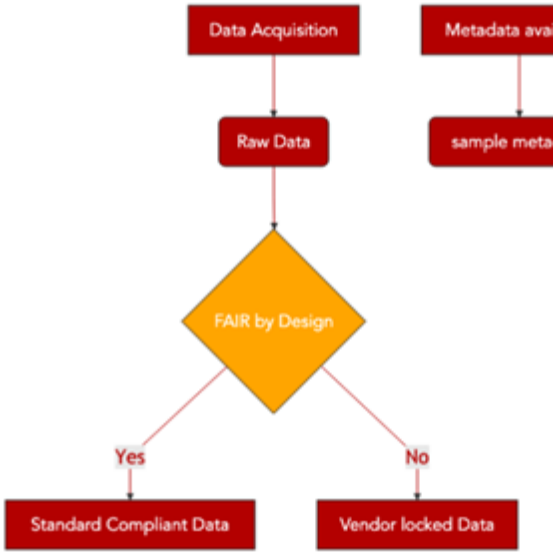


Fig. 12.1 Converting to an open standard file format.

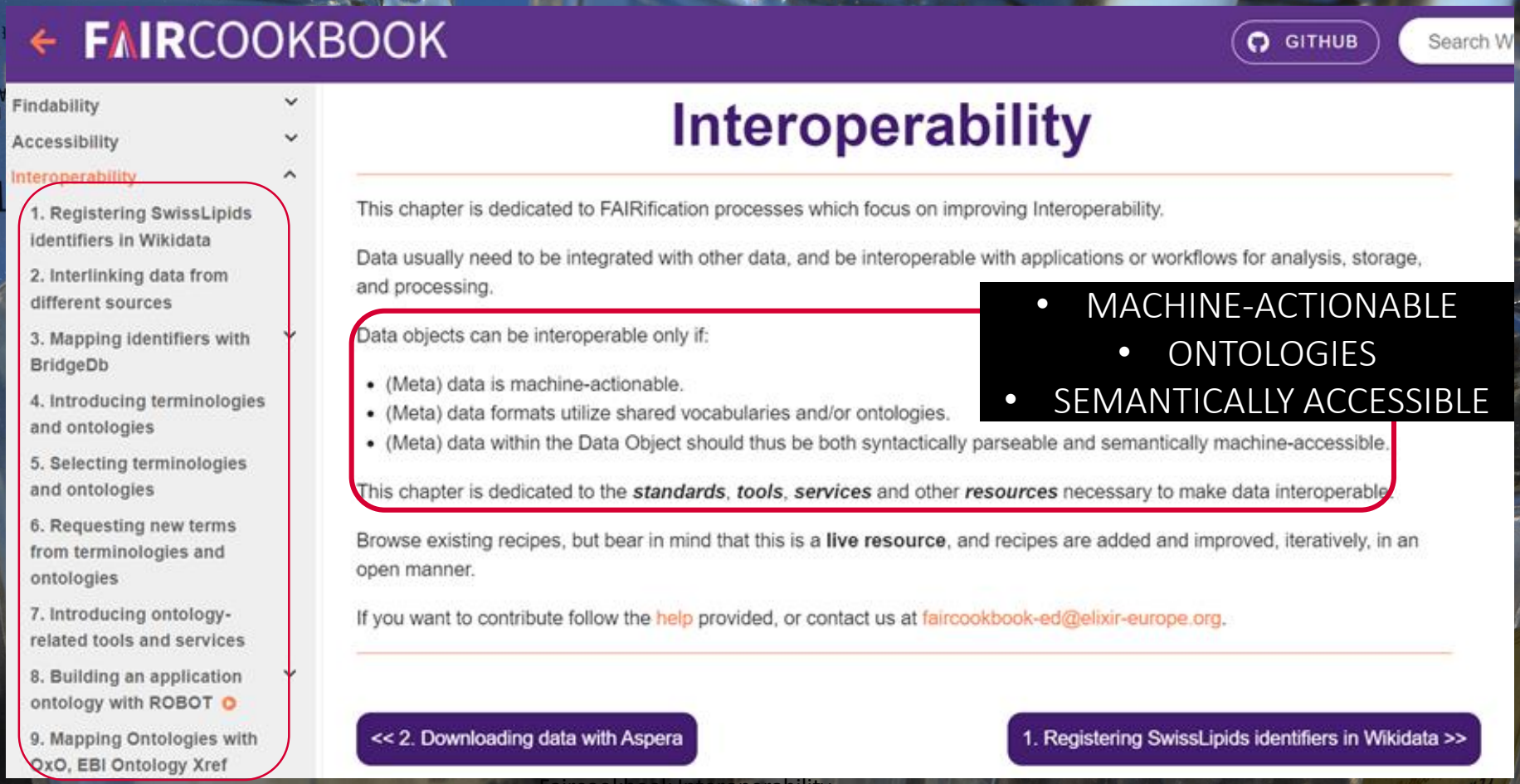
12.5. Ingredients

Tools and Software:

- github
- docker
- python

INTEROPERABLE

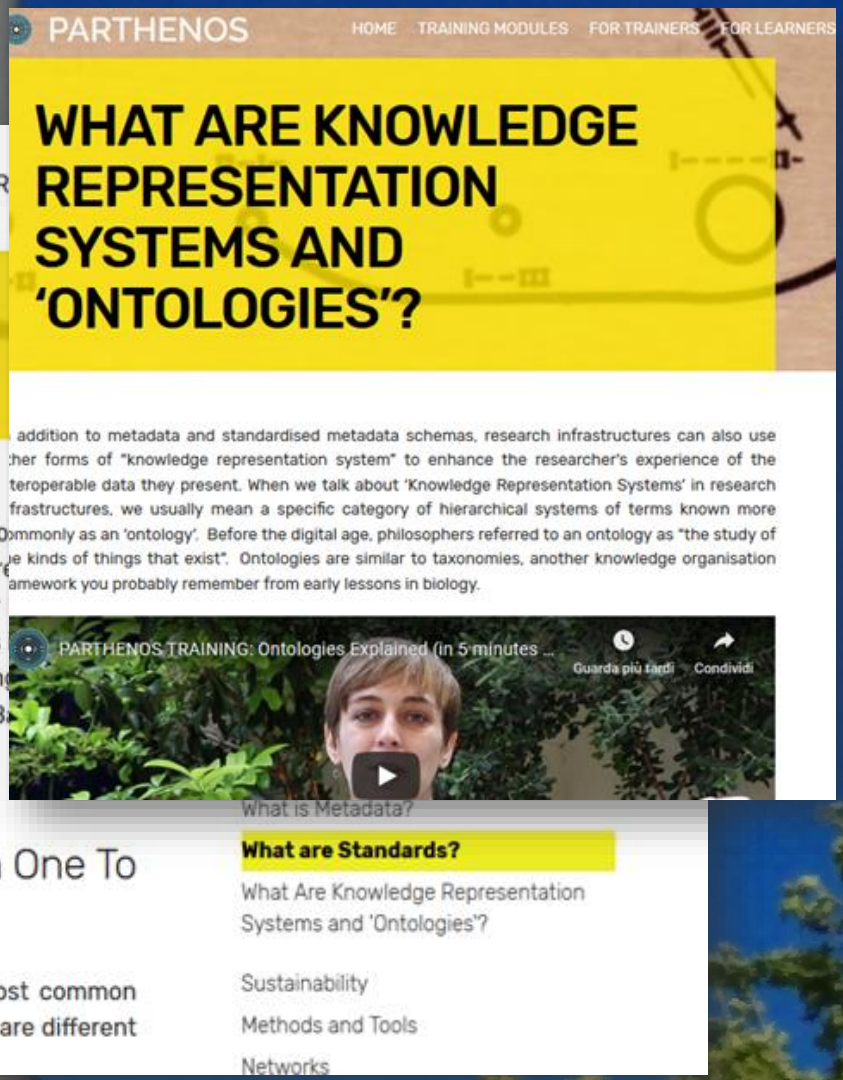
I = Interoperable



- MACHINE-ACTIONABLE
- ONTOLOGIES
- SEMANTICALLY ACCESSIBLE

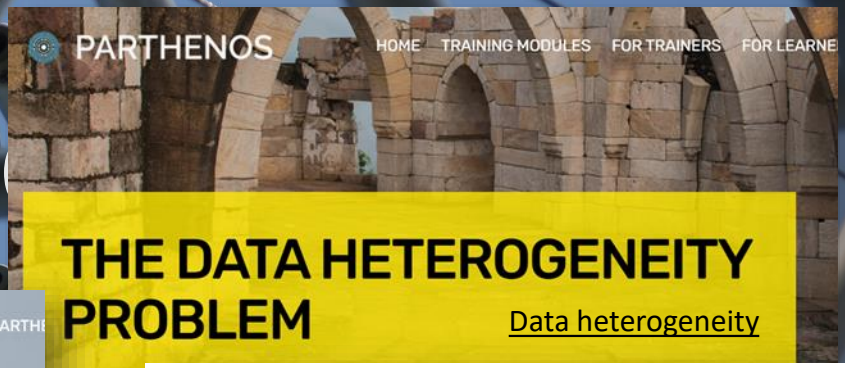
Faircookbook Interoperability

I = Interoperable. Standards

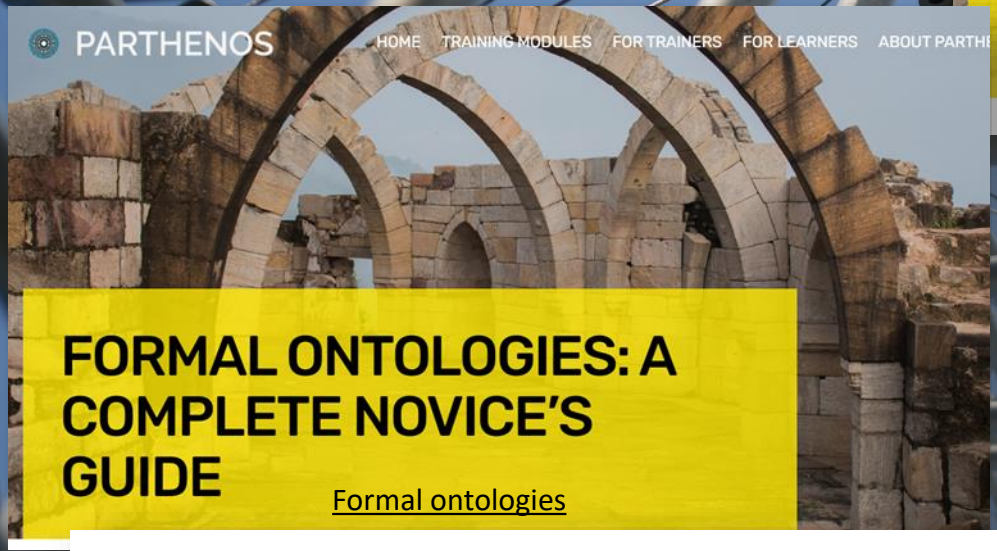


http://training.parthenos-project.eu/sample-page/intro-to-ri/interoperability/what-are-standards/

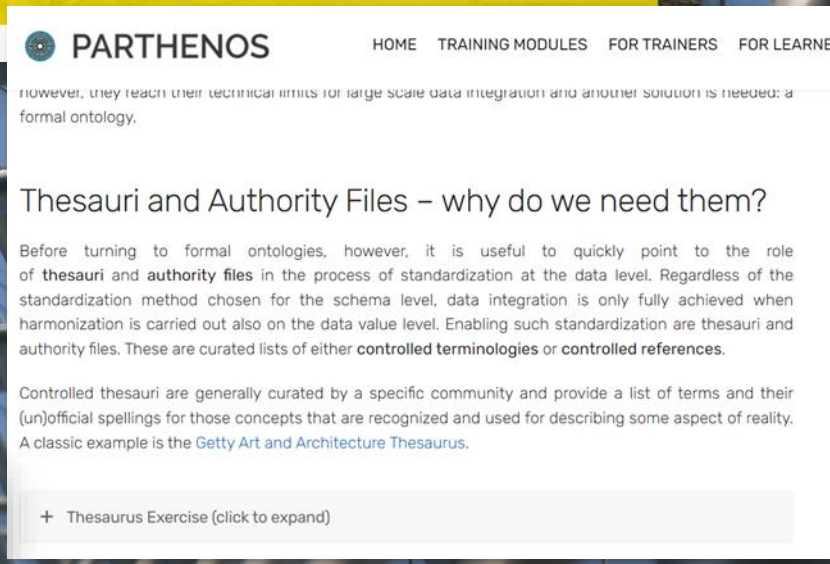
I = Inteoperable.



Data heterogeneity



Formal ontologies



Formal ontologies are proposed as a solution to the data heterogeneity problem because they describe broad ranges of data in a manner that is intellectually consistent and able to cover both general and particular levels of knowledge. Put another way, a formal ontology attempts to build a schema that applies to anyone and no one in a particular field.

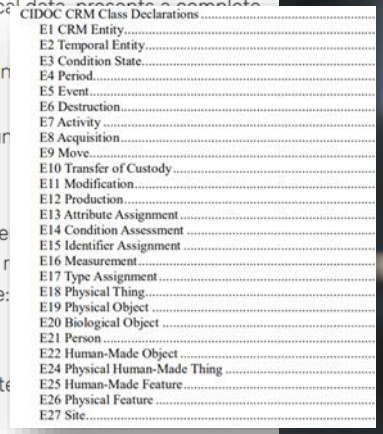
A successful formal ontology must:

- accurately represent the most common information points of interest to a part domain and the relationships that users want to trace between entities.
- offer sufficient abstract classes and relations in order to allow the representati characteristic states of lack of knowledge. That it to say, create information stru which allow representing not just highly accurate information but also more gene uncertain information.

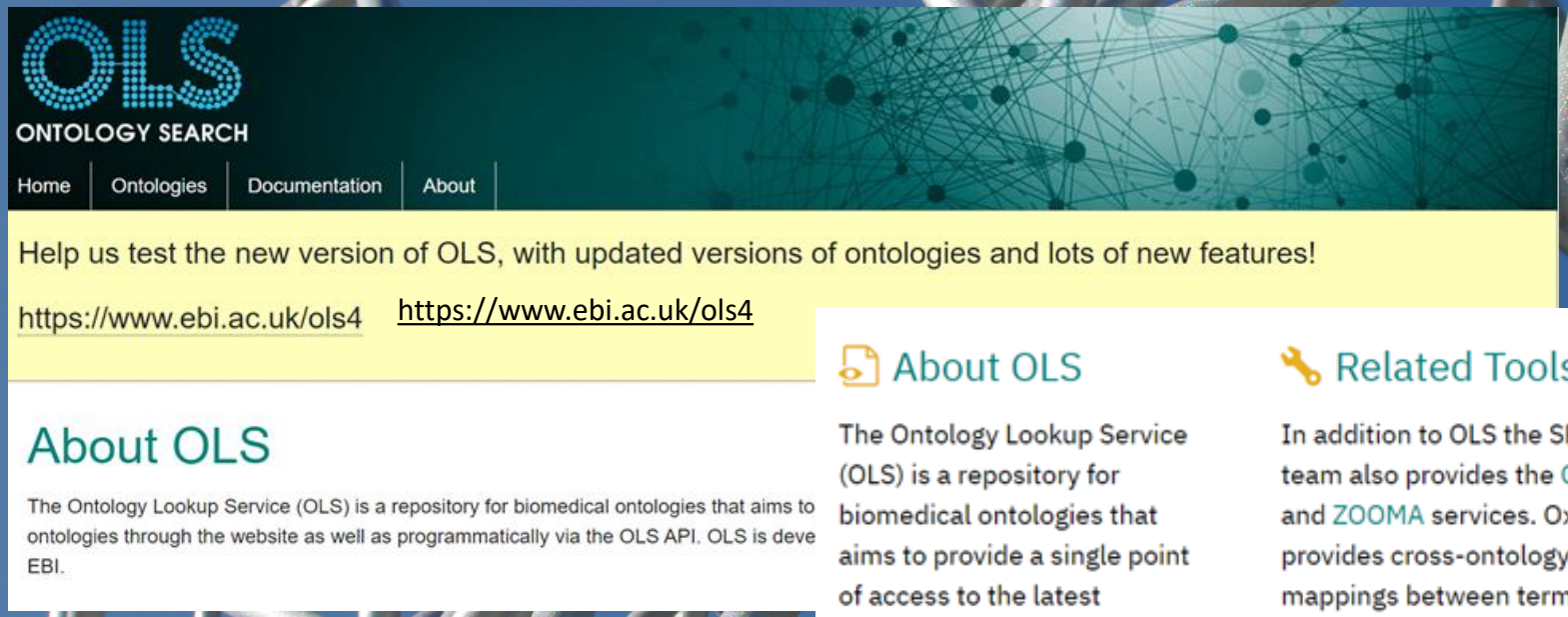
- Basic Formal Ontology: originally used in modelling of medica methodology for data modelling
- DOLCE: was constituted to aid in modelling common sense n language
- CIDOC CRM: originally designed in the museological commun to account for cultural heritage and e-sciences data

On the other hand, other ontologies are designed to address ve ignoring the general aim of interoperability in favour of a more r a problem level. Examples of such focussed ontologies include:

- FOAF: an ontology for tracking social relations
- SPAR: for organizing citation data, article structure and conte
- NeMO: for tracking scholarly process



I = Inteoperable. Ontologies



OLS – ONTOLOGY LOOKUP SERVICE FOR BIOMEDICAL FIELDS

About OLS

The Ontology Lookup Service (OLS) is a repository for biomedical ontologies that aims to provide a single point of access to the latest ontology versions. You can browse the ontologies through the website as well as programmatically via the OLS API. OLS is developed and maintained by the Samples, Phenotypes and Ontologies Team (SPOT) at EMBL-EBI.

Related Tools

In addition to OLS the SPOT team also provides the OxO and ZOOMA services. OxO provides cross-ontology mappings between terms from different ontologies. ZOOMA is a service to assist in mapping data to ontologies in OLS.

I = Inteoperable. Ontologies

Opscidia

Opscidia's ontology generator

Written on 03 March 2021.

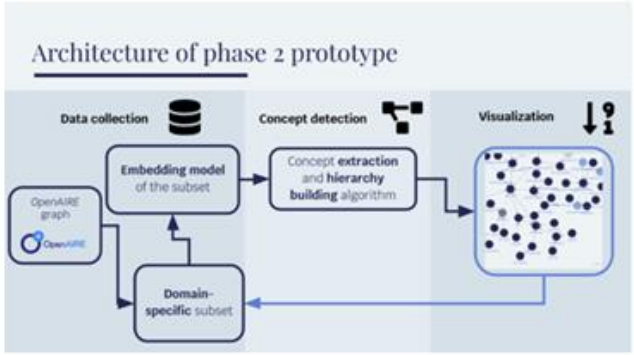
OPSCIDIA

The value of knowledge

ONTOLOGY GENERATOR

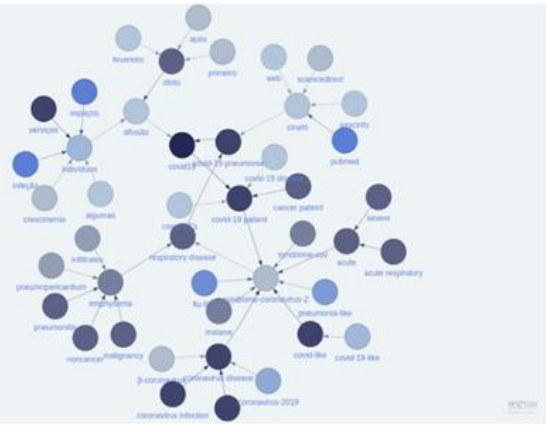
The solution proposed by Opscidia is an ontology generator that consists in three layers:

- **Data collection layer**: here it consists mostly in harvesting the resources (API or Dumps of specific OpenAIRE communities)
- **Concept detection layer**: A simple, unsupervised algorithm extracts and hierarchizes concepts related to seed concept entered by the user. It can easily scale-up both with the amount of data and with the amount users / requests.
- **Visualization layer**: A visualization tool represents graphically the produced ontology and links it back to t documents of the corpus from which the ontology was created.

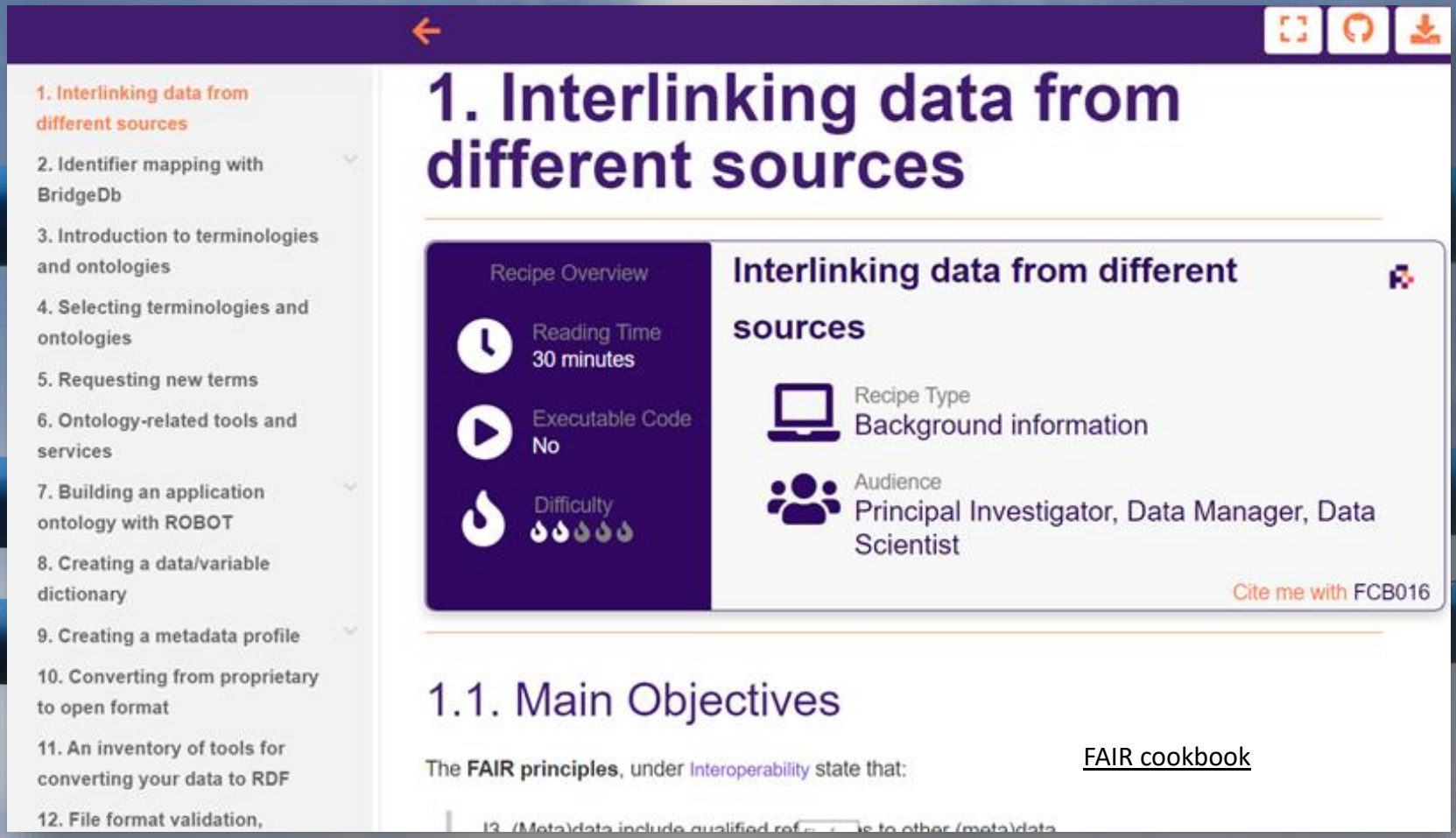


The results of the Ontology Generator

A simple tool for semi-automatic domain specific ontology creation has been built.
It takes a concept as an input and extracts from a subset of OpenAIRE graph a hierarchical list of concepts associated with the user input. This list is displayed using a simple visualization layer and linked back to the scientific literature through OpenAIRE graph.

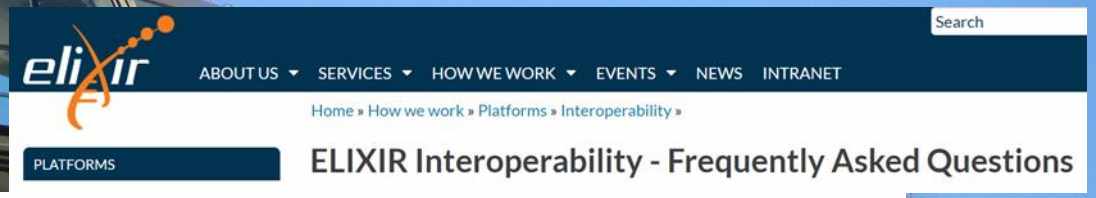


I = Interoperable



FAIR cookbook

I = Interoperable

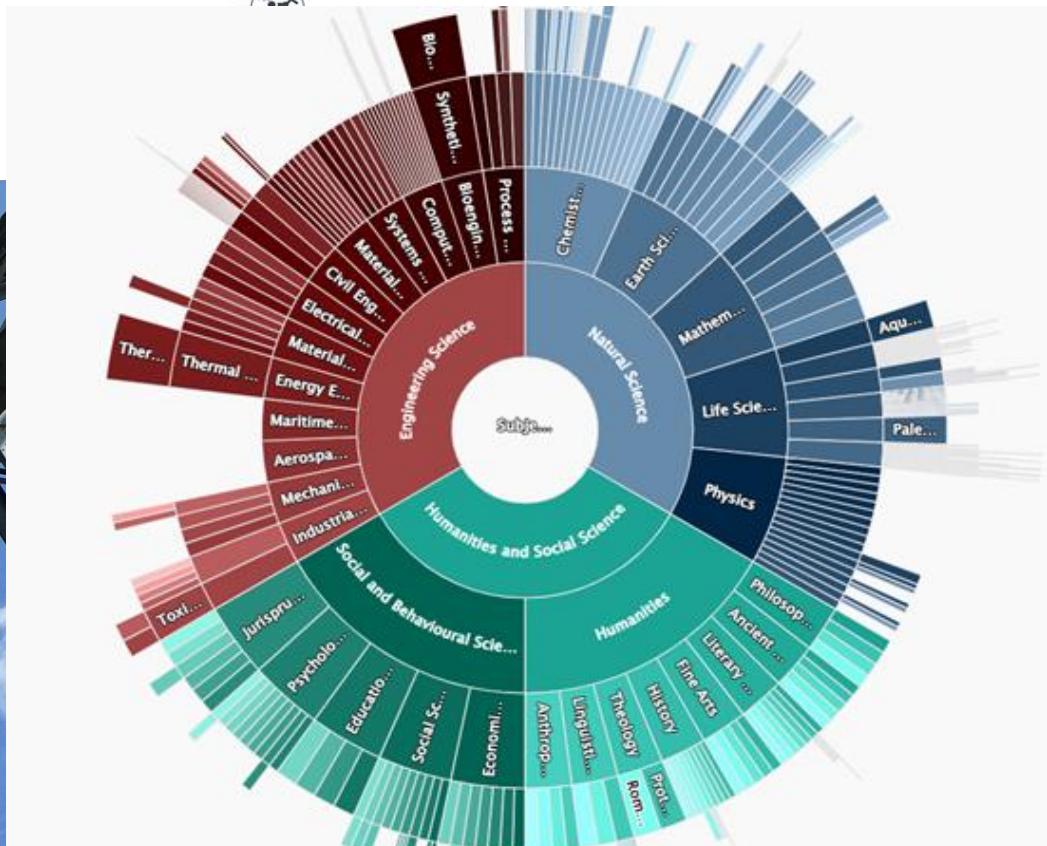


Elixir

- What is the purpose of the ELIXIR Interoperability Platform (EIP)?
- Where can I find information of the past EIP F2F meetings?
- What is an RIR?
- How does a service become an RIR?
- Where can I find information on ELIXIR data management and open science strategy?
- How do I FAIRify my data?
- How can I use Bioschemas?
- What is BrAPI?
- How do I get advice on ontologies?
- Is there a repository of ontology mappings I can donate my mappings to and/or reuse previously mapped identifiers?
- How can FAIRsharing help me?
- How can I get in touch with a specific Interoperability Platform service support team and learn how to use it?
- What is the phone number to speak to an Interoperability Human?
- Can the Interoperability Platform get a Core Data Resource to change its metadata model?

[FAIRsharing. To be interoperable]

https://fairsharing.org/



FAIRSHARING [NEW VERSION] STANDARD REGISTRY

REUSABLE

R = Reusable. Documentation

DOCUMENTATION (README FILE) TO
- AVOID MISUSE/MISINTERPRETATION
- KEEP INTEGRITY

CESSDA guide

cessda TRAINING

Project-level documentation

Project-level documentation explains the aims of the study, what the research questions/hypotheses are, what methodologies were being used, what instruments and measures were being used, etc. In the accordion the questions which your project-level documentation should answer are stated in more detail:

- ⊕ 1. For what purpose was data created
- ⊕ 2. What does the dataset contain
- ⊕ 3. How was data collected
- ⊕ 4. Who collected the data and when
- ⊕ 5. How was the data processed
- ⊕ 6. What possible manipulations were done to the data
- ⊕ 7. What were the quality assurance procedures
- ⊕ 8. How can data be accessed

Data-level documentation

Data-level or object-level documentation provides information at the level of individual objects such as pictures or interview transcripts or variables in a database. You can embed data-level information in data files. For example, in interviews, it is best to write down the contextual and descriptive information about each interview at the beginning of each file. And for quantitative data variable and value names can be embedded within the data file itself.

⊖ Quantitative data

Variable-level annotation should be embedded within a data file itself. If you need to compile an extensive variable level documentation that can be created by using a structured metadata format.

Data-level documentation for quantitative data

For quantitative data document the following:

- **Information about the data file**
  Data type, file type and format, size, data processing scripts.
- **Information about the variables in the file**
  The names, labels and descriptions of variables, their values, a description of derived

R = Reusable. Documentation

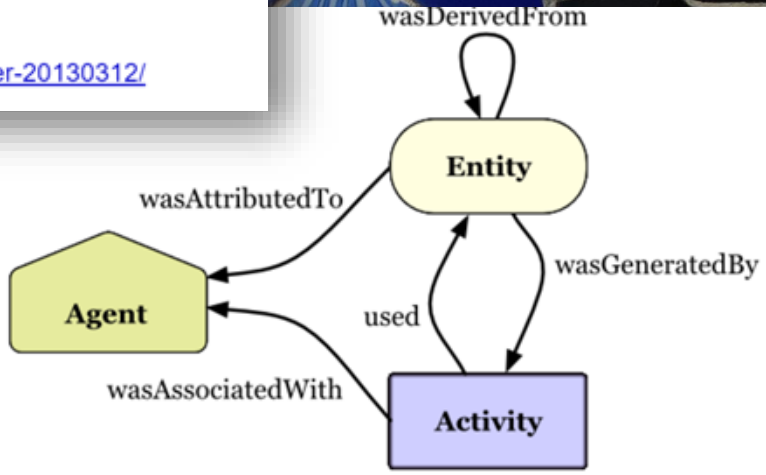
STANDARD FOR PROVENANCE

PROV Model Primer

PROV

W3C Working Group Note 30 April 2013

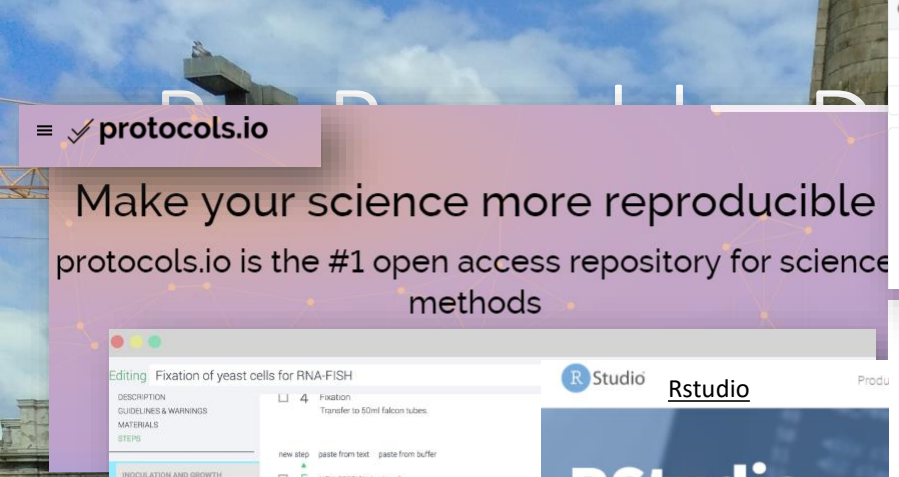
This version:
http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/
Latest published version:
http://www.w3.org/TR/prov-primer/
Previous version:
http://www.w3.org/TR/2013/WD-prov-primer-20130312/
Editors:

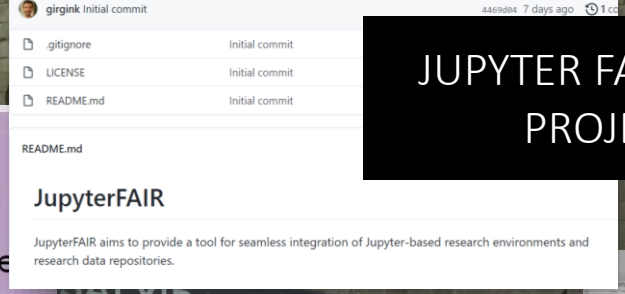


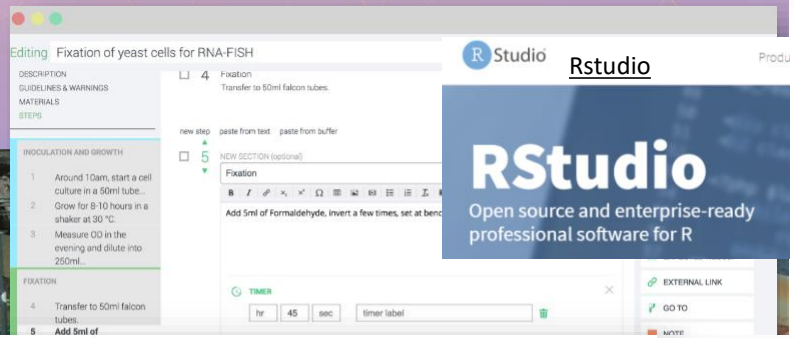
2.1 Entities

In PROV, physical, digital, conceptual, or other kinds of thing are called *entities*. Examples of such entities are a web page, a chart, and a spellchecker. Provenance records can describe the provenance of entities, and an entity's provenance may refer to many other entities. For example, a document D is an entity whose provenance refers to other entities such as a chart inserted into D, and the dataset that was used to create that chart. Entities may be described as having different attributes and be described from different perspectives. For example, document D as stored in my file system, the second version of document D, and D as an evolving document, are three distinct entities for which we may describe provenance.

JUPYTER FAIR, NEW PROJECT





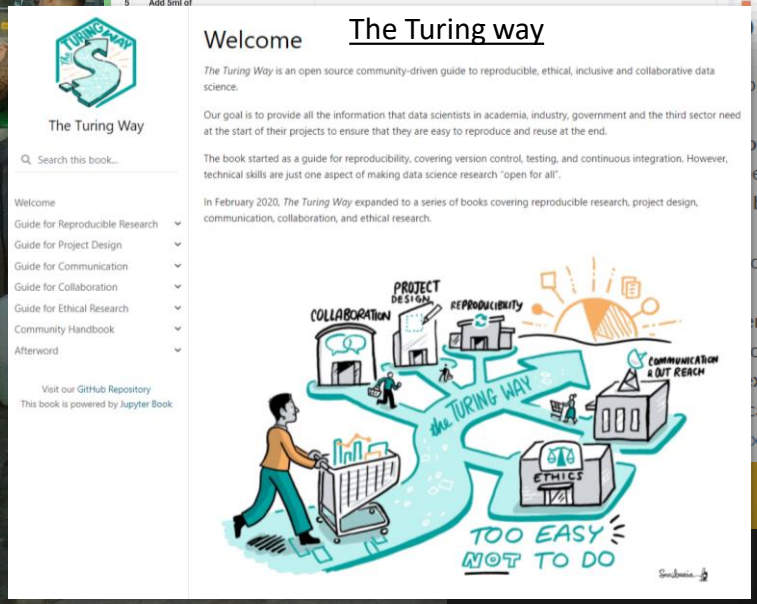


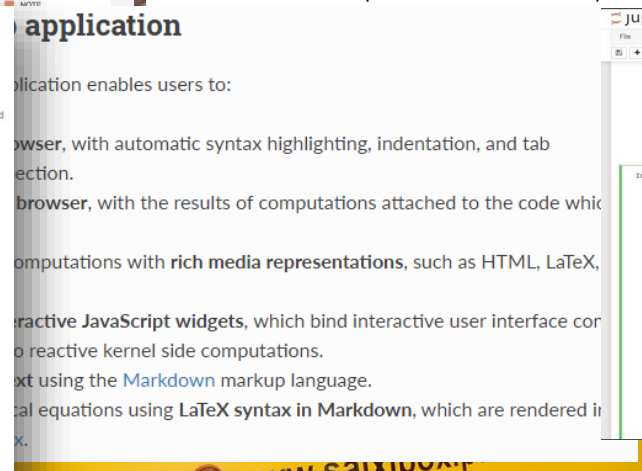
Jupyter

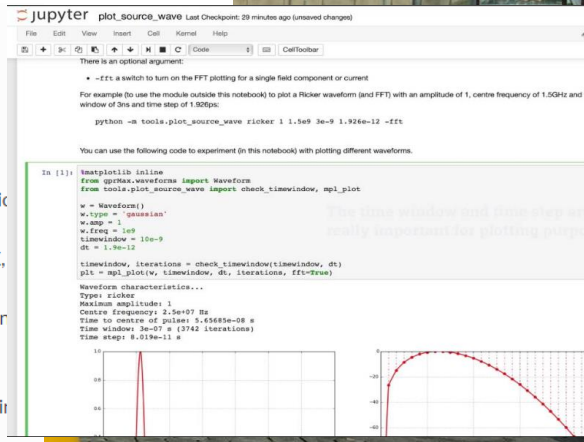
What is an Open Notebook?

Open Notebooks are documents that contain equations, visualisations, narrative text and live code that can be executed independently and interactively, with output visible immediately beneath the input.

They bring together analysis descriptions and results, which can be executed to perform the data analysis in real time.







...WHY NOT?

- PROTOCOLS.IO TO DEPOSIT YOUR METHODS
- OPEN LAB NOTEBOOK TO TRACK ANYTHING YOU DO

[TIME CONSUMING THE FIRST TIME, THEN...]

R= Reusable. Licenses

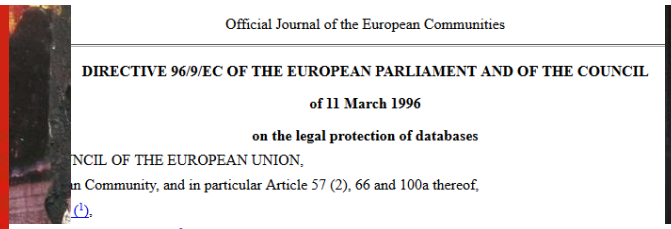
Copyright: protects the STRUCTURE, selection or arrangement of their contents" (Art. 3) NOT THE DATA

Sui generis database right: protects the «substantial effort» in OBTAINING data [NOT «CREATING»]… the right owner often is the institution

KEEP CALM AND

REMEMBER:
RAW DATA ARE NOT PROTECTED BY COPYRIGHT

Database=a collection of independent works, data or other materials arranged in a systematic or methodical way (Art.1)


Official Journal of the European Communities

DIRECTIVE 96/9/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

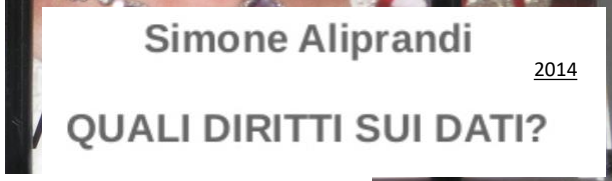
of 11 March 1996

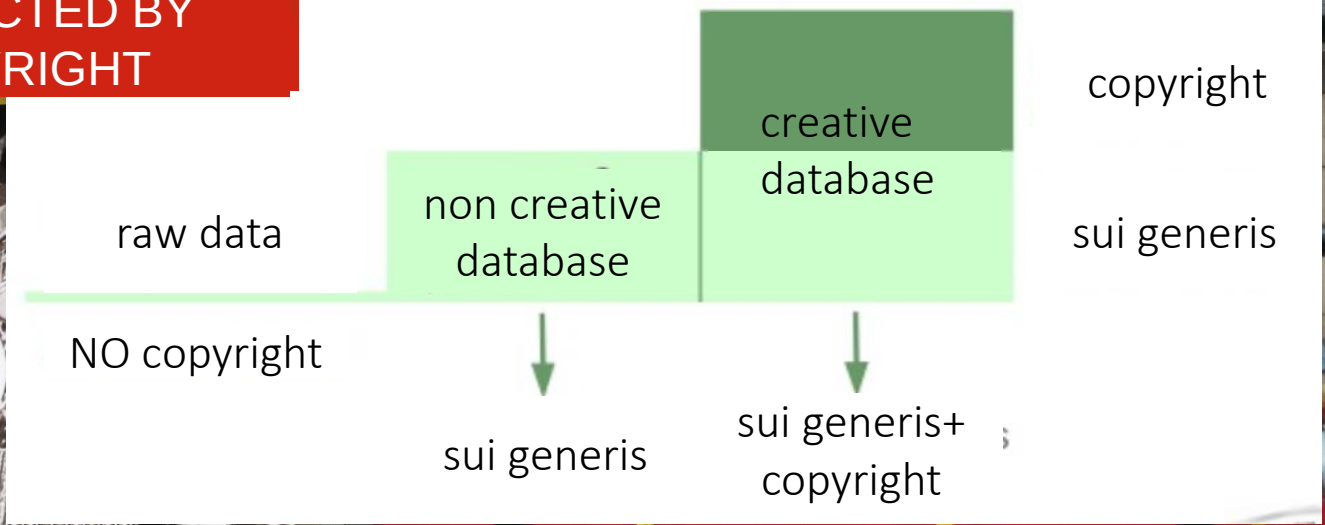
on the legal protection of databases

NCIL OF THE EUROPEAN UNION,

n Community, and in particular Article 57 (2), 66 and 100a thereof,

(1),





R = Reusable – Legal aspects

1. THE PROTECTION OF DATA, DATA SETS AND DATABASES

European Union (EU) law defines "databases", but not data sets or, at least for copyright purposes, data. Databases that meet the legal definition can be protected by copyright i they are original. Data sets, if they correspond to the definition of database, are protected by copyright otherwise not. Data as such are normally excluded from copyright protection [2,3]. It is important to understand that copyright protects original expressions in the "literary and artistic" domain, an expression that has historically included works such as books, musical works, choreographies, cinematographic works, drawings, etc [4]. Ideas, procedures, methods of operation or mathematical concepts as such, news of the day and miscellaneous facts are excluded from copyright protection [4,5,6].

MIT Press Direct 2020

Data Intelligence

Volume 2, Issue 1-2
Winter-Spring 2020

January 01 2020

Licensing FAIR Data for Reuse

Ignasi Labastida, Thomas Margoni

Author and Article Information

Data Intelligence (2020) 2 (1-2): 199–207.
https://doi.org/10.1162/dint_a_00042

Cite PDF Permissions Share

Previous Article Next Article

Article Contents

Abstract

The last letter of the FAIR acronym stands for Reusability. Data and metadata should be made available with a clear and accessible usage license. But, what are the choices? How can researchers share data and allow reusability? Are all the licenses available for sharing content suitable for data? Data can be covered by different layers of copyright protection making the relationship between data and copyright particularly complex. Some research

[webinar]

2020

Access to Scientific Information and Knowledge: A Matter of Democracy

Ludovica Paseri[0000-0002-5818-7969]

CIRSFID, University of Bologna, Via Galliera 3, 40121 Bologna, Italy

LEGAL IMPLICATIONS

2021

January 29, 2021

Project deliverable | Open Access

EOSC-Pillar D4.1 Legal and Policy Framework and Federation Blueprint

Foggetti, Nadina; Gerin Laslier, Maryvonne; Di Giorgio, Sara; Haile Gebreyesus, Netsanet; Müller, Sabine; van Nieuwerburgh, Inge; Romier, Geneviève; Van Wezel, Jos

Development of EOSC is influenced by the parallel development at the national and regional levels. Requirements for open data, data protection and cross border data access rely on a common understanding of existing regulations procedures in countries and their differences.

This deliverable presents the legal and organisational aspects of services delivery in a federated environment and recommends actions that enable service providers to position their services for improved interoperation in the context of the EOSC services landscape. The objectives of this deliverable are:

- a study of the legal and policy state of the art in the involved countries, highlighting commonalities to be leveraged and gaps or challenges to be tackled in order to help harmonise and improve the national policies and strategies related to FAIR data and Open Science,
- proposing recommendations for the rules and procedures with respect to legal issues regarding open access and open data,
- proposing policy recommendations for services management, focusing on the management of service level agreements, and
- delivering a blueprint for EOSC which can be used by service providers as a guideline for legal aspects of service and data provisioning in a European and an international context

The document sketches a policy and legal framework by building upon the existing national policies, delivers recommendations, and considers the aspects that come with agreement on service delivery in a federated IT landscape. These can help to establish a governance structure for service providers and other organisations that handle scientific data.

- NO COPYRIGHT BUT THERE MIGHT BE OTHER LEGAL PROTECTION
- UNDER GDPR, IF YOU DEAL WITH SENSITIVE DATA YOU ALWAYS MUST STATE THE LEGAL GROUND OF YOUR RESEARCH

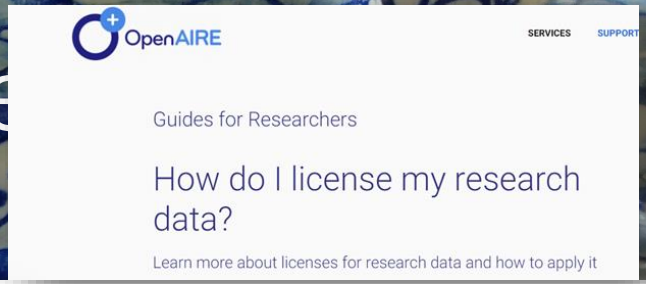
2020

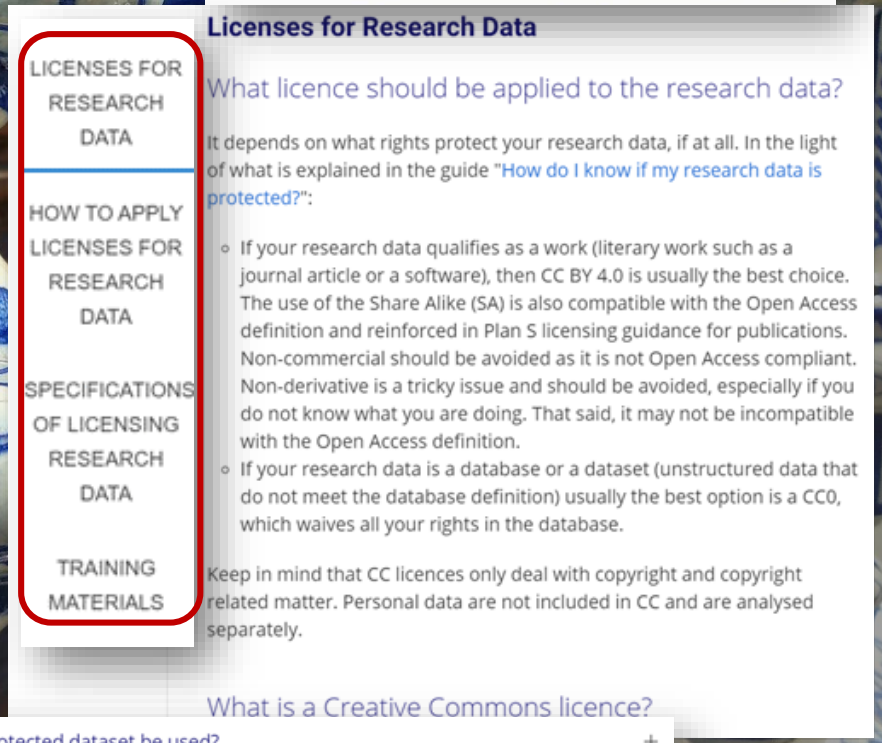
OpenAIRE Legal Policy Webinars

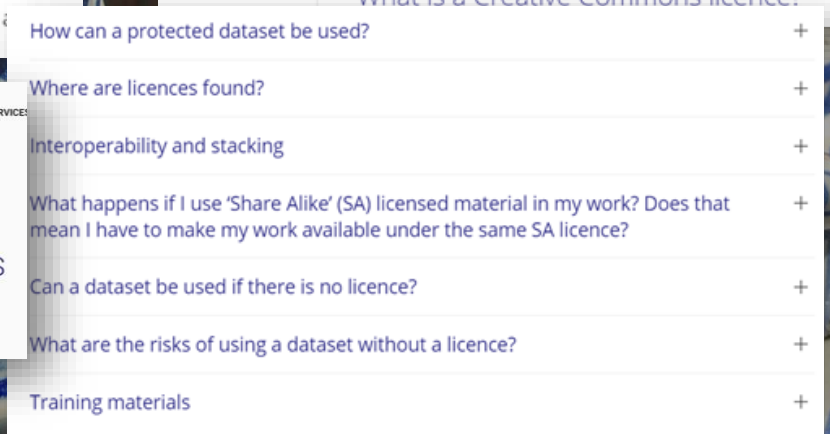
Supporting researchers on the reuse of data: legal aspects to consider

29th April and May 4th, at 2 PM CEST

R = Reusable – Legal aspe







R = Reusable – Legal aspects

2022

Phase1
Research Proposal

Phase2
Research Implementation

Phase3
Research review

Check whether there is background information, data and intellectual property rights brought into the project. More specifically

- Clarify who brings what
- Identify the member state territorial applicability of each r
- Make sure to secure cleara
  - Obtaining any authorisation
  - Agree on rules of ownership
- Aim at avoiding secrecy and at allowing re-use

Define Clearly

- The ownership and/or co-ownership of each research output stemming from
  - The use and re-use of pre-existing background information, data and IPRs,
  - Single or joint research activities within the framework of the project,
  - Single or joint research activities partially within OR outside the framework of the project, if building or depending on project activities.

THE EUROPEAN LEGAL APPROACH TO OPEN SCIENCE AND RESEARCH DATA

Presentata da: Ludovica Paseri

2022

This dissertation proposes an analysis of the governance of the European scientific research, focusing on the emergence of the Open Science paradigm. The paradigm of Open Science indicates a new way of doing science, oriented towards the openness of every phase of the scientific research process, and able to take full advantage of the digital Information and Communication Technologies (ICTs). The emergence of this paradigm is relatively recent, but in the last couple of years it has become increasingly relevant. The European

Creative Co

CC Factsheet

creative commons UK

FACT SHEET ON
CREATIVE COMMONS & OPEN SCIENCE V.0.1

This information guide contains questions and responses to common concerns surrounding open science and the implications of licensing data under Creative Commons licences. It is intended to aid researchers, teachers, librarians, administrators and many others using and encountering Creative Commons licences in their work.

CC0: FROM A LEGAL POINT OF VIEW, THE ONE AND ONLY LICENSE

What is Open Science?

Open Science is the movement to make scientific research and data accessible to all for knowledge dissemination and public reuse.

How should I licence my data for the purposes of Open Science?

We recommend you use the CC0 Public Domain Dedication, which is first and foremost a waiver, but can act as a licence when a waiver is not possible.

CC ZERO LICENCE, 'NO RIGHTS RESERVED' LOGO

By applying CC0 to your data you enable everyone to freely reuse your data as they see fit by waiving (giving up) your copyright and related rights in that data.

You should keep in mind that there are many situations in which data is ***not*** protected as a matter of law. Such data can include facts, names, numbers – things that are considered 'non-original' and part of the public domain thus not subject to copyright protections. Similarly, your database (which is a structured collection of data) might be considered 'non-original' and thus ineligible for copyright, and it might additionally be excluded from other forms of protection (like the EU *sui generis* database right, also known as the 'SGDR', for non-original databases).

In these cases, using a Creative Commons licence such as a CC BY could signal to users that you claim a copyright in the non-original data despite the law, and perhaps despite your real intention.

Finally, if your data is in the public domain worldwide, you might state simply and obviously on the material that no restrictions attach to the reuse of your data and apply a Public Domain Mark.

PUBLIC DOMAIN MARK LOGO

When in doubt, consider which use may be appropriate according to the chart below:

CC0 & PUBLIC DOMAIN LICENCES
WHICH LICENSE TO USE AND WHEN

'Creative arrangement' of data is original, but any copyright has been waived and content is made available copyright-free

'Creative arrangement' of data is not original; the author acknowledges this and communicates the data is in the public domain

nmons and O

But I would like attribution when others use my dataset. In that case, shouldn't I use a CC BY licence?

We recommend that you avoid using a CC BY licence. Here's why:

While attribution is a genuine, recognisable concern, not only might using a CC BY licence be legally unenforceable when no underlying copyright or SGDR protects the work, but it may also communicate the wrong message to the world. A better solution is to use CC0 and simply ask for credit (rather than require attribution), and provide a citation for the dataset that others can copy and paste with ease. Such requests are consistent with scholarly norms for citing source materials.

Legally speaking, datasets that are ***not*** subject to copyright or related rights (and are thus in the public domain) cannot be the object of a copyright licence. Despite this, agreements based in contract law may be enforceable. Creative Commons licences, however, are copyright licences. Therefore, where the conditions for a copyright or related right are not triggered, copyright licences, such as the CC BY licence, are unenforceable.

In some cases, however, rights may exist (like the *sui generis* database right previously mentioned), and permission for others to use your dataset may be legally required. These rights are meant to protect the maker's investment, rather than originality. As such, database rights do not include the moral right of attribution. So by using a CC BY licence, you signal to users that you restrict access to your dataset beyond the protections provided by the law. We are not saying that this cannot be done, we are just saying that if you choose to do this, you should make sure you fully understand what it entails.

USE A CC0

- THEN ASK FOR CREDIT
- PROVIDE A CITATION TO C&P
- BEAR IN MIND IT'S BAD SCIENCE NOT TO CITE THE SOURCE
- CC0 DOES NOT MEAN ACADEMIC UNPOLITENESS

It sounds like you're really pushing for the use of CC0 for open science datasets.

Exactly. Data is only open if anyone is free to use, reuse, and distribute it. This means it must be made available for both commercial and non-commercial purposes under non-discriminatory conditions that allow for it to be modified.

When data is made available for all reuse, others can create new knowledge from combining it. This leads to the enrichment of open datasets and further dissemination of knowledge. Accordingly, CC0 is ideal for open science as it both protects and promotes the unrestricted circulation of data.

And remember, it's bad science not to cite the source of data you use. To help others cite your data include a citation that users can copy and paste to give you credit for your hard work.

cannot be done, we are just saying that if you choose to do this, you should make sure you fully understand what it entails.

I'm uncomfortable with others using my research for commercial purposes. Should I use a non-commercial licence for my dataset?

We recommend you avoid using a non-commercial licence. Here's why:

For legal purposes, drawing a line between what is and is not 'commercial' can be tricky; it's not as black and white as you might think. For example, if you release a dataset under a non-commercial licence, it would clearly prohibit an organisation

I'm uncomfortable permitting use of my research for any and all purposes. Should I use a 'No Derivatives' (ND) licence for my dataset?

We recommend you avoid using a 'No Derivatives' licence. Here's why:

Similar to how a non-commercial licence might restrict meaningful reuse of your dataset, a ND licence can have the same effect: it may prevent someone from recombining and reusing your data for new research. For data to be truly Open Access, it must permit these important types of reuse.

[we are not playing the same music]

Obstacles to the trans-European archiving and sharing of research data

Making research data as openly available as possible is a widely recognised goal. For researchers working on an interdisciplinary project involving several countries, it can be difficult to fully comprehend in which ways open access to research data can be legally obtained. European national laws still diverge.

A report from Knowledge Exchange (Knowledege Exchange, 2011) concludes that it will remain difficult to predict when particular files of research data are protected because of:

- **Diversity in copyright protection**
  Even though most research data will fail to meet the criteria for copyright protection because they are not likely to be considered as "works" (they mainly concern facts), the lack of harmonisation of the criteria for copyright protection in Europe is tricky. E.g., whereas Germany, Denmark and the Netherlands have a relatively similar (higher) originality standard, the UK has a very low standard (skill, judgment and labour) making

- **Diversity in copyright owner**
  If protection applies, the right holder's consent is required for sharing the data. However, the designation of the copyright owner is also different in different jurisdictions. Althoug in many cases the maker of the work will be considered to be the author and therefore th right holder, only Dutch and UK law designate the employer as the right holder if the wor was made in the course of employment.

CESSDA guide

⊕ Czech Republic

⊕ Finland

⊕ Switzerland

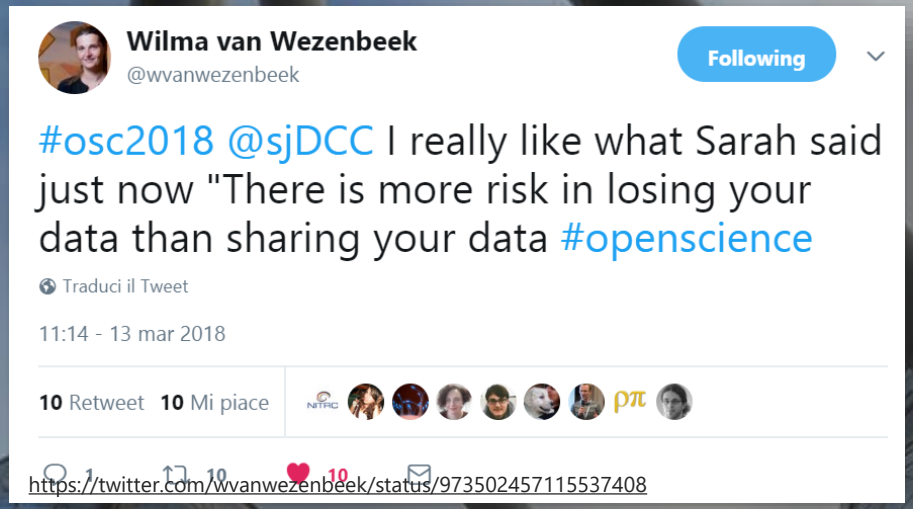
CLARIFY FROM THE BEGINNING POTENTIALLY DIFFERENT OBLIGATIONS FOR THE PARTNERS

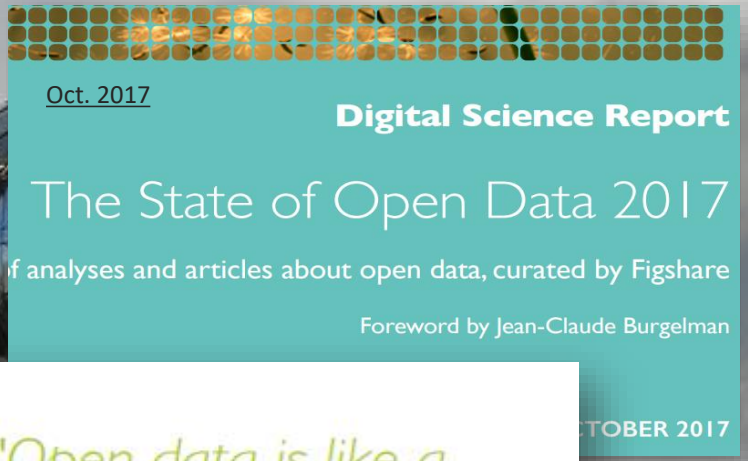
[clear rules]

… SET CLEAR RULES FROM THE BEGINNING
- WHO IS THE RIGHT HOLDER (if)
- WHO HAS THE RESPONSIBILITY OF PRESERVE

3. OPEN DATA

Why Open data?





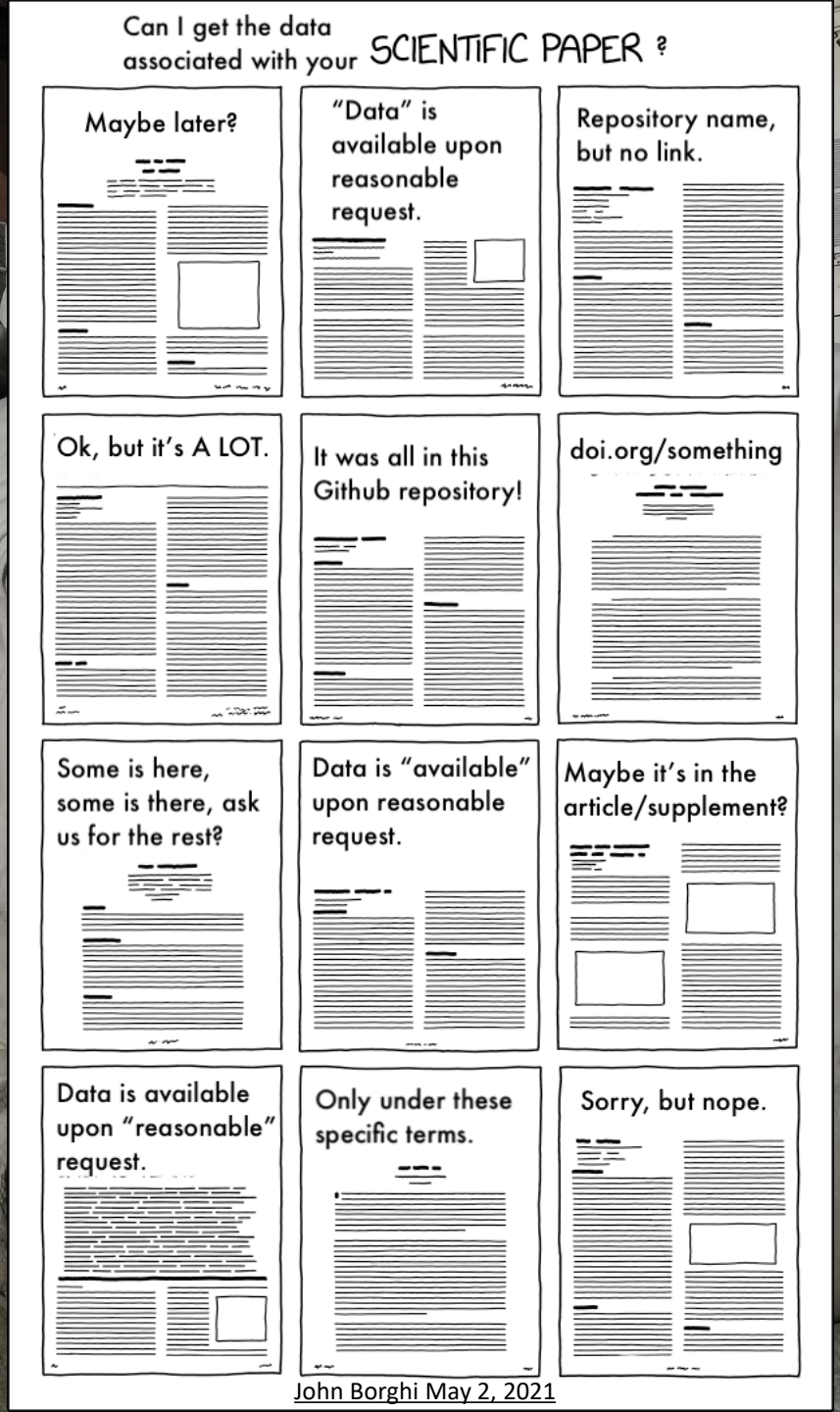
SHARING DATA: GOOD FOR SCIENCE, GOOD FOR YOU

"Open data is like a renewable energy source: it can be reused without diminishing its original value, and reuse creates new value."

Open data saves lives

Open data saves lives. The global pandemic has highlighted beyond anything that came before it the importance of data sharing in solving the big challenges of our time. COVID-19 data may be the most visualized data in history and it was made publicly available on a daily basis to people all over the world. The urgent need to better understand and treat the virus in 2020 brought unprecedented collective and collaborative action from all research stakeholders on an international scale to bring down barriers to research and speed up analysis and testing. These efforts, combined with support from governments and industry, resulted in not one but many vaccines made available by the end of the year. This gives us a glimpse of what incredible research outcomes are possible when we start with collaboration to address a common threat. Imagine how much more we could do, how many more lives we could save, if research data was routinely made open and shared. So, why isn't data sharing the norm? The answers lie in the harmony needed between policies, infrastructure, and practices.

[reality]



John Borghi May 2, 2021

[real


FOR THIS FORM I NEED TO GIVE A REASON WHY YOU'RE NOT SHARING YOUR DATA
I CAN'T SHARE MY DATA BECAUSE THIS GROUP IN BELARUS WILL SEND IT TO MY EX STUDENT IN AUCKLAND WHO WILL INVENT EVERYTHING BEFORE ME. ALSO THE ICELANDIC ARE WATCHING!
OKAY... I'LL JUST TICK THE BOX LABELLED "IRRATIONAL CONSPIRACY"
ERRANTSCIENCE.COM

SCHOLARLY COMMUNICATION IS A CONVERSATION

People will contact me to ask about stuff

Christopher and Alex (C&A) say: "This is usually … objection of people who feel overworked and th… [data sharing] isn't part of their job…" I would a… learning from each other – if a researcher is opposed to the idea of discussing their datasets, collaborating with others, and generally being a good science citizen, then they should be outed by their community as a poor participant.

People will misinterpret the data

IMPOSSIBLE, IF IN «R» IN «FAIR» YOU DOCUMENTED

C&A suggest this: "Document how it s… correct such people; those that misint… help." From the UK Data Archive: "Pro… contextual information for your resea… correctly use and understand your da…

It's worth mentioning, however, a second point C&A make: "Publishing may actu… be useful to counter willful misrepresentation (e.g. of data acquired through Free… of Information legislation), as one can quickly point to the real data on the web t… refute the wrong interpretation."

My data is not very interesting

EHM… SO WHY ARE WE FUNDING YOU WITH PUBLIC MONEY?

C&A: "Let others judge how interesti… people that care about them." I'd als… dataset has value to future research… "climate change" was a research top… documenting and understanding the phenomenon. From the UK Data Archive: "

I might want to use it in a research paper

Anyone who's discussed data sharing with a researcher is familiar with this excuse. The operative word here is *might*. How many papers have we all considered writing, only to have them shift to the back bur… s is a real concern.

EMBARGO PERFECTLY «FAIR»

C&A suggest the embargo route: "One … embargo; require people to archive the… public after X months. You could even g… things that are no longer cared about b… eventually everything can become ope… their datasets, but I would caution to have any restrictions default to sharing. That is, after X months the data are automatically made open by the repository.

I would also add that, as the original collector of the data, you are at a huge advantage compared to others that might want to use your dataset. You have knowledge about your system, the conditions during collection, the nuances of your methods, et cetera that could never be fully described in the best metadata.

I'm not sure I own the data

My data is too complicated.

IMPOSSIBLE, IF IN «R» IN «FAIR» YOU DOCUMENTED

C&A: "Don't be too smug. If it turns out it's not that compl… professional [standing]." I would add that if it's too compl… complicated to reproduce, which means it's arguably not… can be solved by more documentation.

My data is embarrassingly bad

HOW CAN YOU DO RESEARCH WITH «BAD» DATA???

C&A: "Many eyes will help you imp… people will accept your data for what it is." I a… ck end of making the sausage. We know it's … ccept that. Plus it helps you strive will be at m… next collection phase.

It's not a priority and I'm busy

A GROWING NUMBER OF FUNDERS AND JOURNALS ASK FOR DATA…
IT'S A PRIORITY NOW

CARLY STRASSER

Open Science + Data Sharing Advocate

Home
Blog Posts
Previous Research

http://carlystrasser.net/closed-data-excuses-excuses/

Closed Data… Excuses, Excuses

…and still

Digital Science Report

The State of Open Data 2021

The longest-running longitudinal survey and analysis on open data

Foreword by Natasha Simons, Australian Research Data Commons (ARDC)

Nov. 29, 2021

November 2021

What is most striking about this year's State of Open Data report is that while researchers' familiarity and compliance with the FAIR data principles is greater than ever before, there is also more concern about sharing datasets than ever before. In their article on the three key findings of this year's State of Open Data report, Dr. Greg Goodey and Megan Hardeman stress that concern has risen in several key areas, one of which is not receiving enough credit or acknowledgement for data sharing. This points to the uncomfortable tension between the increasing ubiquity of data management and data availability policies and the rareness of rewards and recognition for data sharing. Clearly, the reward and recognition structures of academia are misaligned with ansparency of research

	2018	2019	2020	2021
Concerns about misuse of data	518 36%	2082 37%	1881 38%	1920 43%
Not receiving appropriate credit or acknowledgement	478 33%	1834 32%	1531 31%	1739 39%
Unsure about copyright and data licensing	508 35%	1858 33%	1479 30%	1585 35%
Contains sensitive information or requires consent	416 29%	1274 22%	1535 31%	1366 30%
I am unsure I have permission from my funder or institute	458 32%	1454 26%	1291 26%	1110 25%
Organising data in a presentable and usable way	456 32%	1434 25%	1137 23%	1150 26%
Costs of sharing data	274 19%	1497 26%	1192 24%	1147 26%
Lack of time to deposit data	300 21%	1389 24%	984 20%	942 21%

CONCERNS

- MISUSE
- NO CREDIT
- COPYRIGHT

...concerns

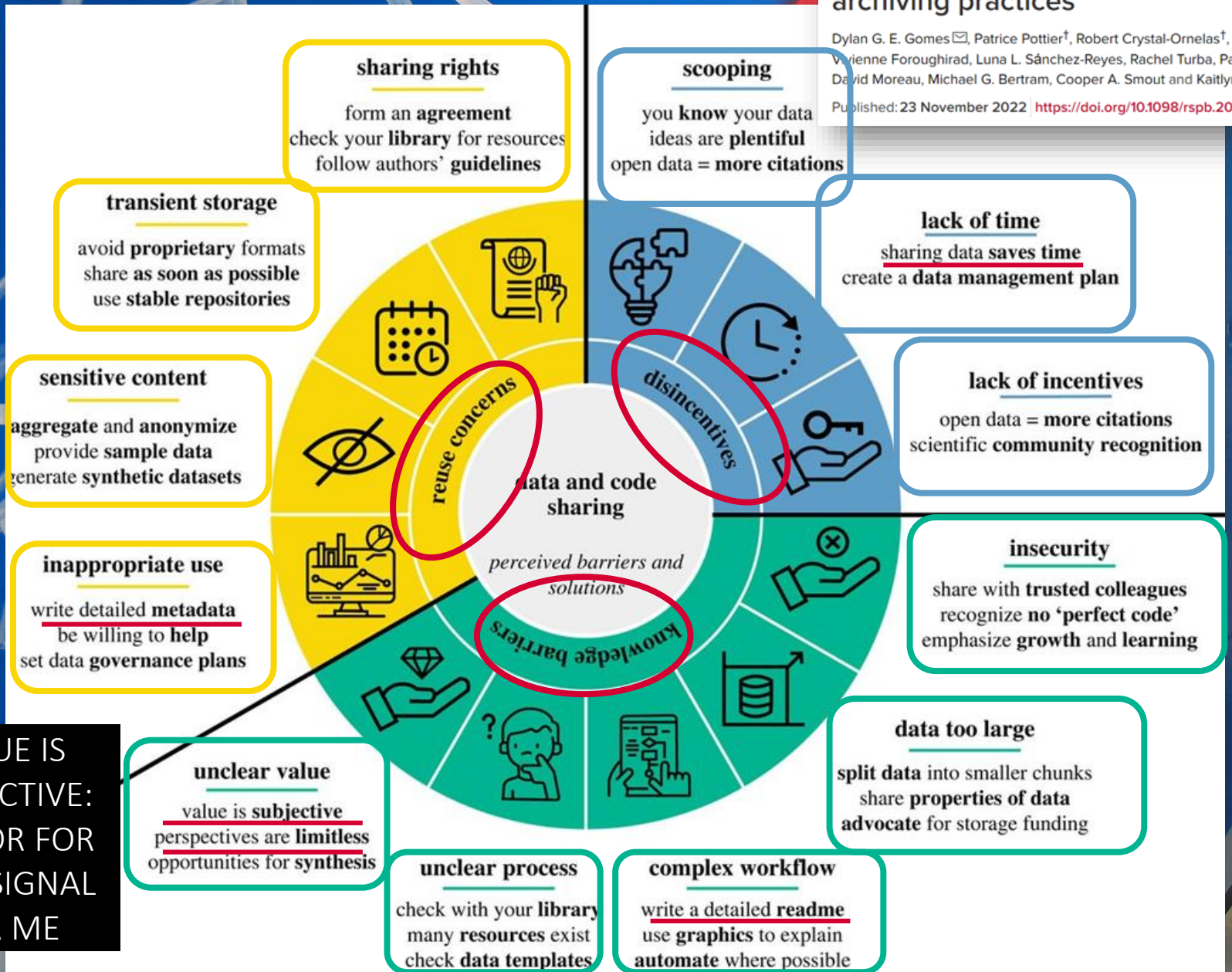
Biological science practices

2022

Why don't we share data and code? Perceived barriers and benefits to public archiving practices

Dylan G. E. Gomes ✉, Patrice Pottier†, Robert Crystal-Ornelas†, Emma J. Hudgins, Vivienne Foroughirad, Luna L. Sánchez-Reyes, Rachel Turba, Paula Andrea Martinez, David Moreau, Michael G. Bertram, Cooper A. Smout and Kaitlyn M. Gaynor

Published: 23 November 2022 https://doi.org/10.1098/rspb.2022.1113



VALUE IS SUBJECTIVE: RUMOR FOR YOU, SIGNAL FOR ME

Pro an

ANSWERS TWO

REASONS NOT TO SHARE DATA

UK Data service p. 11

1/2

	REASONS NOT TO SHARE DATA	REPLIES OR ARGUMENTS IN FAVOUR OF SHARING
1	My data is not of interest or use to anyone else.	It is! Researchers want to access data from all kinds of studies, methodologies and disciplines. It is very difficult to predict which data may be important for future research. Who would have thought that amateur gardener's diaries would one day provide essential data for climate change research? Your data may also be essential for teaching purposes. Sharing is not just about archiving your data but about sharing them amongst colleagues.
2	I want to publish my work before anyone else sees my data.	Data sharing will not stand in the way of you first using your data for your publications. Most research funders allow you some period of sole use, but also want timely sharing. Also remember that you have already been working with your data for some time so you undoubtedly know the data better than anyone coming to use them afresh. If you are still concerned you can embargo your data for a specific period of time.
3	I have not got the time or money to prepare data for sharing	It is important to plan data management early in the research data lifecycle. Data management ideally becomes an integral part of your research practice, reduces time and financial costs and greatly enhancing the quality of the data for your use too.
4	If I ask my respondents for consent to share their data then they will not agree to participate in the study.	Don't assume that participants will not participate because data sharing is discussed. Talk to them – they may be less reluctant than you might think, or less concerned over data sharing! Make it clear that it is entirely their decision, whereby they can decide whether their data can be shared, independent of them participating in the research. Explain clearly what data sharing means, and why it may be important. But they are still free to consent or not. You can always explain what data archiving means in practice for their data. If you have not asked permission to share data during the research, then you can always return to gain retrospective permission from participants.
5	I am doing highly sensitive research. I cannot possibly make my data available for others to see.	The first thing is to ask respondents and see if you can get consent for sharing in the first instance. Anonymisation procedures can help to protect identifying information. If these first two strategies are not appropriate then consider controlling access to the data or embargoing for a period of time. Also data that is held in the UK Data Archive is not publically available. Only registered researchers can gain access to the data.
6	I am doing quantitative research and the combination of my variables discloses my participant's identity.	Quantitative data can be anonymised through processes of aggregation, top coding, removal of variables, or controlled access to certain variables (i.e. postcodes).
7	I have collected audiovisual data and I cannot anonymise them, therefore I cannot share these data.	Visual data can be anonymised through blurring faces or distorting voices, but this can be time consuming and costly to carry out. It can mean losing much of the value of the data. It is better to ask for consent to share data from participants in an unanonymised form,
8	I have made promises to destroy my data once the project finishes.	Why were such promises made? Always avoid making unnecessary promises to destroy data. There is usually no legal or ethical need to do so, except in the case of personal data. But that certainly would not apply to research data in general. Also consider where you have received this advice from? You may need to negotiate with research ethics committee or ethics boards about this agreement.

ARGUMENTS IN FAVOUR OF SHARING

Data Management Plans: the pillars of your research

… DMP?

DATA MANAGEMENT PLANS ARE YOUR FIRST RESEARCH «PRODUCT» IF YOU WANT YOUR DATA TO BE AVAILABLE AND REUSABLE (EVEN BY YOURSELVES!!!)

BROWSE PUBLISH ABOUT

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

EDITORIAL

Ten simple rules for getting and giving credit for data

Elisha M. Wood-Charlson, Zachary Crockett, Chris Erdmann, Adam P. Arkin, Carly B. Robinson

Published: September 29, 2022 • https://doi.org/10.1371/journal.pcbi.1010476

2022

Rule 3: Data management plans are your first research product

Now that you have mastered the complexity (or at least scratched the surface) of what it takes to create FAIR, comparable, and reproducible data, we need to talk about data management plans (DMPs). These are often required by funders as supplementary documents to research grants, where you outline when, where, and how data from the project will be preserved and shared. We won't go into best practices for creating a DMP, as that is well articulated by Michener [28]. However, we do want to emphasize that DMPs are no longer just supplementary pdfs. They can (and should) be created as FAIR, machine-actionable, living documents [29]. DMPs establish the initial node in your upcoming research product network (data, code, etc.). DMPs connect the people and data to the funding agency and put a stake in the ground for the


IT IS A STRUCTURED WAY TO THINK OF YOUR DATA
CLEAR RULES, LESS MISTAKES FROM THE BEGINNING
IT'S A FORMAL DOCUMENT ABOUT HOW YOU ARE GOING TO MANAGE YOUR DATA
IT'S A «LIVING DOCUMENT», IT GROWS WITH THE PROJECT
A NEW WAY OF THINKING TO YOUR RESEARCH, FROM THE PERSPECTIVE OF YOUR DATA
IT IS THE RIGHT VENUE TO JUSTIFY YOUR CHOICES ON OPEN/CLOSED
...LET'S BE CLEAR:
THE ISSUE HERE IS NOT «LEARNING» HOW TO DRAFT A DMP
BUT LEARNING HOW TO RESPONSIBLY MANAGE FAIR DATA.
DMP IS ITS PRACTICAL DECLARATION
IT IS CRUCIAL TO ENSURE FUNDING TO COVER THE COSTS OF DATA MANAGEMENT
DATA MANAGEMENT PLAN

Tips and tricks

SINTETIC AND SPECIFIC

DO NOT COPY/PASTE

BEING GENERIC IS USELESS
[we expect a huge size of data; data will be available]

EVERY DATASET IS UNIQUE, EVERY INFRASTRUCTURE IS DIFFERENT, EVERY RESEARCH HAS DIFFERENT PARTNERS/POLICIES

- LET'S USE TABLES AND BULLET POINTS
- BE CLEAR, SHORT SENTENCES. IT'S NOT A DISSERTATION

- IF YOU DON'T KNOW IT, SAY IT [THEN YOU'LL UPDATE]
- IF NOT, IT SEEMS YOU ARE NOT AWARE [SAME DIFFERENCE BETWEEN BLANK CELL AND A CELL WITH N.A.]

WHAT YOU STATE IN THE DMP THEN HAS TO BE DONE…
DON'T SHOW OFF
DON'T DECLARE SOMETHING YOU CAN'T GET
e.g. PSEUDONIMYZED DATA, not ANONIMYZED

Some help

OA@unito.it

Come scrivere un DMP

In UniTO | Come | Cos'è utile | Perché è importante | Editori e Politiche Open Access (EPOcA) | Eventi

Come scrivere un Data Management Plan

Il Data Management Plan (DMP) è un documento strutturato, vivo, che cresce con il progetto. Serve a dichiarare come si producono i dati, come li si conserverà e come li si condividerà (se possibile).

Pensatelo come le "Istruzioni per l'uso" dei vostri dati.

Deve essere

- **sintetico**: evitate sproloqui, non è una dissertazione. Frasi
- **schematico**: utilizzate il più possibile tabelle e punti elenco
- **preciso**: evitate frasi (viste davvero) tipo "we expect a huge
far perdere tempo a chi lo scrive e achi lo legge. Quantifica

Preparing a Data Management Plan (DMP)

A Data Management Plan is a document specifying how research data will be **handled both during and after a research project**. It identifies key actions and strategies to ensure that research data are of a high-quality, secure, sustainable, and – to the extent possible – accessible and reusable.

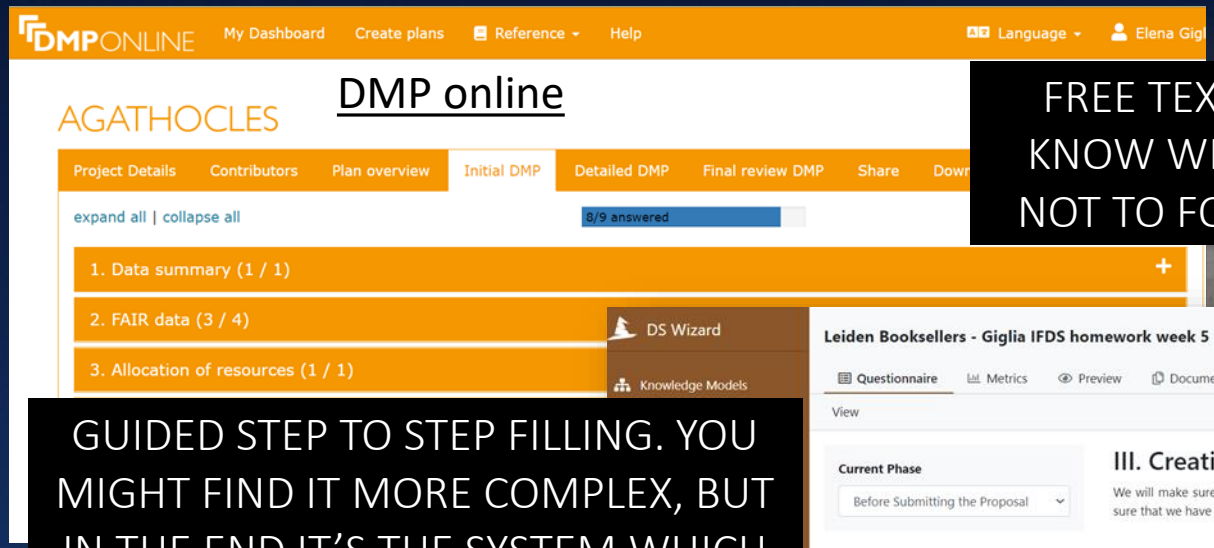
Preparing a DMP

Why develop a DMP?

Creating a DMP is **considered good practice** for any research project using or generating data. After all, planning is the first step towards proper research data management.

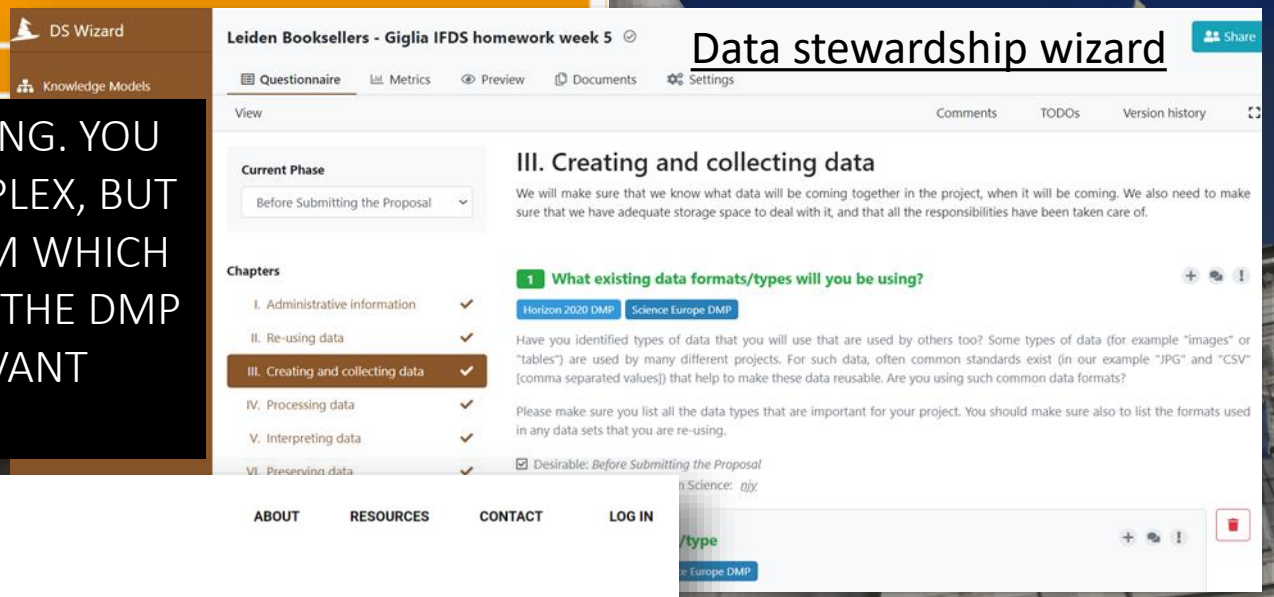
Decisions made early on affect what you can do later, so good and timely planning can **save you a lot of time and problems** in the longer run. It also helps you consider the necessary **resources and** costs for data management, so you can include these in your grant applications.

In addition, you may be **required** to draft a DMP, for example by your research funder.



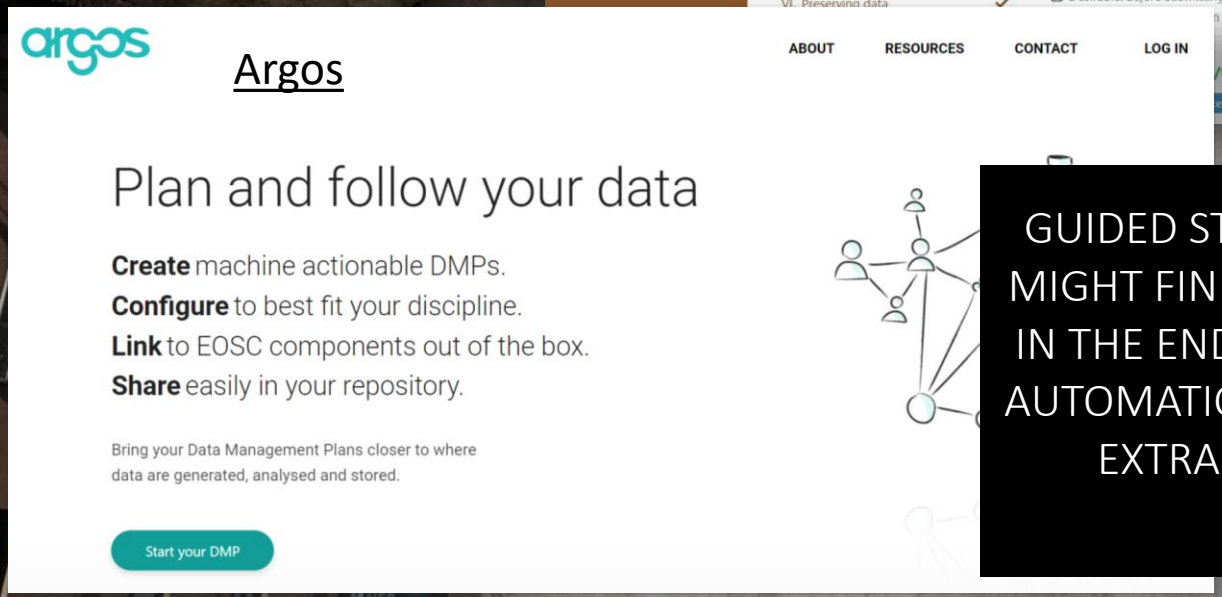
DMP online

FREE TEXT. YOU HAVE TO KNOW WHAT TO ADDRESS NOT TO FORGET ANYTHING



Data stewardship wizard

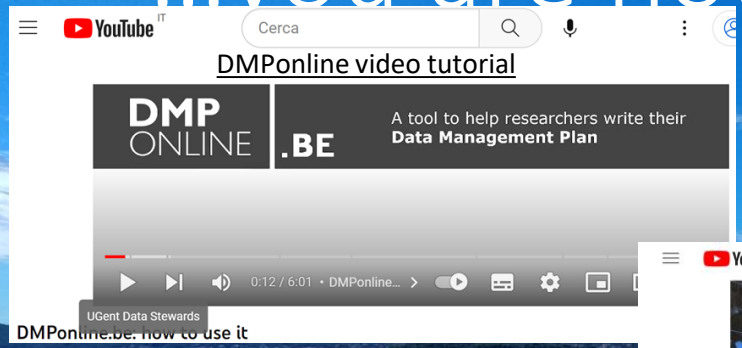
GUIDED STEP TO STEP FILLING. YOU MIGHT FIND IT MORE COMPLEX, BUT IN THE END IT'S THE SYSTEM WHICH AUTOMATICALLY GENERATE THE DMP EXTRACTING THE RELEVANT INFORMATION



Argos

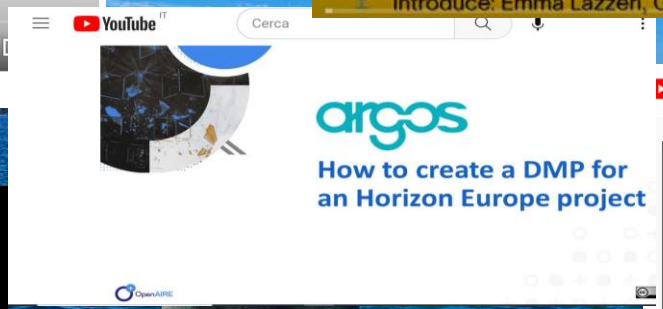
GUIDED STEP TO STEP FILLING. YOU MIGHT FIND IT MORE COMPLEX, BUT IN THE END IT'S THE SYSTEM WHICH AUTOMATICALLY GENERATE THE DMP EXTRACTING THE RELEVANT INFORMATION

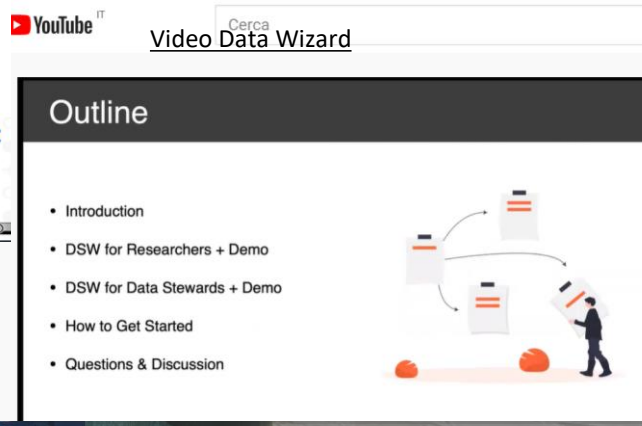
...you are not alo



DMPonline video tutorial

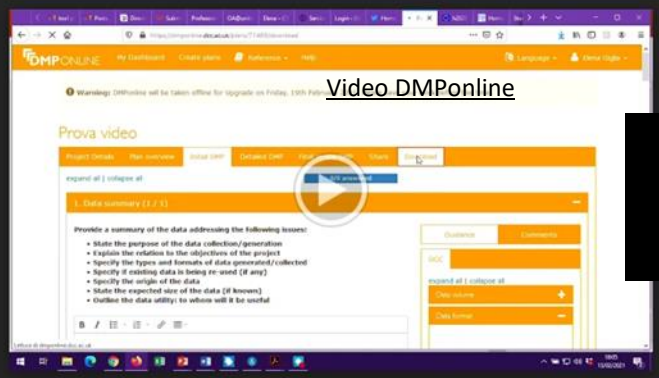
DMPonline.be: how to use it





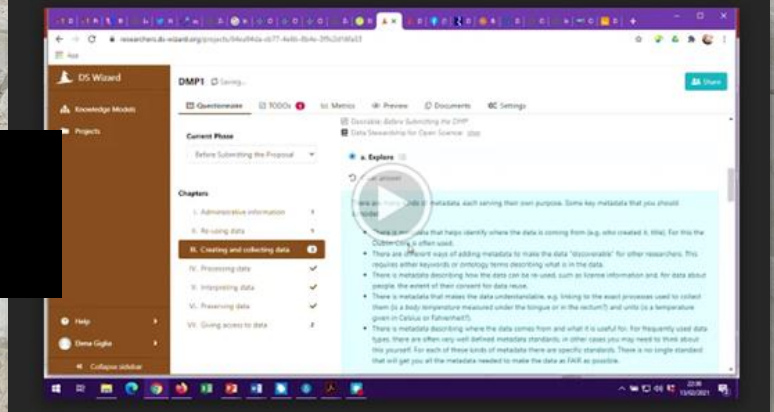
Video Data Wizard

VIDEO TUTORIALS



Video DMPonline

TUTORIAL (IT)



«How to» in Horizon Europe

Open Science in HEU

OPEN SCIENCE PRACTICES
EVALUATED UNDER
«EXCELLENCE»
a) MANDATORY
b) RECOMMENDED
BOTH TO BE EMBEDDED IN
THE PROPOSAL

Open science

Open science in Horizon Europe

Open science is an approach based on open cooperative work and systematic sharing
knowledge and tools as early and widely as possible in the process. It has the potenti
to increase the quality and efficiency of research and accelerate the advancement
knowledge and innovation by sharing results, making them more reusable an
improving their reproducibility. It entails the involvement of all relevant knowledg
actors.

Horizon Europe moves beyond open access to open science for which it feature
a comprehensive policy implemented from the proposal stage to project reporting. Th
Horizon Europe Regulation sets the legal basis for the open science obligations an
incentives that apply to Horizon Europe beneficiaries. The Annotated Grant Agreemer
provides guidance on how to comply with the open science obligations required in th
Model Grant Agreement. **The present guide complements the information**
pro
the

In Horizon Europe, open science practices are considered in the evaluation of proposals, under 'excellence' and under the 'quality and efficiency of implementation'.[17] There are mandatory open science practices, which are required for all projects through the Model Grant Agreement and/or through the work programme or call conditions, and recommended practices (all open science practices that are not mandatory). Recommended open science practices are incentivised through their the evaluation at the proposal stage. Proposers should be aware of both mandatory and recommended practices and integrate them into their proposals.

V.1 June 17 2021

Horizon Europe

Programme Guide

Open Science in HEU

IN EXCELLENCE – METHODOLOGY /QUALITY OF IMPLEMENTATION

1) EXPLAIN HOW YOU WILL IMPLEMENT MANDATORY OS PRACTICES
2) HOW YOU WILL ADOPT RECOMMENDED OS PRACTICES – GETTING A HIGHER SCORE!
3) JUSTIFY IF YOU RECKON NO OPEN SCIENCE PRACTICE FITS IN YOUR PROPOSAL

Open science practices are evaluated under the **'Excellence'** criterion (in particular under methodology) and under the **'Quality and efficiency of implementation'** award criterion. Proposers should address open science practices in the relevant section on open science under methodology[20].

Proposers will have to provide concrete information on **how** they plan to comply with the **mandatory open science** practices. Failure to sufficiently address this, will result in a lower evaluation score.

A clear explanation of how they will adopt **recommended practices**, as appropriate for their projects, will result in a higher evaluation score.

If proposers believe that none of the open science practices (mandatory or recommended) apply to their project, then they have to provide a **justification**.

Under the 'excellence' part of their proposals, in the section on methodology, proposers should describe how open science practices (mandatory and recommended, as appropriate) are implemented as an integral part of the methodology and show how their implementation is adapted to the nature of their work, therefore increasing the chances of the project delivering on its objectives. Information relevant to the specific area of the proposal should be provided in no more than one page. If open science practices are not applicable to the proposal, justifications should be provided sp that, if

Horizon Europe


V.1 Feb 2021

Horizon Europe (HORIZON)
Euratom Research and Training Programme (EURATOM)

General Model Grant Agreement
EIC Accelerator Contract


Pre-draft July 2021

EU Grants

AGA – Annotated Model Grant Agreement

EU Funding Programmes 2021-2027

ART. 6.2 SPECIFIC ELIGIBILITY CONDITIONS FOR EACH BUDGET CATEGORY C.3 OTHER GOODS [P.30]
ART. 17 COMMUNICATION, DISSEMINATION AND VISIBILITY [P.49]
ANNEX 5, TO ART. 17, OPEN SCIENCE [P.107-109]

- ART. 6.2.C.3 OTHER COSTS (DISSEMINATION) P.[69]
- ART.17 COMMUNICATION& DISSEMINATION [P.113-115]
- ANNEX 5 IPR RULES [P.124-125 E 133-146 EXPLOITATION&PROTECTION]
- ANNEX 5 DISSEMINATION & OPEN SCIENCE [P.153-161]

DEFINTION OF «TRUSTED REPOSITORY» P. 156

- ANNEX 5 DISSEMINATION PLAN [P. 162]

Horizon Europe

- DISSEMINATION & IPR MANAGEMENT [P.30-37]
- OPEN SCIENCE [P.38-52]

RIGHTS RETENTION CLAUSE [P.49] AND USEFULT TOOLS

- CITIZIEN SCIENCE [P.52-54]

PART A – LIST OF PUBLICATIONS (OPEN ACCESS) [P.12]

PART B – 1.EXCELLENCE – 1.2 METHODOLOGY (OPEN SCIENCE+DATA MANAGEMENT) [P.8]

PART B – 2.IMPACT

PART B – 3.2 CONSORTIUM CAPACITY [P.15]

Open Science in Horizon Europe

MANDATORY AND RECOMMENDED PRACTICES TO BE ADAPTED TO YOUR PROJECT – EVALUATED AT THE PROPOSAL STAGE

Open Science in Horizon Europe RIA/IA/CSA

IN THE METHODOLOGY YOU NEED TO ADDRESS BOTH:
1) HOW YOU WILL COMPLY WITH THE MANDATORY PRACTICES
2) HOW YOU WILL ADOPT RECOMENDED PRACTICES

RECOMMENDED PRACTICES

IN THE **LIST OF ACHIEVEMENTS:** 5 RELEVANT OUTPUTS (publications, data) OPENLY ACCESSIBLE + PERSISTENT IDENTIFIER + «AS OPEN AS POSSIBLE»

IN THE PROJECT **METHODOLOGY**
1) EMBEDDED OPEN SCIENCE PRACTICES
2) FAIR DATA MANAGEMENT + DMP SCHEMA

MAXIMIZING IMPACT USING OPEN SCIENCE (OS IS AMONG KEY PATHWAY INDICATORS) + SCHEMA OF DISSEMINATION PLAN (DELIVERABLE M6)

OPEN SCIENCE PRACTICES/SKILLS IN PREVIOUS PROJECTS TO EVALUATE **QUALITY OF IMPLEMENTATION AND CONSORTIUM CAPACITY**

MANDATORY PRACTICES

DEPOSIT+ IMMEDIATE ACCESS (ZERO EMBARGO + CC BY) =
1. OPEN RESEARCH EUROPE
2. OA JOURNAL
3. TRADTIONAL JOURNAL [RETAINING RIGHTS]

1. RESPONSIBLE MANAGEMENT ACCORDING TO **FAIR PRINCIPLES**
2. DATA AND OTHER OUTPUTS **«AS OPEN AS POSSIBLE, AS CLOSED AS NECESSARY»**
3. DATA MANAGMENT PLAN BY M6

INFORMATION ON OUTPUTS/TOOLS AND **ACCESS** TO DATA/RESULTS FOR **VALIDATION OF RESEARCH**

LIST OF ACHIEVEMENTS	EXCELLENCE	IMPACT	QUALITY OF IMPLEMENTATION	OPEN SCIENCE	OPEN SCIENCE	ENSURE REPRODUCIBILITY
Template PartA	Template PartB	Template PartB	Template PartB	Publications	FAIR data	

PROJECT PROPOSAL WILL BE EVALUATED ON
a) HOW IT WILL ADOPT RECOMMENDED PRACTICES AND b) HOW IT WILL BE COMPLIANT TO MANDATORY ONES

Giglia 2021

Horizon Europe

How do I address open science in my proposal?

HORIZON EUROPE

Open science (OS) takes a central place in Horizon Europe and open science practices are considered in the evaluation of Horizon Europe proposals. If not applicable to the proposal, justifications should be provided so that, if evaluators agree, open science will not be taken into consideration in the evaluation.

...in a nutshell...

Infographic created by Open science team, Ghent University Library and adapted by Elena Giglia

Giglia 2021

Open Science in Horizon Europe

EXAMPLES OF MANDATORY/RECOMMENDED PRACTICES

Open Science practices

What?	How?	Mandatory in all calls/recommended
Early and open sharing of research	Preregistration, registered reports, preprints, etc.	Recommended
Research output management	Data management plan (DMP)	**Mandatory**
Measures to ensure reproduciblity of research outputs	Information on outputs/tools/instruments and access to data/results for validation of publications	**Mandatory**
Open access to research outputs through deposition in trusted repositories	• Open access to publications • Open access to data • Open access to software, models, algorithms, workflows etc.	• **Mandatory** for peer-reviewed publications • **Mandatory** for research data **but** with exceptions ('as open as possible…') • Recommended for other research outputs
Participation in open peer-review	Publishing in open peer-reviewed journals or platforms	Recommended
Involving all relevant knowledge actors	Involvement of citizens, civil society and end-users in co-creation of content (e.g. crowd-sourcing, etc.)	Recommended

Slide courtesy Victoria Tsoukala, EC

Mandatory/recommended

IN THE PROPOSAL YOU NEED TO ADDRESS BOTH:

1. HOW YOU WILL BE COMPLIANT TO THE MANDATORY
2. HOW YOU WILL ADOPT THE RECOMMENDED

MANDATED OPEN SCIENCE PRACTICES ARE DETAILED IN THE GRANT AGREEMENT:

- OPEN ACCESS TO PUBLICATIONS
- OPEN ACCESS TO DATA
- RESEARCH OUTPUTS MANAGEMENT
- REPRODUCIBILITY

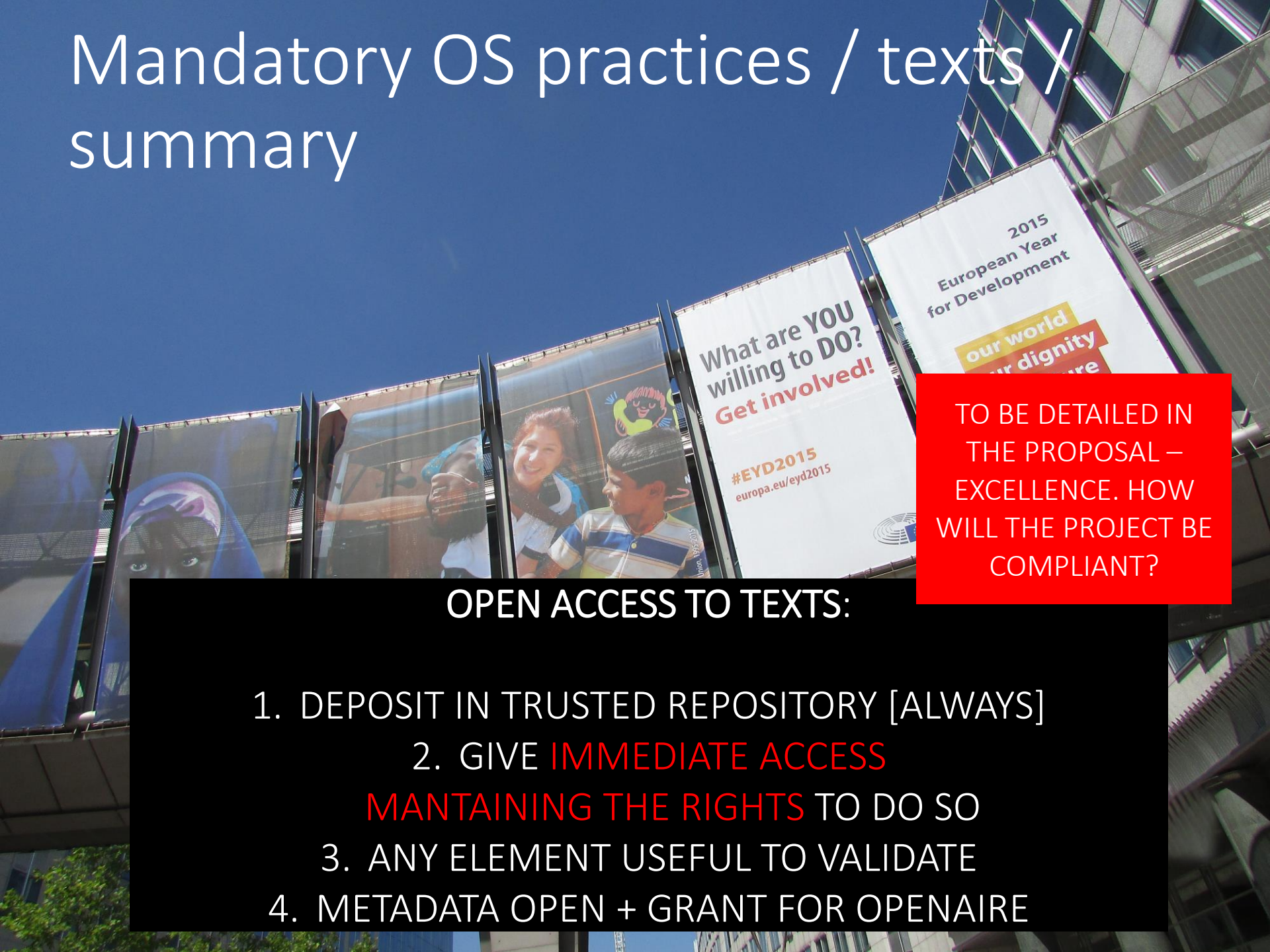
RECOMMENDED OPEN SCIENCE PRACTICES :

e.g. open peer review, pre-registration, citizen science…

SOME CALLS COULD MANDATE MORE PRACTICES

«RECOMMENDED» BUT THE PROPOSAL IS EVALUATED ALSO ON THIS BASIS

Mandatory OS practices / texts / summary



TO BE DETAILED IN THE PROPOSAL – EXCELLENCE. HOW WILL THE PROJECT BE COMPLIANT?

OPEN ACCESS TO TEXTS:

1. DEPOSIT IN TRUSTED REPOSITORY [ALWAYS]
2. GIVE IMMEDIATE ACCESS MANTAINING THE RIGHTS TO DO SO
3. ANY ELEMENT USEFUL TO VALIDATE
4. METADATA OPEN + GRANT FOR OPENAIRE

[Patents and Open Science]

European Commission

IP Helpdesk

Home | Services | Regional helpdesks | IP management and resources | About | News & Events

European Commission > IP Helpdesk > News & Events > News > Open Science vs. IPR in Horizon Europe – which one wins?

NEWS ARTICLE | 17 September 2021 | European Innovation Council and SMEs Executive Agency

Open Science vs. IPR in Horizon Europe – which one wins?

1) MANDATORY TO PROTECT (IF THE CASE)
2) MANDATORY TO DISSEMINATE IN OPEN ACCESS DOES NOT MEAN «MANDATORY TO PUBLISH».

IF YOU PUBLISH, IT MUST BE OPEN

Our enquirer's concerns were the following: is it possible to first file for a patent (his proposed project would involve the development of a new invention), and only then to proceed to the dissemination of results via an open access article? Or does the Open Science policy applicable in Horizon Europe prevail over IPR protection, and imposes the disclosure of the invention in an open access journal as soon as possible?

To answer this, it is essential to keep in mind that in Horizon Europe (including MSCA), grant beneficiaries have the **obligation to protect their results** - see Annex 5 to the model GA for Unit ants incl. MSCA (page 88 onwards).

the other hand, Open Science practices, while compulsory in Horizon Europe, are not ompatible with this obligation… even though they may seem so. Indeed, the open access gation (for example) is NOT an obligation to publish. Simply, if/when fellows publish a scientific cle, it will have to be in open access.

ther words, Open Science obligations in Horizon Europe are NOT a general obligation to seminate. **They are even less an obligation to surrender IP rights, and for this reason uld not be construed in opposition to IP protection**. The dissemination of Horizon results can postponed to allow the appropriate protection of results beforehand - see the grant agreement uses on dissemination (annex 5 to the MGA for Unit Grants, pp.94-95) according to which **the semination obligation is made subject to any restrictions linked to the protection of ellectual property**.

is is confirmed by the European Commission in the annotated model grant agreement for Horizon Europe (see page 153).

To sum up: not only is it possible for fellows and beneficiaries to protect their results first (e.g. via a patent filing), but **it is also necessary to ensure compliance with the obligation to protect the project results**. This is something that can be explained in the proposal – that the strategy is, first, to secure IP protection, and that once this is completed, dissemination obligations will be fulfilled, including via open access if publications are foreseen.

3 ways to be compliant

1. PUBLISH IN ORE – OPEN RESEARCH EUROPE

NO COSTS

2. PUBLISH IN AN OPEN ACCESS JOURNAL + DEPOSIT [IN HE ALWAYS NEEDED]

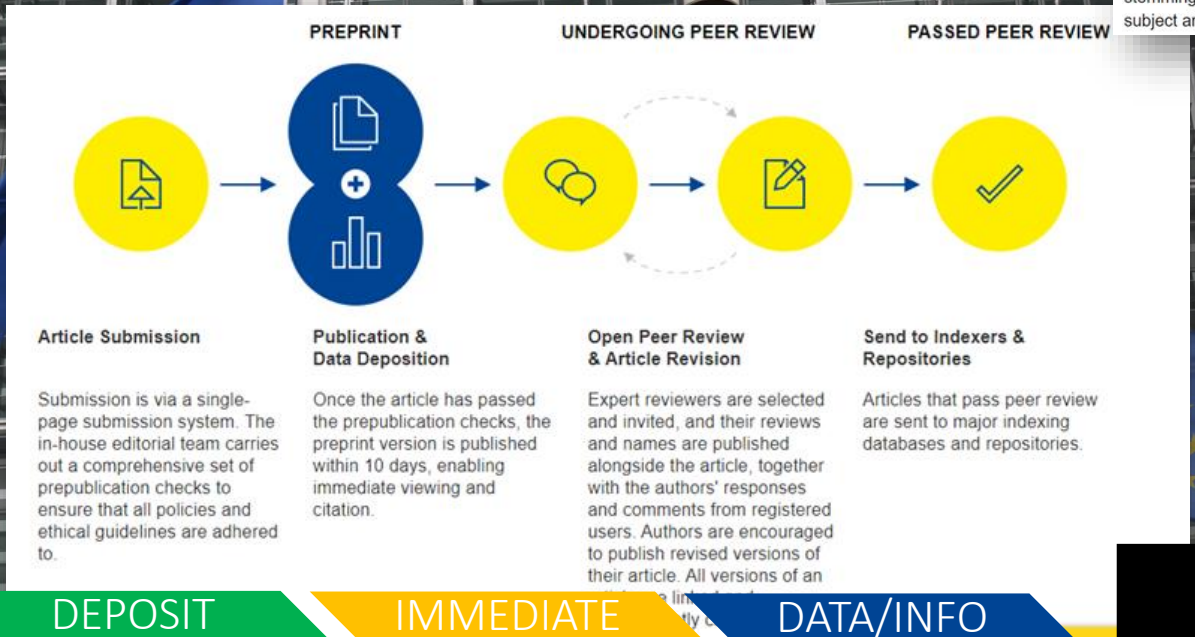
POSSIBLE APC - REIMBURSED

NO REIMBURSE FOR HYBRID

3. PUBLISH IN A SUBSCRITPION BASED JOURNAL + RETAIN RIGHTS TO DEPOSIT+ IMMEDIATE ACCESS

1. Publishing in ORE

ORE

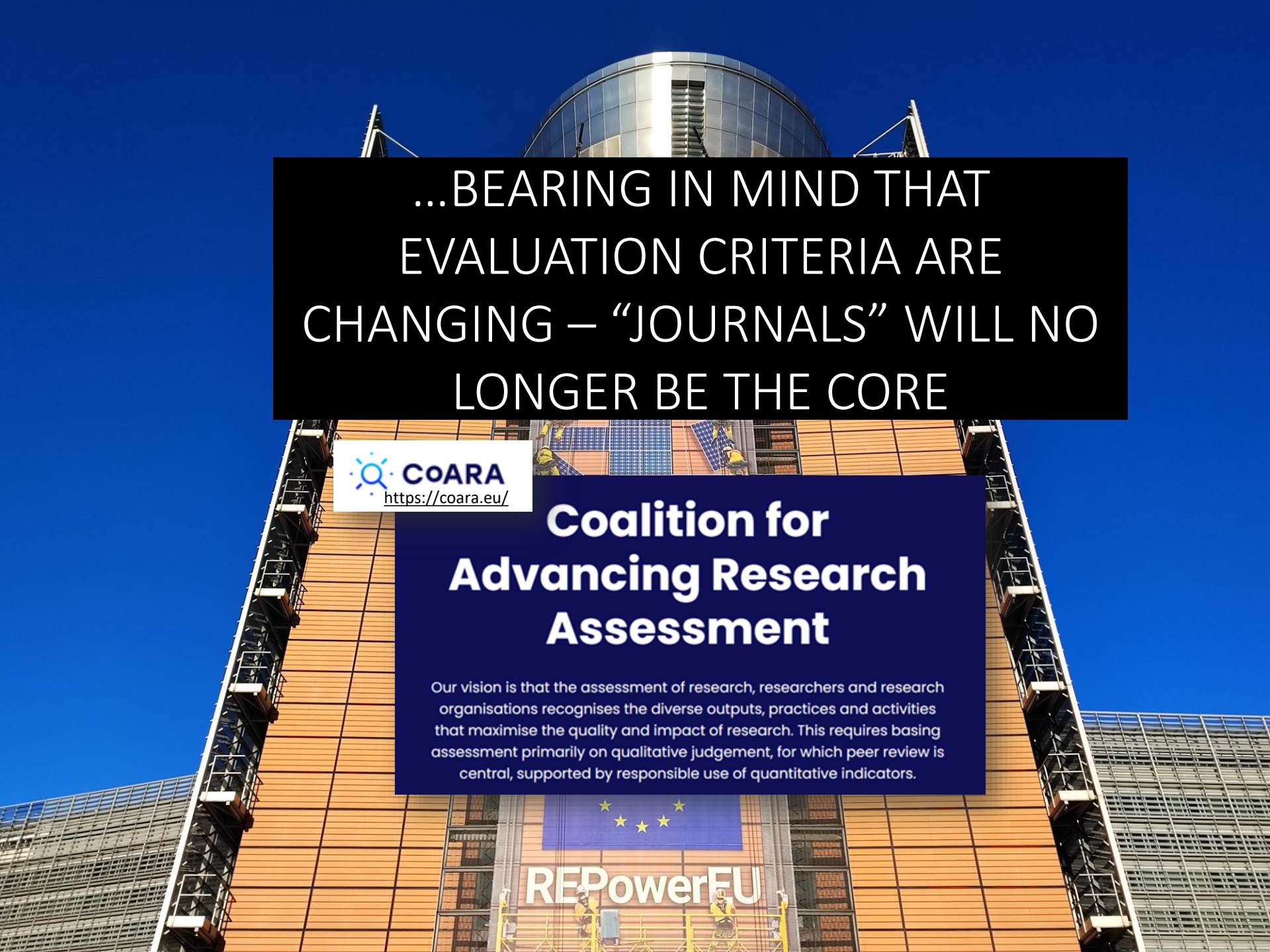


DEPOSIT [INCLUDED] → IMMEDIATE OPEN → DATA/INFO [INCLUDED] → COMPLIANT

PLUS:

- GRATIS
- OPEN PEER REVIEW
- INDEXING
- NOT TO BE IN THE BUDGET
- OPEN PRACTICE
- MAX IMPACT

…BEARING IN MIND THAT EVALUATION CRITERIA ARE CHANGING – “JOURNALS” WILL NO LONGER BE THE CORE



2. Publishing on an Open Access journal [Gold o Diamond]

Three tips to choose a publishing venue using the Directory of Open Access Journals (DOAJ)

Published on January 11, 2021

Jan. 11, 2021

Andrea Chiarelli
Senior Consultant at Research Consulting | Enhancing the effectiveness and impact of research
4 articles ✓ Following

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS
Apply News About For Publishers API

DOAJ

> 17.000
FULL OPEN ACCESS

DEPOSIT
[UP TO YOU]

- IR
- ZENODO

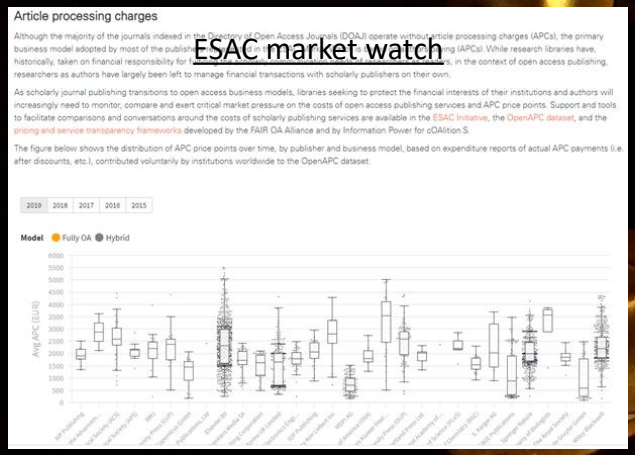
IMMEDIATE OPEN

COSTS
?

DATA/INFO
[UP TO YOU]

- ZENODO
- [RE3DATA]

COMPLIANT



ESAC market watch

29% ASK FOR APCs
250-2900 $

- COSTS TO BE INCLUDED INTO YOUR BUDGET
- MEAN COST IN ESAC MARKET
- CHECK YOUR SPECIFIC JOURNAL

ELIGIBLE ONLY COSTS FOR
- FULL OPEN ACCESS (NO HYBRID)
- DIGITAL (NO PRINT FOR BOOKS)

3. Publishing on a traditional journal (subscritpion based)

DEPOSIT [UP TO YOU]

- IR
- ZENODO

IMMEDIATE OPEN

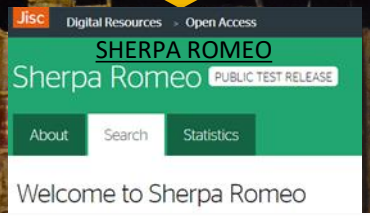
CAN I?

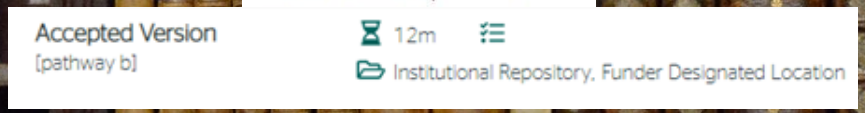
DATA/INFO [UP TP YOU]

- ZENODO
- [RE3DATA]

COMPLIANT

CHECK FOR EMBARGO (SHERPA ROMEO)





IF AN EMBARGO IS REQUIRED, YOU MUST RETAIN RIGHTS TO GIVE IMMEDIATE ACCESS IN THE ARCHIVE

IT'S A PRIOR OBLIGATION TO THE FUNDER

IN THE PROGRAMME GUIDE P.49 CLAUSE TO BE ADDED

3. Traditional journal

HYBRID APCs NOT ELIGIBLE

⚠ Publishing fees (including page charges or colour charges) for publications in other venues, for example in subscription journals (including hybrid journals) or in books that contain some scholarly content that is open and some that is closed are NOT eligible costs. Publishing fees for open access books may be eligible to the extent that they cover the first digital open access edition of the book (which could include different formats such as html, pdf, epub, etc.). Printing fees for monographs and other books are NOT eligible.

PRINT COSTS NOT ELIGIBLE («OPEN» ONLINE)

[reminder]

"WE DO NOT TELL RESEARCHERS WHERE TO PUBLISH, SO NOTHING IS PROHIBITED. HOWEVER, WE DO CARE WHERE WE SPEND TAXPAYER MONEY"

«TRUSTED REPOSITORY»

Trusted repositories are:

- Certified repositories (e.g. CoreTrustSeal, nestor Seal DIN31644, ISO16363) or disciplinary and domain repositories commonly used and endorsed by the research communities. Such repositories should be recognised internationally.
- General-purpose repositories or institutional repositories that present the essential characteristics of trusted repositories, i.e.:
  - display specific characteristics of organisational, technical and procedural quality such as services, mechanisms and/or provisions that are intended to secure the integrity and authenticity of their contents, thus facilitating their use and re-use in the short- and long-term. Trusted repositories have specific provisions in place and offer explicit information online about their policies, which define their services (e.g. acquisition, access, security of content, long-term sustainability of service including funding etc.).
  - provide broad, equitable and ideally open access to content free at the point of use, as appropriate, and respect applicable legal and ethical limitations. They assign persistent unique identifiers to contents (e.g. DOIs, handles, etc.), such that the contents (publications, data and other research outputs) are unequivocally referenced and thus citeable. They ensure that contents are accompanied by metadata sufficiently detailed and of sufficiently high quality to enable discovery, reuse and citation and contain information about provenance

facilitate mid- and long-term preservation of the deposited material. They have mechanisms or provisions for expert curation and quality assurance for the accuracy and integrity of datasets and metadata, as well as procedures to liaise with depositors where issues are detected. They meet generally accepted international and national criteria for security to prevent unauthorized access and release of content and have different levels of security depending on the sensitivity of the data being deposited to maintain privacy and confidentiality.

- INTEGRITY
- PRESERVATION
- SECURITY
- IDENTIFIERS
- REUSE/LICENSES

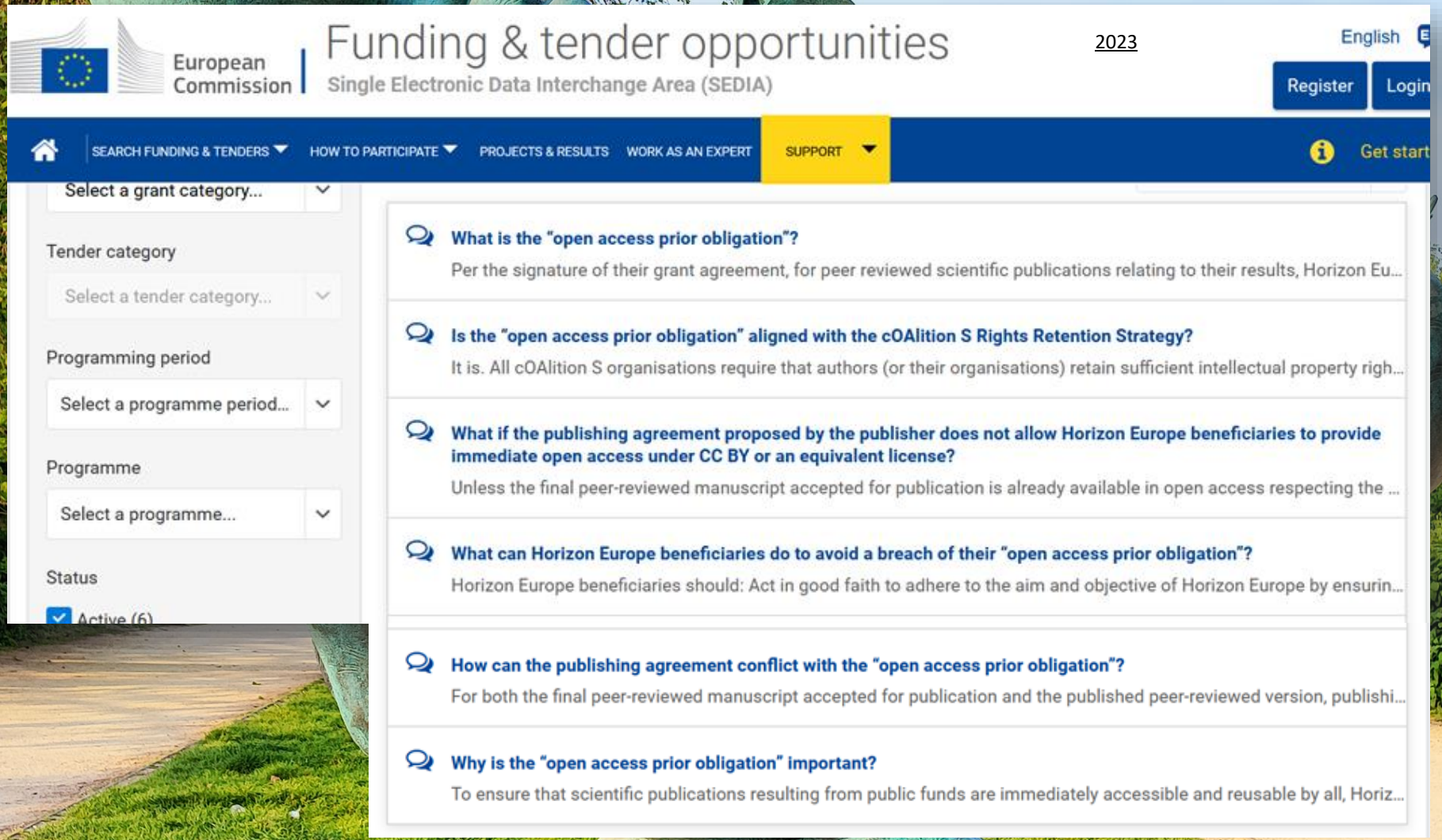
Rights retention clause

CLAUSE TO BE USED UPON SUBMISSION
[PRIOR OBLIGATION]

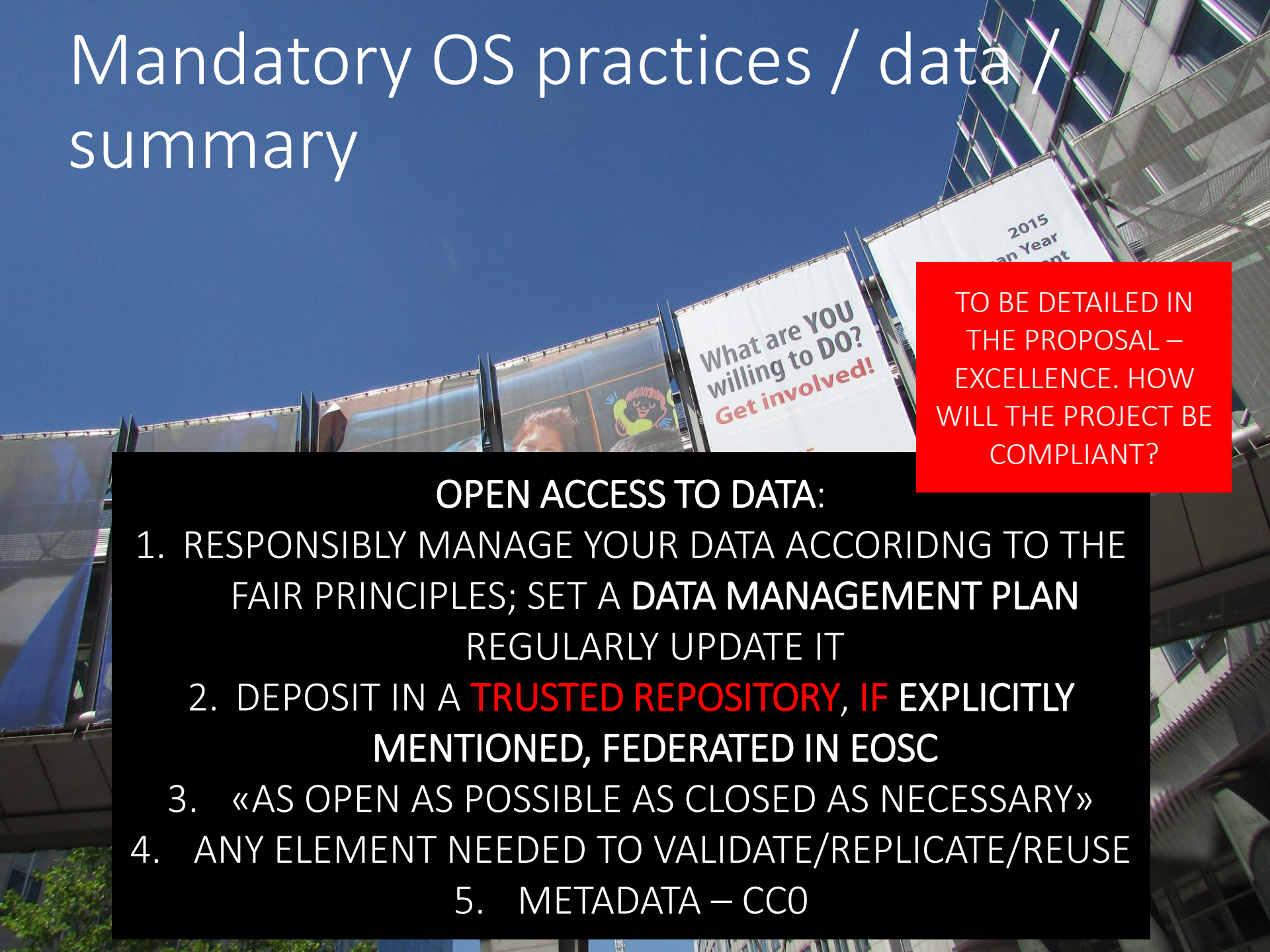
beneficiaries/researchers are encouraged to notify publishers of their grant agreement obligations (including the licensing requirements) already at manuscript submission. For example, by adding the following statement to their manuscript: "*This work was funded by the European Union under the Horizon Europe grant [grant number]. As set out in the Grant Agreement, beneficiaries must ensure that at the latest at the time of publication, open access is provided via a trusted repository to the published version or the final peer-reviewed manuscript accepted for publication under the latest available version of the Creative Commons Attribution International Public Licence (CC BY) or a licence with equivalent rights. CC BY-NC, CC BY-ND, CC BY-NC-ND or equivalent licenses could be applied to long-text formats.*" If the publishing agreement is contrary to the grant agreement obligations, authors should negotiate its terms and, alternatively, look for a different publishing venue/options.

IF THE PUBLISHERS REFUSES, LOOK FOR A DIFFERENT ONE!

Still in doubt?



Mandatory OS practices / data / summary



TO BE DETAILED IN THE PROPOSAL – EXCELLENCE. HOW WILL THE PROJECT BE COMPLIANT?

OPEN ACCESS TO DATA:

1. RESPONSIBLY MANAGE YOUR DATA ACCORIDNG TO THE FAIR PRINCIPLES; SET A **DATA MANAGEMENT PLAN** REGULARLY UPDATE IT
2. DEPOSIT IN A TRUSTED REPOSITORY, IF **EXPLICITLY MENTIONED, FEDERATED IN EOSC**
3. «AS OPEN AS POSSIBLE AS CLOSED AS NECESSARY»
4. ANY ELEMENT NEEDED TO VALIDATE/REPLICATE/REUSE
5. METADATA – CC0

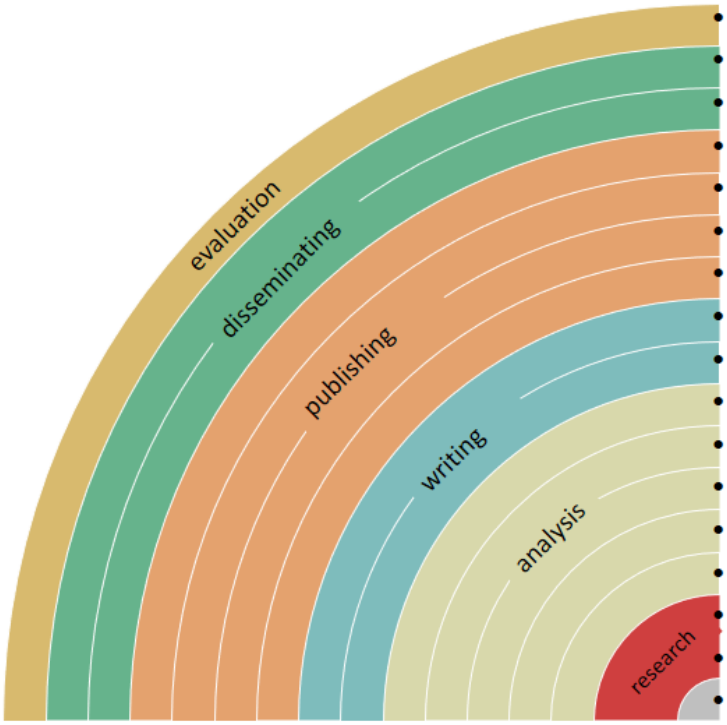
Mandatory OS practices / validation / summary

MEASURES TO ENSURE REPRODUCIBILITY:

1. INFORMATION ON TOOLS AND MATERIALS NEEDED TO VALIDATE THE RESULTS PUBLISHED IN THE PAPER
2. INFORMATION ON TOOLS AND MATERIALS NEEDED TO VALIDATE AND REUSE THE DATA
3. WHENEVER POSSIBLE, GIVE ACCESS TO IT

Recommended Open Science practices

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. Creative Commons BY
- self archiving in archives or publishing on Open journals
- using open peer review, e.g. at PubPeer o F1000
- sharing preprints, e.g. at OSFpreprint, arXiv o biorXiv
- using actionable formats, e.g. with Jupyter o CoCalc
- open XML-drafting, e.g. at Overleaf o Authorea
- sharing protocols & workflows, e.g. at Protocols.io
- sharing notebooks, e.g. at OpenLabNotebook
- sharing code, e.g. at GitHub licensing GNU/MIT
- sharing data, e.g. at Dryad, Zenodo o Dataverse
- pre-registering, e.g. at OSFregistry o AsPredicted
- commenting openly, e.g. with Hypothes.is o Pund.it
- using shared reference libraries, e.g. with Zotero
- sharing (grant) proposals, e.g. with RIO Journal

Interactive rainbow 2021

[Guide]

PROGRAMME GUIDE, p.41-42

- EARLY SHARING
- FAIR DATA MANAGEMENT
- REPRODUCIBILITY
- OPEN ACCESS
- OPEN PEER REVIEW
- CITIZIEN SCIENCE

Early and open sharing: Provide specific information on whether and how you will implement early and open sharing and for which part of your expected output. For example, you may mention what type of early and open sharing is appropriate for your discipline and project, such as preprints or preregistration/registration reports, and which platforms you plan to use.

Research data management (RDM): RDM is mandatory in Horizon Europe for projects generating or reusing data. If you expect to generate or reuse data and/or other research outputs (except for publications), you are required to outline in a maximum of one page how these will be managed. Further details on this are provided

Reproducibility of research outputs: you should outline the measures planned in the project that tend to increase reproducibility. Such measures may already be interweaved in other parts of the methodology of a proposal (such as transparent research design, the robustness of statistical analyses, addressing negative results, etc) or in mandatory/non-mandatory open science practices *(e.g. the DMP, early sharing through preregistration and preprints, open access to software, workflows, tools, etc)* to be implemented. More detailed suggestions on good practices for enhancing

Open access: Offer specific information on how you will meet the open access requirements, that is deposition and immediate open access to publications and open access to data (the latter with some exceptions and within the deadlines set in the DMP) through a trusted repository, and under open licenses. You may elaborate on the (subscription-based or open access) publishing venues that you will use. You may also

Open peer review: Anytime it is possible, you are invited to prefer open peer review for your publications over traditional ('blind' or 'closed') peer review. When the case, you should provide specific information regarding the publishing venues you envisage to make use of, and highlight the venues that would qualify as providing open peer review.

Citizen, civil society and end-user engagement: Provide clear and succinct information on how citizen, civil society and end-user engagement will be implemented in your project, where/if appropriate. The kinds of engagement activities will depend on the type of R&I activity envisaged and on the disciplines and sectors implicated.

MSCA Application form – Part A

2022

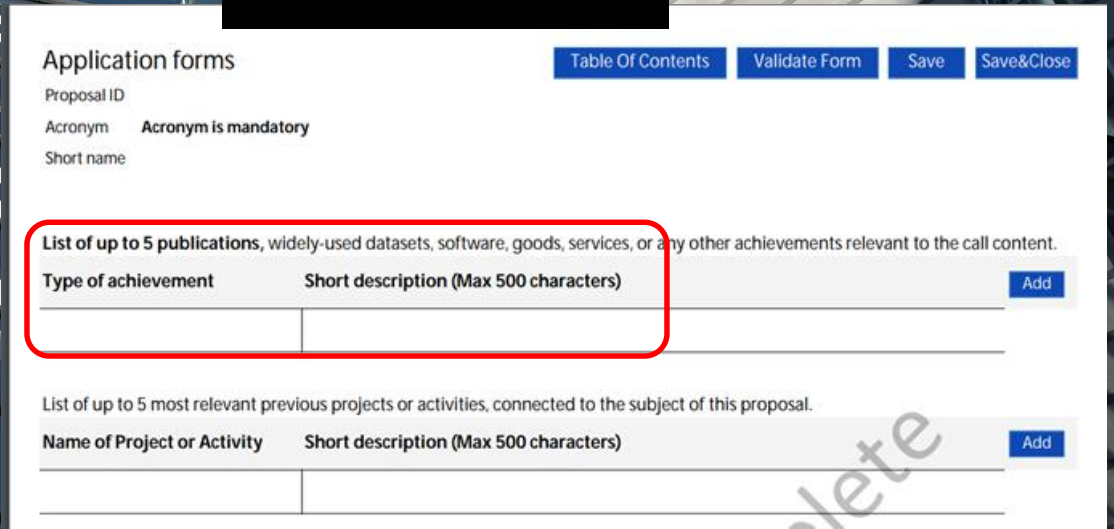
Horizon Europe Programme

Marie Skłodowska-Curie Actions
Postdoctoral Fellowships (HE MSCA PF)

Application form (Part A)
Project proposal – Technical description (Part B)

Version 1.1
5 May 2022

PART A – 5 ACHIEVEMENTS



MSCA App

2022

Horizon Europe Programme

Marie Skłodowska-Curie Actions
Postdoctoral Fellowships (HE MSCA PF)

Application form (Part A)
Project proposal – Technical description (Part B)

Version 1.1
5 May 2022

Part B-1

1. Excellence

1.1 Quality and pertinence of the project's research and innovation objectives (and the extent to which they are ambitious, and go beyond the state of the art)

At a minimum, address the following aspects:

- Describe the quality and pertinence of the R&I objectives; are the objectives measurable and verifiable? Are they realistically achievable?
- Describe how your project goes beyond the state-of-the-art, and the extent to which the proposed work is ambitious.

1.2 Soundness of the proposed methodology (including interdisciplinary approaches, consideration of the gender dimension and other diversity aspects if relevant for the research project, and the quality of open science practices)

At a minimum, address the following aspects:

- Open science practices: Describe how appropriate open science practices are implemented as an integral part of the proposed methodology. Show how the choice

gy, including the
ain how this will
nt challenges you
end to overcome

that will increase the chances of the project delivering on its objectives *[e.g. up to 1/2 page, including research data management]*. If you believe that none of these practices are appropriate for your project, please provide a justification here.

Open science is an approach based on open cooperative work and systematic sharing of knowledge and tools as early and widely as possible in the process. Open science practices include early and open sharing of research (for example through pre-registration, registered reports, pre-prints, or crowd-sourcing); research output management; measures to ensure reproducibility of research outputs; providing open access to research outputs (such as publications, data, software, models, algorithms, and workflows); participation in open peer-review; and involving all relevant knowledge actors including citizens, civil society and end users in the co-creation of R&I agendas and contents (such as citizen science).

Please note that this does not refer to outreach actions that may be planned as part of the communication, dissemination and exploitation activities. These aspects should instead be described below under 'Impact'.

MSCA Application form – Part B1

- Open science practices: Describe how appropriate open science practices are implemented as an integral part of the proposed methodology. Show how the choice of practices and their implementation is adapted to the nature of your work in a way

[2] Interdisciplinarity means the integration of information, data, techniques, tools, perspectives, concepts or theories from two or more scientific disciplines.

2022

Horizon Europe Programme

Marie Skłodowska-Curie Actions
Postdoctoral Fellowships (HE MSCA PF)

Application form (Part A)
Project proposal – Technical description (Part B)

Version 1.1
5 May 2022

1-2 PAGES ON APPROPRIATE OPEN SCIENCE PRACTICES

that will increase the chances of the project delivering on its objectives *[e.g. up to 1/2 page, including research data management]*. If you believe that none of these practices are appropriate for your project, please provide a justification here.

Open science is an approach based on open cooperative work and systematic sharing of knowledge and tools as early and widely as possible in the process. Open science practices include early and open sharing of research (for example through pre-registration, registered reports, pre-prints, or crowd-sourcing); research output management; measures to ensure reproducibility of research outputs; providing open access to research outputs (such as publications, data, software, models, algorithms, and workflows); participation in open peer-review; and involving all relevant knowledge actors including citizens, civil society and end users in the co-creation of R&I agendas and contents (such as citizen science).

Please note that this does not refer to outreach actions that may be planned as part of the communication, dissemination and exploitation activities. These aspects should instead be described below under 'Impact'.

[...real life]

..THE RISK IS HAVING A DMP ILLUSTRATING HOW AND WHEN A PRIVACY CONSENT FORM WILL BE SIGNED... INA A RESEARCH WITH MICE...[A POLISH COLLEAGUE TOLD ME]

...I HOPE IT'S CLEAR THIS IS PRECISELY WHAT YOU **DO NOT** HAVE TO DO...

«COME ON, IT'S LIKE HORIZON 2020... UHM NO? DID THEY ADD OPEN SCIENCE? SO YOU WRITE A PAGE ON OPEN SCIENCE AND THEN WE ALL COPY/PASTE, YOU WRITE A DMP SCHEMA AND THEN WE ALL COPY/PASTE»

MSCA Application form – Part B1

2022

Horizon Europe Programme

Marie Skłodowska-Curie Actions
Postdoctoral Fellowships (HE MSCA PF)

Application form (Part A)
Project proposal – Technical description (Part B)

Version 1.1
5 May 2022

- Research data management and management of other research outputs: Applicants generating/collecting data and/or other research outputs (except for publications) during the project must explain how the data will be managed in line with the FAIR principles (Findable, Accessible, Interoperable, Reusable).

- *For guidance on open science practices and research data management, please refer to the relevant section of the HE Programme Guide on the Funding & Tenders Portal.*

INCLUDING FAIR DATA MANGEMENT
[IF YOU COLLECT OR GENERATE DATA]

MSCA action Application form

2022

2. Impact

2.1 Credibility of the measures to enhance the career perspectives and employability of the researcher and contribution to his/her skills development

At a minimum, address the following aspects:

- **Expected** skill development of the researcher.
- **Expected** impact of the proposed research and training activities on the researcher's career perspectives inside and/or outside academia.

2.2 Suitability and quality of the measures to maximise expected outcomes and impacts, as set out in the dissemination and exploitation plan, including communication activities

At a minimum, address the following aspects:

- Plan for the dissemination and exploitation activities, including communication activities:* Describe the planned measures to maximize the impact of your project by providing a first version of your 'plan for the dissemination and exploitation including communication activities'. Describe the dissemination, exploitation measures that are planned, and the target group(s) addressed (e.g. scientific community, end users, financial actors, public at large). Regarding communication measures and public engagement strategy, the aim is to inform and reach out to society and show the

- Strategy for the management of intellectual property, foreseen protection measures if relevant, discuss the strategy for the management of intellectual property, foreseen protection measures, such as patents, design rights, copyright, trade secrets, etc., and how these would be used to support exploitation.

DISSEMINATION & EXPLOITATION PLAN

[THERE IS NO CONFLICT BETWEEN OPEN/PATENTS]

MSCA Application form – Part B1

Be specific, referring to the effects of your project, and not R&I in general in this field. State the target groups that would benefit.

- Expected scientific impact(s): e.g. contributing to specific scientific advances, across and within disciplines, creating new knowledge, reinforcing scientific equipment and instruments, computing systems (i.e. research infrastructures);
- Expected economic/technological impact(s): e.g. bringing new products, services, business processes to the market, increasing efficiency, decreasing costs, increasing profits, contributing to standards' setting, etc.
- Expected societal impact(s): e.g. decreasing CO2 emissions, decreasing avoidable mortality, improving policies and decision-making, raising consumer awareness.

2022

Horizon Europe Programme

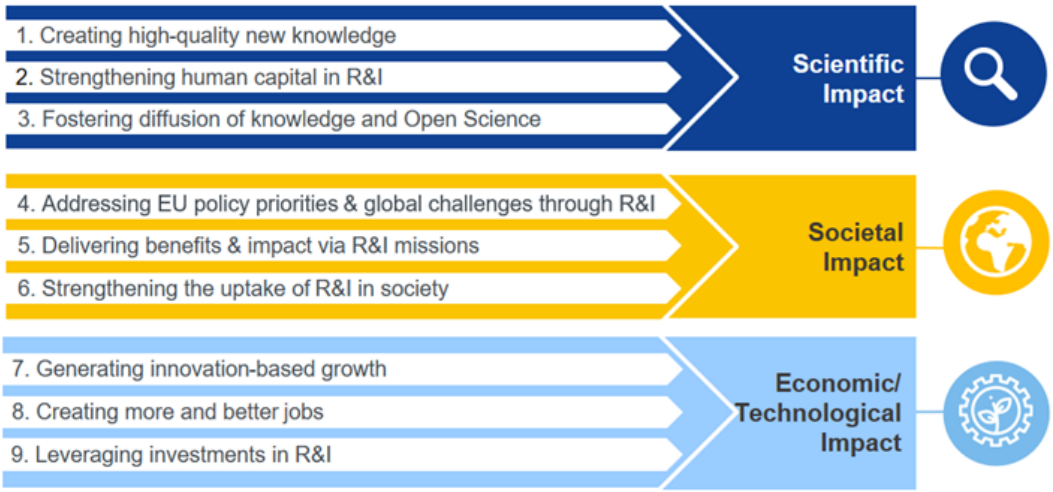
Marie Skłodowska-Curie Actions
Postdoctoral Fellowships (HE MSCA PF)

Application form (Part A)
Project proposal – Technical description (Part B)

Version 1.1
5 May 2022

March 24, 2021

HORIZON EUROPE LEGISLATION defines three types of impact, tracked with Key Impact Pathways



Article 50 & Annex V *'Time-bound indicators to report on an annual basis on progress of the Programme towards the achievement of the objectives referred to in Article 3 and set in Annex V along impact pathways'*

European Commission

EXPECTED IMPACT – DO NOT FORGET THE IMPACT PATHWAYS
[AMONG WHICH, OPEN SCIENCE!]

MSCA application form – Part B2


2022

Horizon Europe Programme

Marie Skłodowska-Curie Actions
Postdoctoral Fellowships (HE MSCA PF)

Application form (Part A)
Project proposal – Technical description (Part B)

Version 1.1
5 May 2022

Part B2 (no overall page limit applied)

4. CV of the researcher (indicative length: 5 pages)

Any information provided in Parts A and B of the proposal should be fully consistent. Always mention full dates (using format: dd/mm/yyyy). The CV should include the standard academic and research record. Any research career gaps and/or unconventional paths should be clearly explained.

At a minimum, the CV should contain:

a) The name of the researcher;
b) Professional experience (most recent first, with exact dates in format dd/mm/yyyy);
c) Education, including PhD award date (most recent first, with exact dates in format: dd/mm/yyyy).

CV should include information on:

Publications in peer-reviewed scientific journals, peer-reviewed conference proceedings, and/or monographs (they are expected to be open access either published or through repositories) and other outputs such as data, software, algorithms significant for your research path (they are expected to be open access in appropriate repositories to the extent possible; they should be accompanied by a very short qualitative assessment of their scientific significance and not by the Journal Impact Factor);

Invited presentations to internationally established conferences and/or international advanced schools;

IN YOUR CV
- PUBLICATIONS ARE SUPPOSED TO BE OPEN (PUBLISHED OR DEPOSITED)
- SCIENTIFIC IMPACT NOT BY IMPACT FACTOR

…once the project is approved…

BY MONTH 6 YOU HAVE TO PROVIDE
A DATA MANAGEMENT PLAN

What if I generate no data?

Research data management (RDM): RDM is mandatory in Horizon Europe for projects generating or reusing data. If you expect to generate or reuse data and/or other research outputs (except for publications), you are required to outline in a maximum of one page how these will be managed. Further details on this are provided

YOU SIMPLY DON'T HAVE TO DRAFT ANY DATA MANAGEMENT PLAN!
JUST STATE IN THE PROPOSAL THAT YOUR PROJECT IS NOT GOING TO GENERATE DATA

IF YOU GENERATE SOFTWARE, THEN THIS IS AN OUTPUT TO BE DEPOSITED (IN GITHUB?) AND ADDRESSED IN A SHORT DMP

THANK YOU!