Innovation Acta FOR GlyCanDrug

Data management basics

Elena Giglia

Soft Skills courses

Sevilla 3rd conference – November 20th/21st, 2025

Data management basics

Innovation Acta

GlyCanDrug

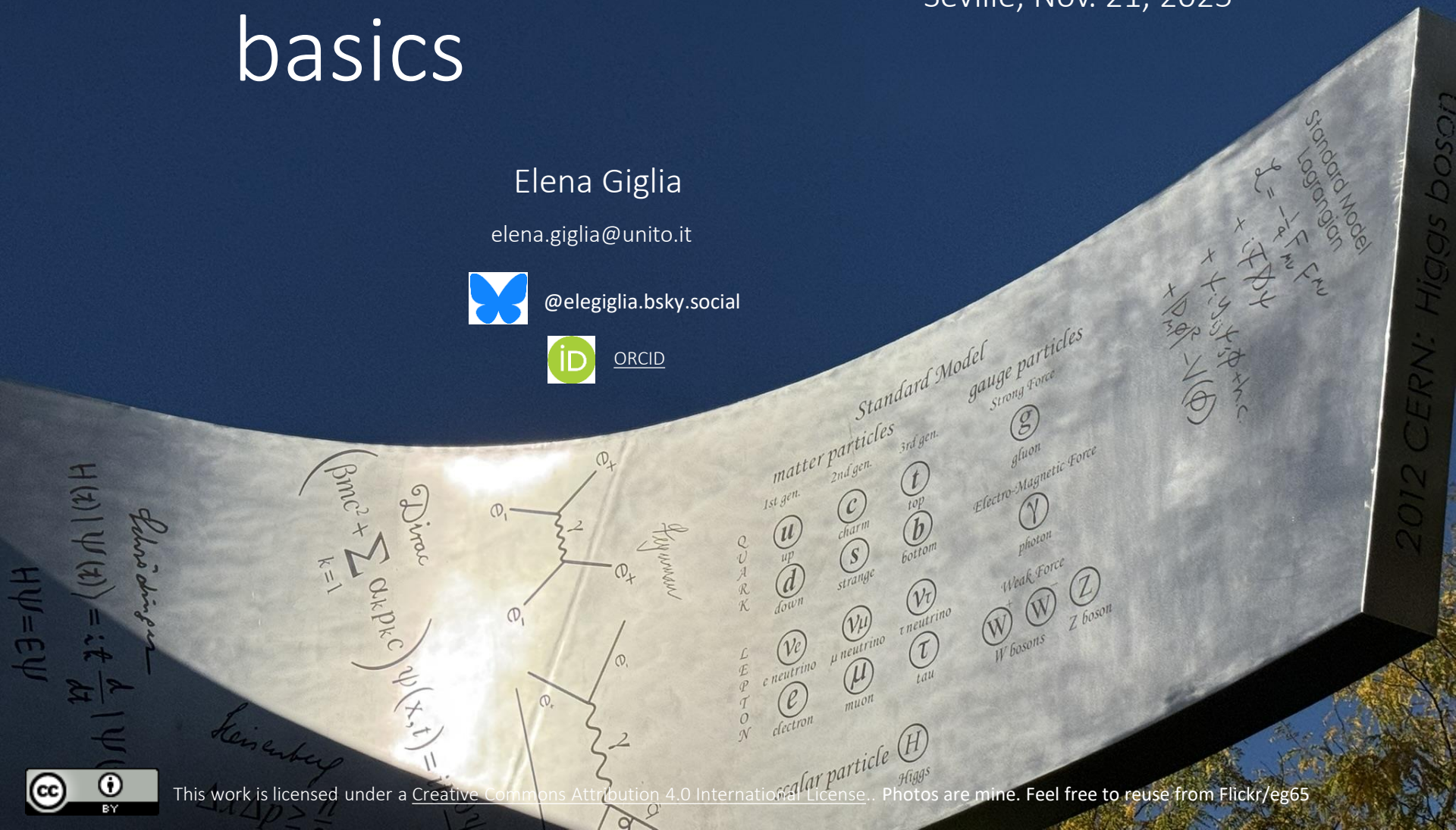
Seville, Nov. 21, 2025

Elena Giglia

elena.giglia@unito.it

@elegiglia.bsky.social

ORCID



This work is licensed under a Creative Commons Attribution 4.0 International License.. Photos are mine. Feel free to reuse from Flickr/eg65

Housekeeping

TARGET:

- GIVE AN OVERVIEW OF THE COMPLEXITY OF DATA MANAGEMENT [MISSION IMPOSSIBLE IN 3 HOURS!]
- PROVIDE SOME BASIC TOOLS

QUESTIONS WILL BE TAKEN AT THE END OF THE FIRST PART, BUT YOU CAN NOTE THEM DOWN WHILE I'LL BE SPEAKING

SLIDES ARE AVAILABLE ON ZENODO

Why are we here today?

OPEN SCIENCE IS THE «NEW NORMAL»

AND FAIR DATA ARE EOSC BUILDING BLOCKS

FAIR Open data, just an administrative burden?

NO. OPEN DATA 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)

Nov. 29, 2021

November 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

…that’s it.
Thank you for
your undivided
attention

The framewo ence

OPEN SCIENCE≠OPEN ACCESS

OPENING EVERY STEP
- AS **EARLY** AS POSSIBLE
- AS **OPEN** AS POSSIBLE
- AS **FAIR** AS POSSIBLE

FOCUS ON THE ENTIRE PROCESS, NOT ONLY THE FINAL SYNTHESIS (ARTICLE)

TODAY WE ARE DEALING WITH **FAIR OPEN DATA**

OPEN SCIENCE IS NOT THE TARGET. THE GOAL IS A SOUNDER SCIENCE, MORE TRANSPARENT AND RESPONSIVE TO SOCIETAL NEEDS

…to cut it short…

4 nov 2025

Enrica Porcari • 1st

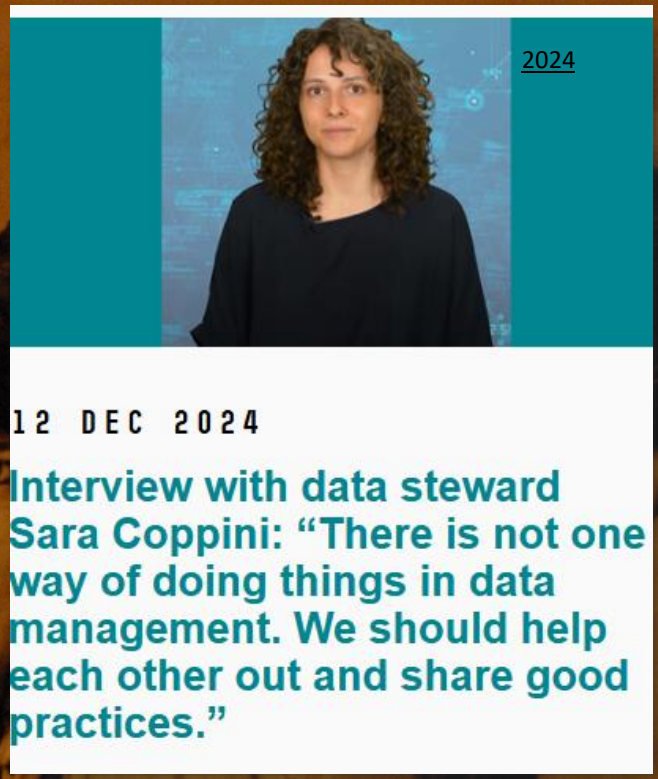
Head of IT, CERN | 30+ Years in Technology &…

1d • Edited •

The future of science will be open — or it won't be at all.

The question is no longer whether we can afford to invest in open science —
It's whether we can afford not to.

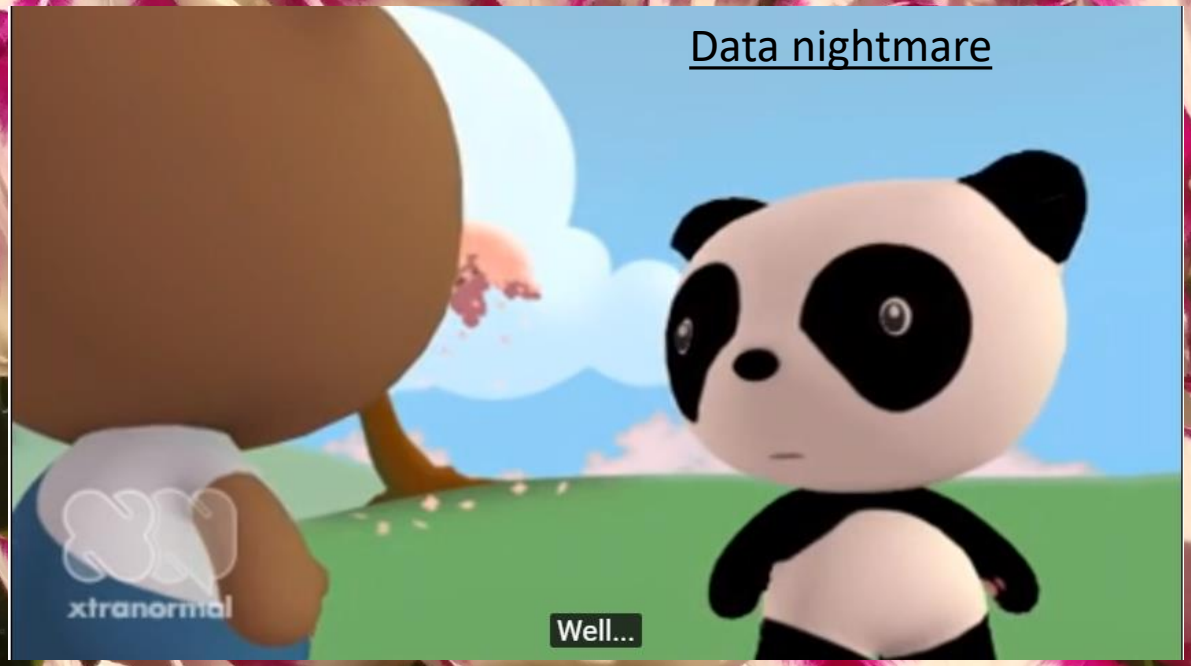
… I see…



1) MANAGING DATA IS NOT AN EASY TASK
2) THERE IS NO RECIPE AS DATA ARE UNIQUE
3) SO MANY ASPECTS TO BE CONSIDERED
4) SO MANY TOOLS TO GET USED TO
5) IT SEEMS SOOOOOOOO TIME CONSUMING
6) BUT THE BENEFITS ARE HUUUUUUUUUUGE

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:

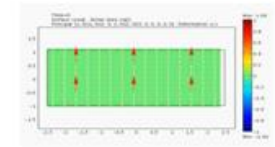
- DATA GENERATED WITH PUBLIC FUNDS
- PUBLISHED IN «SCIENCE» (DATA POLICY)
- REQUESTED FROM A DIFFERENT DISCIPLINE

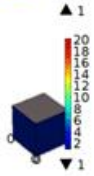
Why should we care about data? A personal view

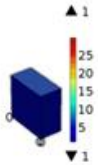
Past scientific interests

Mathematical models for soft-active materials

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

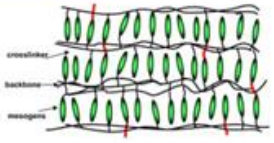


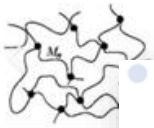




M. de Luca, A. DeSimone. Elastomeric Gels: A Model and First Results. Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems. Lecture Notes in Applied and Computational Mechanics, vol 81. Springer, Cham. (2016) https://doi.org/10.1007/978-3-319-39022-2_4

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

This work © 2023 by Mariarita de Luca is licensed under CC BY 4.0

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?

HARVARD UNIVERSITY

Vision

Research Data Management @Harvard

Vision

Research data are an important asset to our University and our researchers.

AS THEY ARE THE FOUNDATION OF RESPONSIBLE RESEARCH

DCC because good research needs good data

…that’s it.
Thank you for your undivided attention
[AGAIN!]

…ever happened…

TO LOSE YOUR DATA?

TO HAVE THE DATA ON THE PHD COMPUTER – AND THEN THE PHD LEFT (WITH THE DATA)??

TO OPEN A DATASET AND NOT BEING ABLE TO UNDERSTAND THE DATA??

TO ASK DATA AFTER READING AN ARTICLE AND THE AUTHOR WAS NOT ABLE TO FIND THEM??

NOT BEING ABLE TO OPEN A DATASET BECAUSE OF AN OBSOLETE FORMAT?

…OR SIMPLY BEING SO HAPPY WHEN YOU FIND A REUSABLE DATABASE?

Why should

1. ABSOLUTE SELF-INTEREST. YOU SAVE TIME AND STREAMLINE YOUR JOB

How does good RDM support researchers in their daily work?

Good RDM practices can save researchers time facilitate collaboration, promote data quality and reproducibility, and help ensure compliance with funder and publisher requirements, all of which can contribute to more effective and impactful research.

2024

- **Saving you time** Doing the work now to plan for your expected data, back them up and document them in detail preserves time that is otherwise lost in searching for, recovering and deciphering data in the future.

2025

Why should we care about data?

2.1 WE HAVE TO.
OPEN DATA DIRECTIVE + OPEN DATA ACT

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

- Stimulate the publishing of dynamic data and the uptake of Application Programme Interfaces (APIs).
- Limit the exceptions which currently allow public bodies to charge more than the marginal costs of dissemination for the re-use of their data.
- Enlarge the scope of the Directive to:
  - data held by public undertakings, under a specific set of rules. In principle, the Directive will only apply to data which the undertakings make available for re-use. Charges for the re-use of such data can be above marginal costs for dissemination;
  - research data resulting from public funding – Member States will be asked to develop policies for open access to publicly funded research data. New rules will also facilitate the re-usability of research data that is already contained in open repositories.
- Strengthen the transparency requirements for public–private agre... arra...

Official Journal of the European Union EN Series L

Data act

2023/2854 22.12.2023

REGULATION (EU) 2023/2854 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 13 December 2023

on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act)

(Text with EEA relevance)

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

2024

Council of the European Union

Brussels, 14 May 2024
(OR. en)

9097/1/24
REV 1

POLCOM 167
COMER 68
RELEX 547
DUAL USE 35
RECH 173
ENER 188
ENV 433

LEGISLATIVE ACTS AND OTHER INSTRUMENTS

Subject: COUNCIL RECOMMENDATION on enhancing research security

TO BE COUPLED WITH RECOMMENDATION ON RESEARCH SECURITY

«OPEN DATA» DIRECTIVE ENLARGED TO RESEARCH DATA.
THEY MUST BE OPEN [AS OPEN AS POSSIBILE]

Why should we care about data?

ANNEX 5

SPECIFIC RULES

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:

- establish a data management plan ('DMP') (and regularly update it)
- as soon as possible and within the deadlines set out in the DMP, deposit the data in a trusted repository; if required in the call conditions, this repository must be federated in the EOSC in compliance with EOSC requirements
- as soon as possible and within the deadlines set out in the DMP, ensure open access — via the repository — to the deposited data, under the latest available version of the Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a licence with equivalent rights, following the principle 'as open as possible as closed as necessary', unless providing open access would in particular:
  - be against the beneficiary's legitimate interests, including regarding commercial exploitation, or
  - be contrary to any other constraints, in particular the EU competitive interests or the beneficiary's obligations under this Agreement; if open access is not provided (to some or all data), this must be justified in the DMP

V.1 Feb 2021

Horizon Europe (HORIZON)
Euratom Research and Training Programme (EURATOM)

General Model Grant Agreement
EIC Accelerator Contract

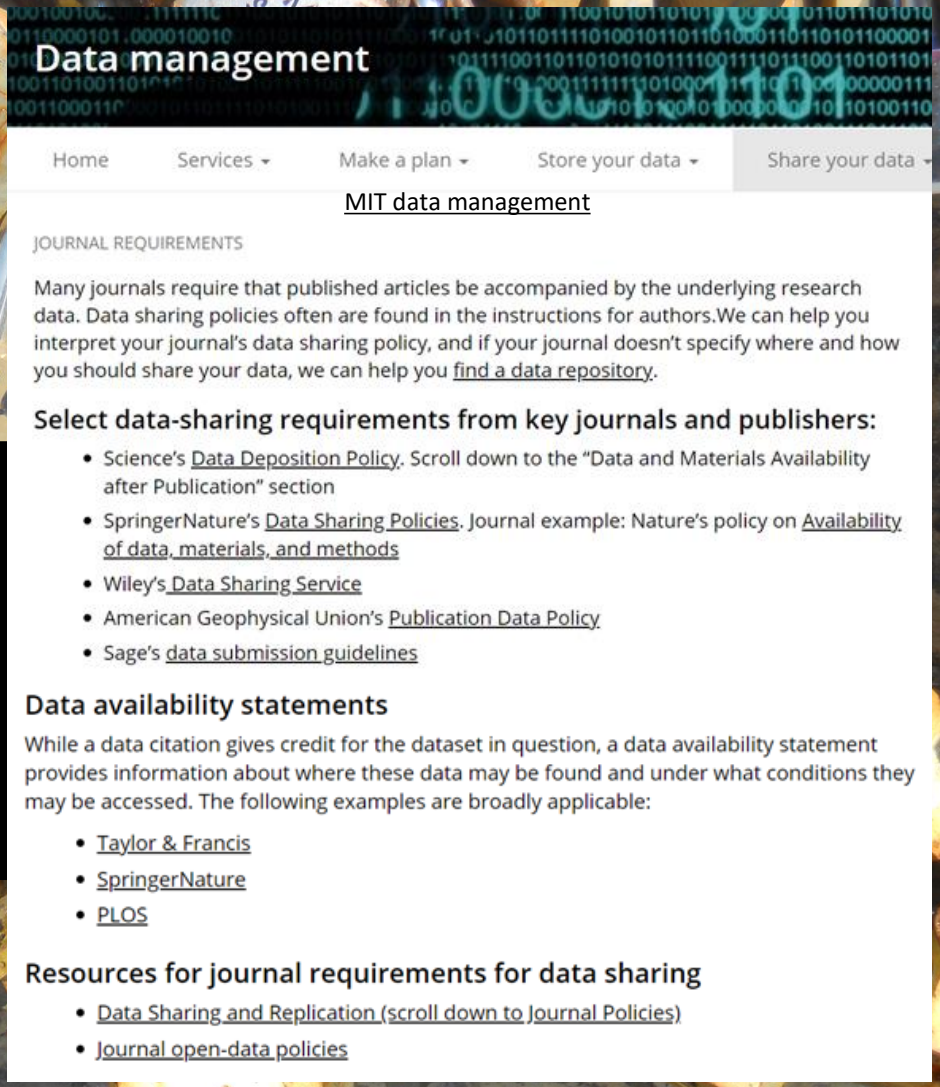
(HE MGA — Multi & Mono)

2.2 WE HAVE TO.
HORIZON EUROPE

RESPONSIBLY MANAGED [DMP]
FAIR IN A TRUSTED REPOSITORY [IN EOSC]
AS OPEN AS POSSIBLE AS CLOSED AS NECESSARY

Why should we care about data?

2.3. WE HAVE TO.
A GROWING NUMBER OF
JOURNALS/PUBLISHERS
ASK FOR DATA TO BE PUBLISHED
WITH THE ARTICLE FOR
REPRODUCIBILITY AND
TRANSPARENCY PURPOSES



Data management

Home | Services | Make a plan | Store your data | Share your data

MIT data management

JOURNAL REQUIREMENTS

Many journals require that published articles be accompanied by the underlying research data. Data sharing policies often are found in the instructions for authors.We can help you interpret your journal's data sharing policy, and if your journal doesn't specify where and how you should share your data, we can help you find a data repository.

Select data-sharing requirements from key journals and publishers:

- Science's Data Deposition Policy. Scroll down to the "Data and Materials Availability after Publication" section
- SpringerNature's Data Sharing Policies. Journal example: Nature's policy on Availability of data, materials, and methods
- Wiley's Data Sharing Service
- American Geophysical Union's Publication Data Policy
- Sage's data submission guidelines

Data availability statements

While a data citation gives credit for the dataset in question, a data availability statement provides information about where these data may be found and under what conditions they may be accessed. The following examples are broadly applicable:

- Taylor & Francis
- SpringerNature
- PLOS

Resources for journal requirements for data sharing

- Data Sharing and Replication (scroll down to Journal Policies)
- Journal open-data policies

[why we need FAIR Open data and Open Science?]

Retraction Watch
6 gennaio alle ore 10:00 ·

Science paper retracted by authors: All but one figure lacked "supporting data."
https://www.science.org/doi/10.1126/sciadv.adv4615

Biotech company agrees to pay $4 million to settle data falsification allegations

2025

A biotech company whose CEO faced allegations of manipulating data in papers used in NIH grant applications will pay a settlement of $4 million to resolve those allegations, the Department of Justice announced January 6.

Journals investigating dozens of papers by leading Canadian urologists

2024 Dec 23

A high-profile Canadian urologist received an editorial expression of concern for one of his papers this month, after anonymous comments on PubPeer flagged suspected data duplication in dozens of his articles.

By our count, sleuths have flagged 30

Research Article

Money and inflation: a case study of the v transparency

2025

Gary Smith
Published online: 20 Dec 2024

Cite this article | https://doi.org/10.1080/01603477.2024.2444246 | Check for updates

Full Article | Figures & data | Read this article

TRANSPARENCY, RESEARCH INTEGRITY, REPRODUCIBILITY

Abstract

Hanke reported that there is "a tight linkage between changes in the money supply (M2) and inflation." That paper illustrates clearly why journals should require authors to make all of their data and methods publicly available *before* their papers are published—indeed before their papers are sent out for review.

Why sh

UNLOCK THE FULL POTENTIAL OF RESEARCH DATA
TO ACCELERATE DISCOVERIES AND INNOVATION

eosc EOSC Strategy – Status Current Thinking

What

EOSC is a web of FAIR data and related services for research
Research data that is easy to find, access, interoperate and reuse (FAIR)
Trusted and sustainable research outputs are available within and across scientific disciplines

Why

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

How

Access and interoperability of research data and results

- Define ownership, authorship and responsibility of data and research outputs
- Ensure long-term preservation of data throughout its lifecycle
- Enable the creation of standards for all research domains
- Make data machine-actionable
- …able new scientific discovery methods
- …d science disciplines
- …in researchers on adopting FAIR …nciples as an integral part in their activity

A sustainable coordinated infrastructure

- Establish and maintain a coordinated federated reference architecture
- Implement an operational infrastructure framework that is long term sustainable
- Ensure high quality of data and services
- Ensure secure access to data and services
- Define clear standards for API and interoperability of data and services
- Apply user friendly practices
- Inspire EOSC ambassadors to assist in on-boarding of researchers

Inspired people and robust governance

- Communicate an inspiring EOSC vision and strategy
- Implement an unambiguous and clearly mandated governance structure
- Establish a framework to engage human capital in institutions, countries and scientific communities
- Enable disciplinary and cross-disciplinary transnational research to find new insights from existing and new research data and outputs

3. BECAUSE WE HAVE EOSC

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…

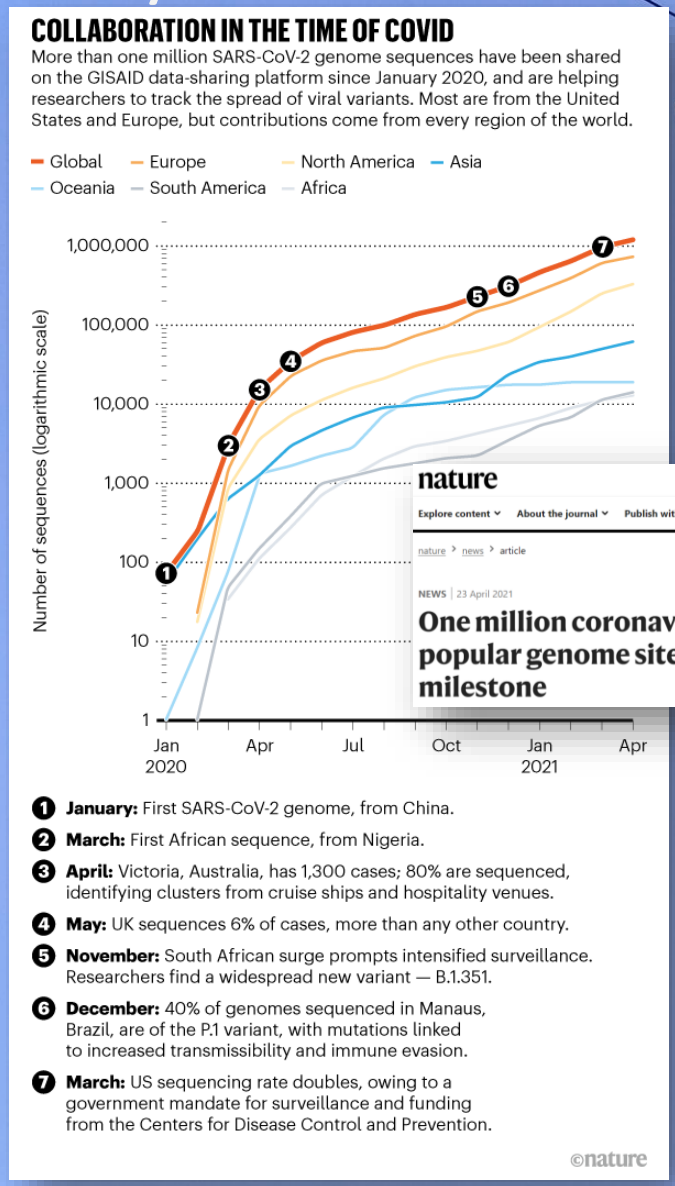
A SUPPORTING ENVIRONMENT FOR OPEN SCIENCE AND NOT AN «OPEN CLOUD» FOR SCIENCE

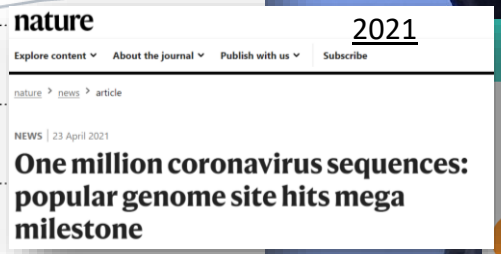
YOU DON'T «UPLOAD» YOUR DATA INTO EOSC

AND GIVE SEAMLESS ACCESS TO 20 M EU RESEARCHERS

[Nov. 3rd 2025]

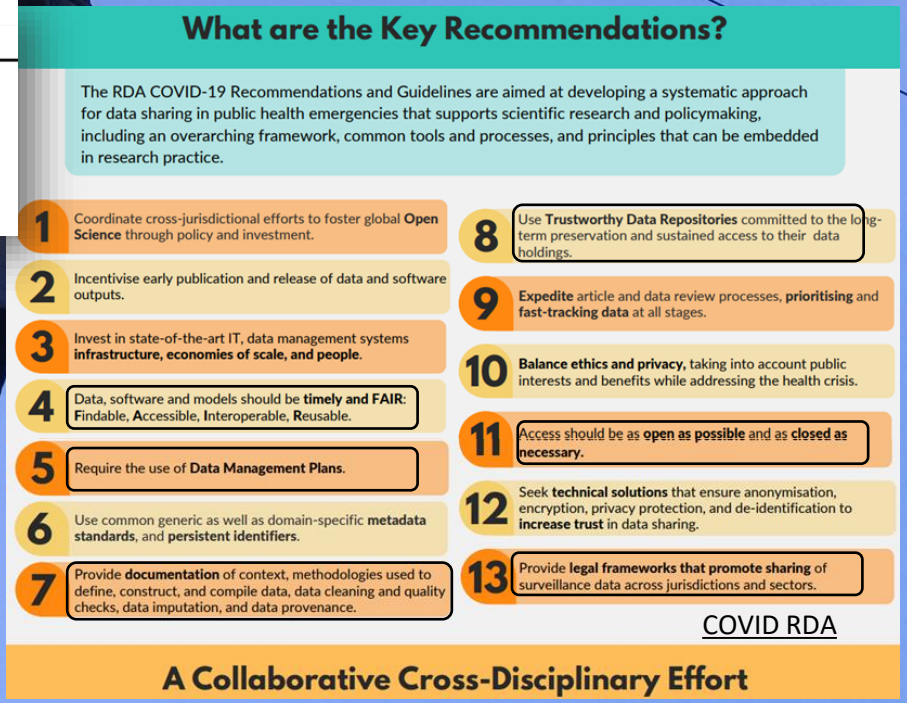
Why should we care about data?





2021

4. IT'S IN OUR INTEREST.
IT WAS CLEAR DURING THE COVID CRISIS



COVID RDA

Why should we care about data?

5. IT'S IN OUR INTEREST

LOST VALUE IF DATA ARE MISSING:

- AT BEST: EXPENSIVE RESEARCH IS OF LITTLE OR NO VALUE
- AT WORST: RESULTS OF INVALID RESEARCH ARE PUT INTO CLINICAL USE

Great values lost by sharing data

Lack of reproducibility well known problem in medical research.

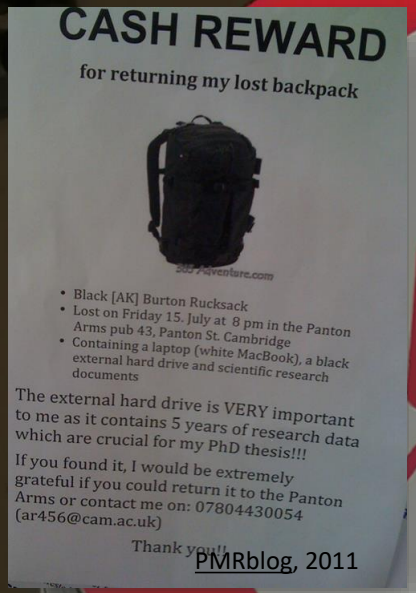
Investigations in the US: Up to 50% of studies not reproducible. 25% of this caused by unavailability of data.

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?



PMRblog, 2011

6. IT'S IN OUR INTEREST. DATA GET LOST

March 10, 2021

S.Aliprandi, Sicurezza dati e privacy (le norme) 2017

Why should we care about data?

Scientists losing data at a rapid rate

2013

Decline can mean 80% of data are unavailable after 20 years.

Elizabeth Gibney & Richard Van Noorden

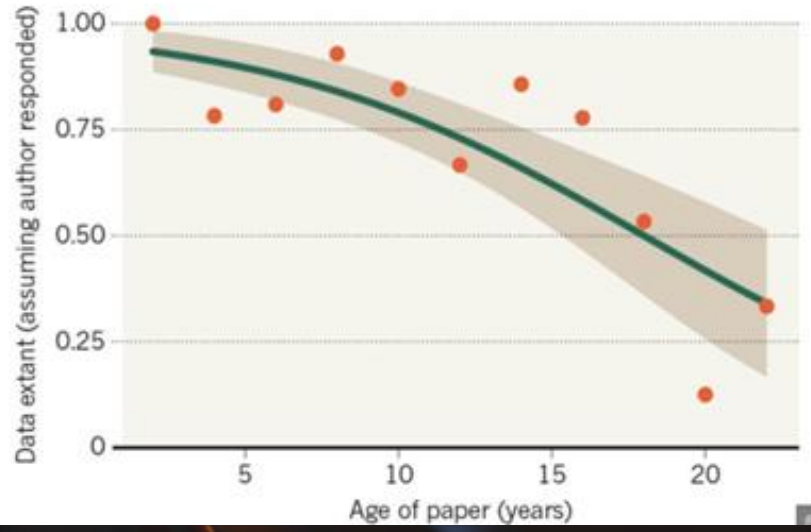
19 December 2013

Rights & Permissions

80% LOST IN 20 YEARS

MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.



7. BECAUSE DATA ARE FRAGILE

…THAT'S WHY YOU NEED A DATA MANAGEMENT PLAN.

IT'S NOT THE UMPTHENTH ADMINISTRATIVE BURDEN

Why should we care about data?

Nikolai Ivanovich Yezhov was head of the People's Commissariat for Internal Affairs until fell from Stalin's favor and power. Among art historians, he also has the nickname "The Vanishing Commissar" because after his execution, his likeness was retouched out of an official press photo; he is among the best-known examples of the Soviet press making someone who had fallen out of favor "disappear".

The Newseum (1 September 1999). " "The Commissar Vanishes" in The Vanishing Commissar". Archived from the original on 8 February 2007.

8. TO ENSURE INTEGRITY.
DATA CAN BE MANIPULATED.
…IT'S IN YOUR OWN INTEREST
TO AVOID IT

Lazzeri-Di Donato RDM, 2020

Why should we care about data?

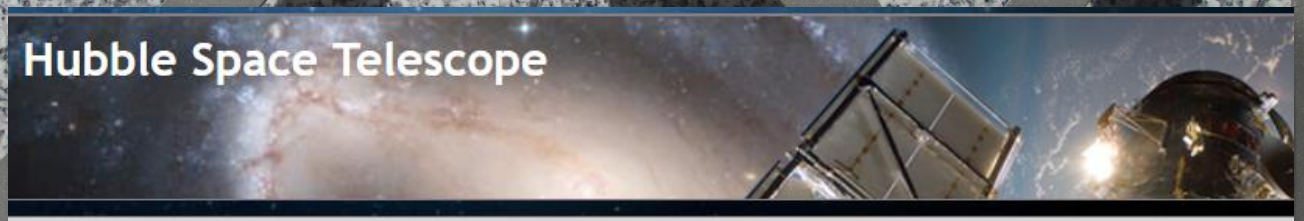
9. BECAUSE SOME DATA ARE UNIQUE AND NOT REPRODUCIBLE (EARTHQUAKES, WEATHER EVENTS)

Why should we care about data?

10. BECAUSE THEY CAN BE REUSED

… OFTEN IN UNPREDICTABLE WAYS

«THE COOLEST THING TO DO WITH YOUR DATA WILL BE THOUGHT OF BY SOMEONE ELSE» [R.POLLOCK]



News

Text Size

Astronomers Find Elusive Planets in Decade-Old Hubble Data **10.06.11**

In a painstaking re-analysis of Hubble Space Telescope images from 1998, astronomers have found visual evidence for two extrasolar planets that went undetected back then.

Finding these hidden gems in the Hubble archive gives astronomers an invaluable time machine for comparing much earlier planet orbital motion data to more recent observations. It also demonstrates a novel approach for planet hunting in archival Hubble data.

Exoplanet HR 8799 System

Why should we care about data?

11. AS ACCESS TO DATA BOOSTERS INNOVATION

OECD data

Why should we care about data?

12. TO AVOID
GARBAGE IN/GARBAGE OUT
FOR ARTIFICIAL INTELLIGENCE

CAREER COLUMN | 01 February 2024 2024

In the AI science boom, beware: your results are only as good as your data

Machine-learning systems are voracious data consumers – but trustworthy results require more vetting both before and after publication.

By Hunter Moseley

13. TO BE RESILIENT AGAINST CENSORSHIP

...re about data?

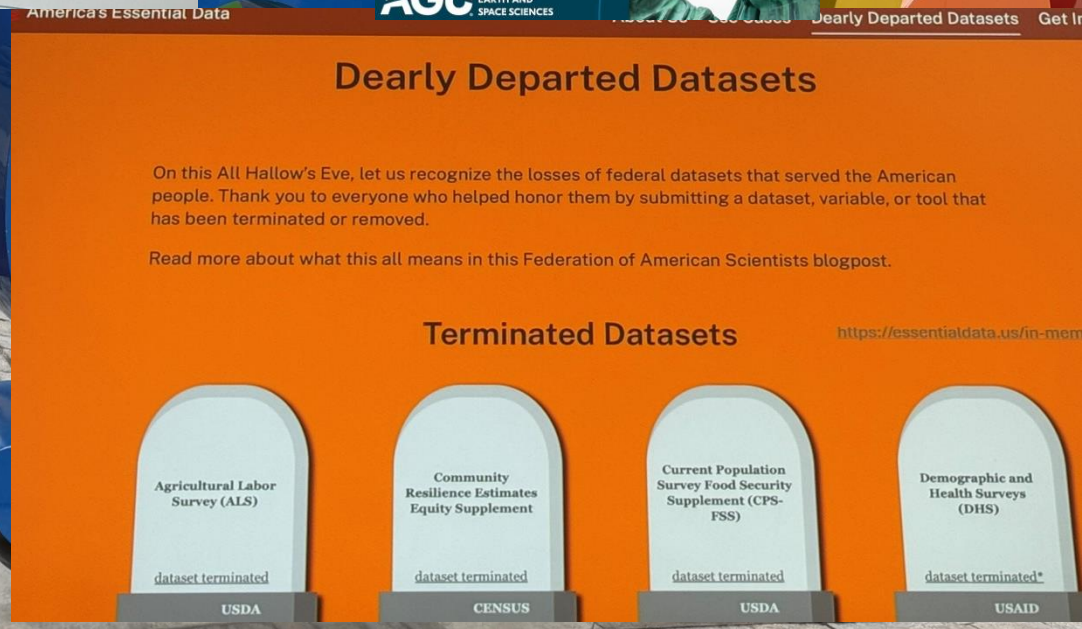
2025

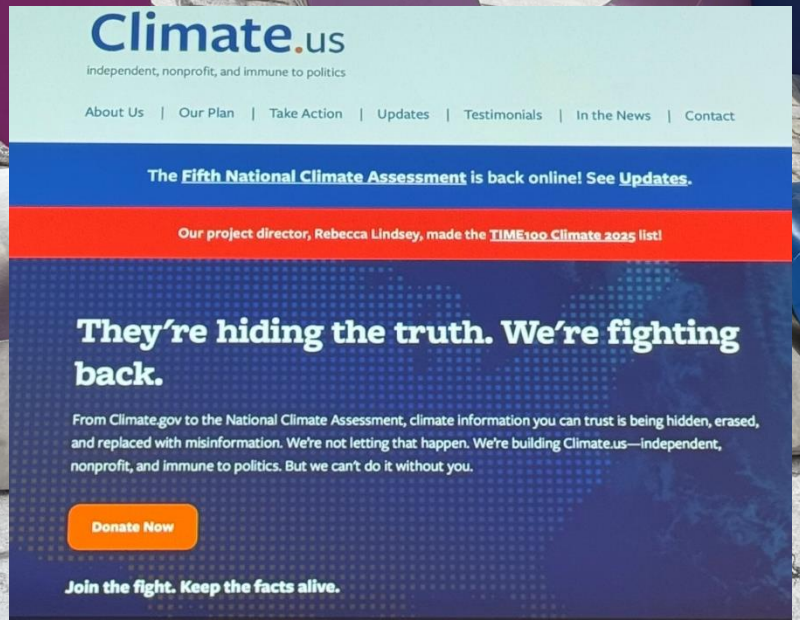
Starting in January 2025...a few examples

Removal of data elements not aligned with the administration's priorities.
Re: The terms "climate change" and "gender identity", and more.

Repurposing of Climate.gov (NOAA) website and removal of full-time federal and contractor staff supporting the platform.
U.S. premier platform for climate information for the public.

Removal of Globalchange.gov (USGCRP) website
Hosts five legislatively mandated reports on the state of the climate.





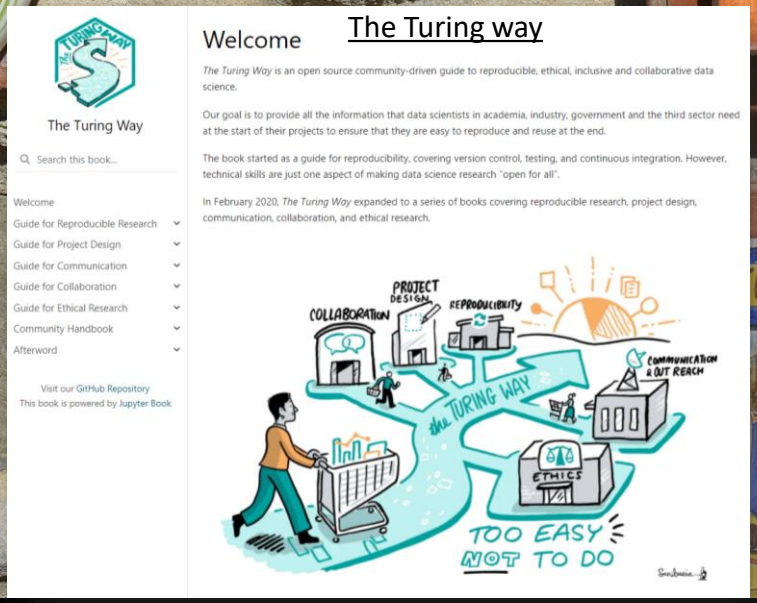
Why should we care about data?

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

14. DATA
CREATES BRIDGES…

Why should we care about data?



The Turing way

15. TO BE REPRODUCIBLE

16. FOR RESEARCH INTEGRITY

Why should we care ab

thebmjopinion Latest Authors Topic

Jan. 4, 2021

Peter Doshi: Pfizer and Moderna's "95% effective" vaccines—we need more details and the raw data

January 4, 2021

Five weeks ago, when I raised questions about the results of Pfizer's and Moderna's covid-19 vaccine trials, all that was in the public domain were the study protocols and a few press releases. Today, two journal publications and around 400 pages of summary data are available in the form of multiple reports presented by and to the FDA prior to the agency's emergency authorization of each company's mRNA vaccine. While some of the additional details are reassuring, some are not. Here I outline new concerns about the trustworthiness and meaningfulness of the reported efficacy results.

17. TO ALLOW FOR VALIDATION AND CHECKS (AND DETECT ERRORS/FRAUDS)

Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff

Thomas Herndon* Michael Ash Robert Pollin

April 15, 2013

Herndon, 2013

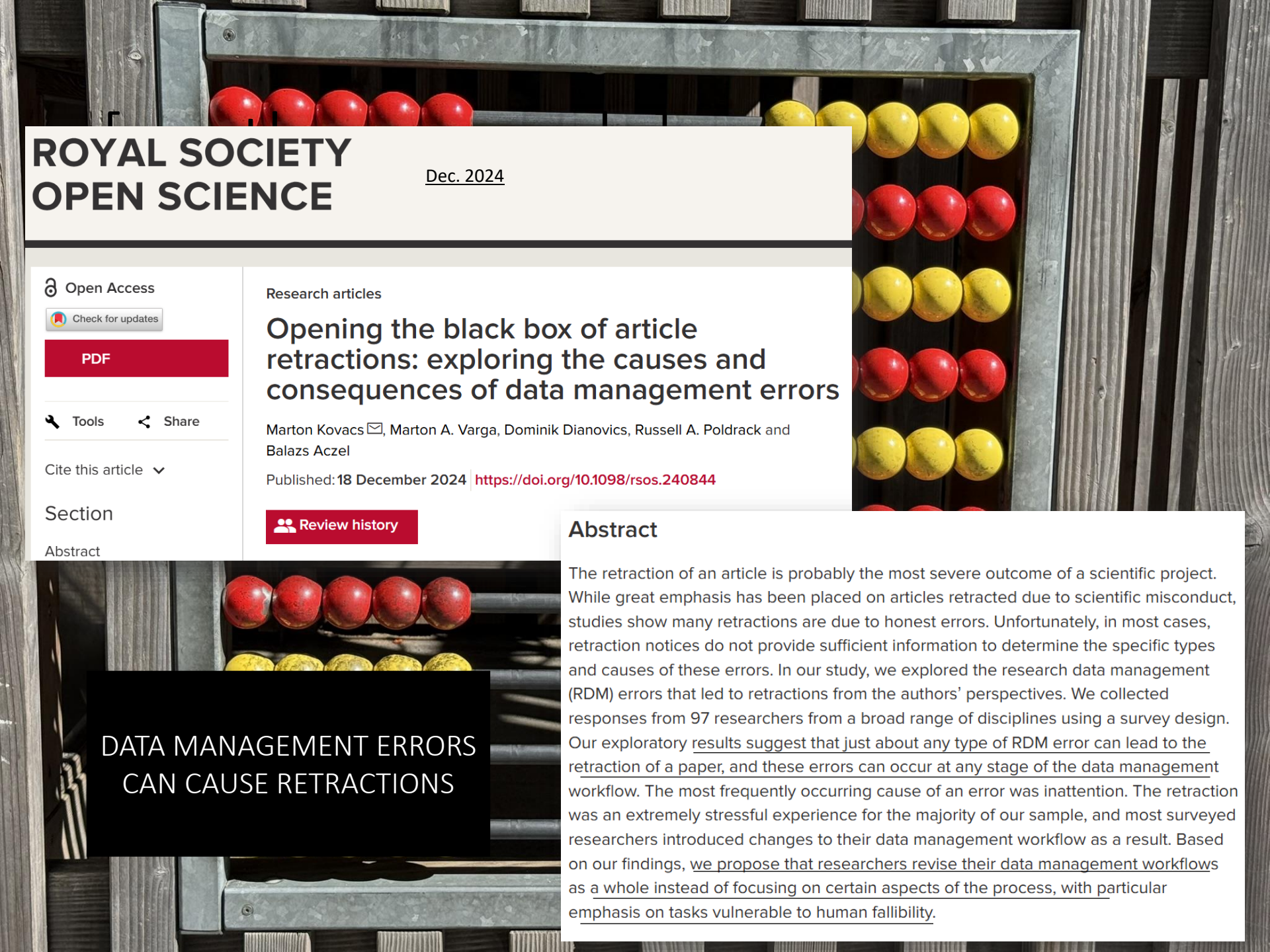
JEL CODES: E60, E62, E65

Abstract

We replicate Reinhart and Rogoff (2010a and 2010b) and find that coding errors, selective exclusion of available data, and unconventional weighting of summary statistics lead to serious errors that inaccurately represent the relationship between public debt and GDP growth among 20 advanced economies in the post-war period. Our finding is

debt loads greater than 90 percent of GDP consistently reduce GDP growth.

- SELECTIVE EXCLUSION OF DATA
- UNCONVENTIONAL SCHEMA OF WEIGHTING DATA
- A MISTAKE IN THE SPREADSHEET



ROYAL SOCIETY OPEN SCIENCE

Dec. 2024

Open Access

Check for updates

PDF

Tools Share

Cite this article

Section

Abstract

Research articles

Opening the black box of article retractions: exploring the causes and consequences of data management errors

Marton Kovacs, Marton A. Varga, Dominik Dianovics, Russell A. Poldrack and Balazs Aczel

Published: 18 December 2024 https://doi.org/10.1098/rsos.240844

Review history

DATA MANAGEMENT ERRORS CAN CAUSE RETRACTIONS

Abstract

The retraction of an article is probably the most severe outcome of a scientific project. While great emphasis has been placed on articles retracted due to scientific misconduct, studies show many retractions are due to honest errors. Unfortunately, in most cases, retraction notices do not provide sufficient information to determine the specific types and causes of these errors. In our study, we explored the research data management (RDM) errors that led to retractions from the authors' perspectives. We collected responses from 97 researchers from a broad range of disciplines using a survey design. Our exploratory results suggest that just about any type of RDM error can lead to the retraction of a paper, and these errors can occur at any stage of the data management workflow. The most frequently occurring cause of an error was inattention. The retraction was an extremely stressful experience for the majority of our sample, and most surveyed researchers introduced changes to their data management workflow as a result. Based on our findings, we propose that researchers revise their data management workflows as a whole instead of focusing on certain aspects of the process, with particular emphasis on tasks vulnerable to human fallibility.

Why should we care about data?

18. FOR A
SOUNDER SCIENCE

AN ARTICLE WITHOUT DATA
IS ONLY «ADVERTISING»
OF THE RESEARCH
YOU CAN'T VERIFY THE PROCESS

1995

WaveLab and Reproducible Research

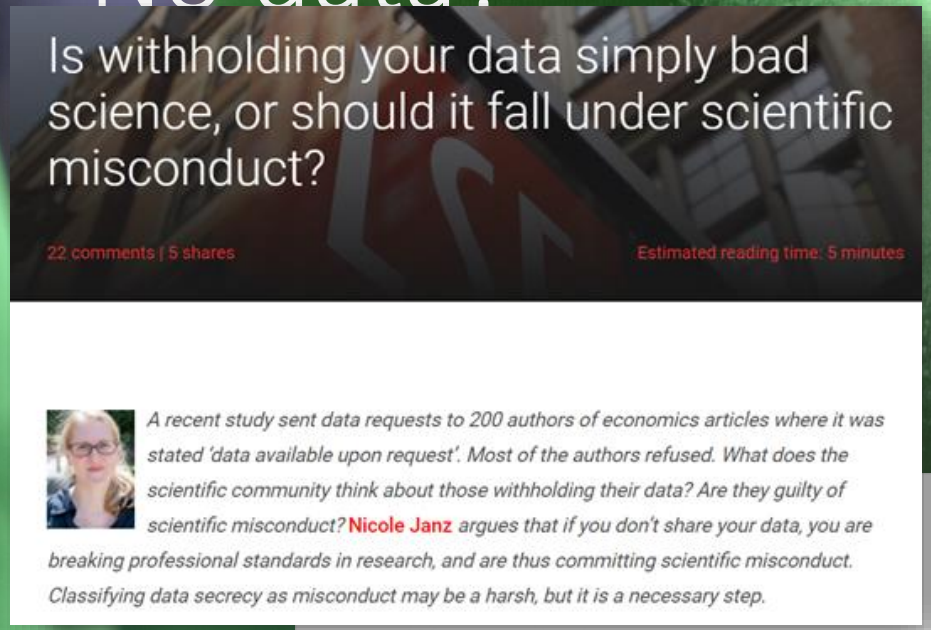
Jonathan B. Buckheit and David L. Donoho

Stanford University, Stanford CA 94305, USA

An article about computational science in a scientific publication is **not** *the scholarship itself, it is merely* **advertising** *of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.*

https://memegenerator.net/instance/64979477/case-closed-judge-judy-data-or-it-didnt-happen

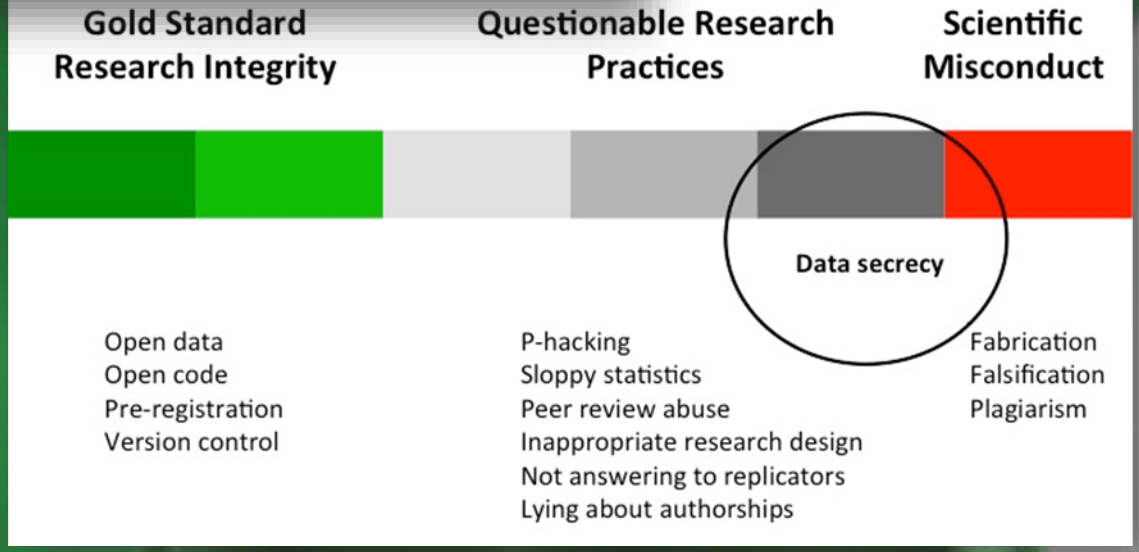
No data?



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

NO DATA?
LAZINESS OR FRAUD?

2015



Wilma van Wezenbeek
@wvanwezenbeek

Following

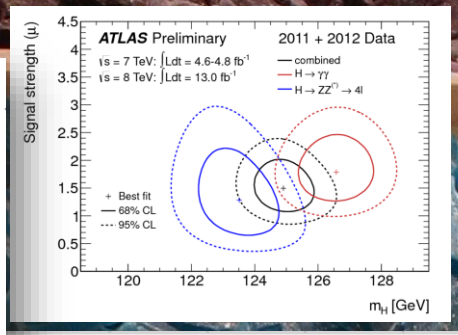
#osc2018 Wolfram Horstmann wants us to talk about datadiversity, like we do with biodiversity #openscience

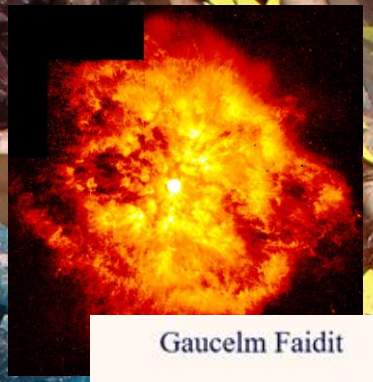
Traduci il Tweet

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.

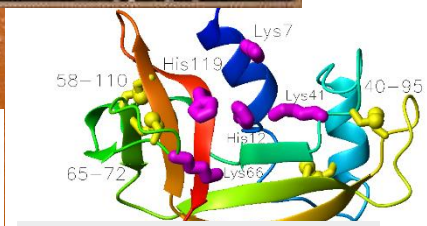
	A01_TP3	A03_TP1	A03_TP3	A04_TP1	A04_TP2	A04_TP3	A05_T
	1206	523	2131	25707	64473	60665	
	117	43035	206	119	1152	539	
					22858	1898	
					1457	29	
					19	85	
					2646	214	
					292	37	
					18	170	
					6	4	

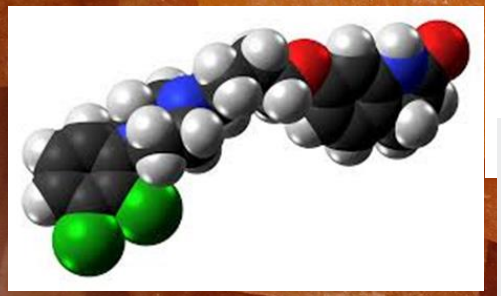
OTU_5	601	16
OTU_16	65	17
OTU_33	0	0

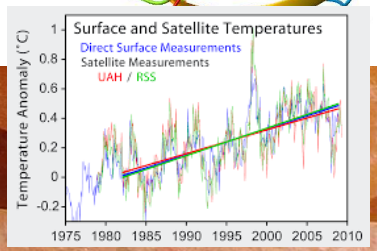
Gaucelm Faidit

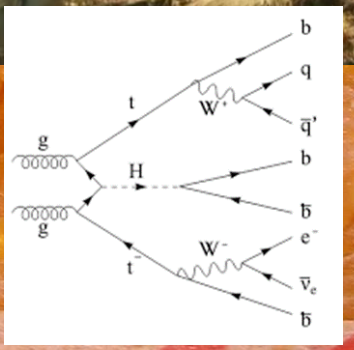
I.

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.

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

- **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

[we need data stewards]

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 500.000 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;

…data stewards

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.

DATA STEWARDS ARE NOT TO BE INTENDED AS «SUPPORTING STAFF». THEY ARE PART OF THE RESEARCH TEAM

DATA DOMAIN COMPETENCES+ TRANSVERSAL SKILLS

KØBENHAVNS UNIVERSITET

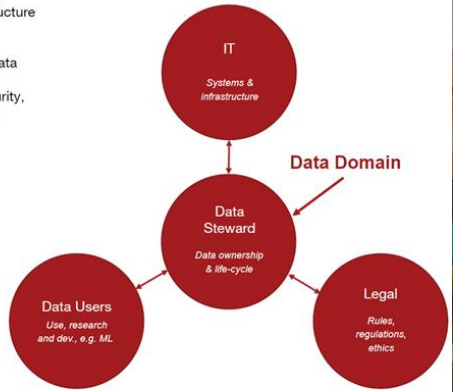
Competence Profile

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 data internally and externally in the organisation.

Competence profile examples

- Domain-specific data understanding
- Ability to ensure that structured and unstructured data and meta data is modelled, harvested, stored, and maintained in a 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.



Copenhagen Univ. June 17 2020

[data stewards]

- **Data Quality:** Ensuring that data is accurate, complete, and up-to-date. This involves regular audits, cleaning, and validation processes to maintain high data standards.
- **Data Access and Security:** Managing who can access data, ensuring that sensitive information is protected, and implementing appropriate security measures to prevent unauthorized access.
- **Data Documentation and Metadata Management:** Creating and maintaining documentation that describes data sources, structure, and usage, which helps other users understand the data and its context.
- **Compliance and Ethical Use:** Ensuring that data practices comply with relevant laws and regulations (e.g., GDPR, HIPAA) and promoting the ethical use of data throughout the organization.
- **Collaboration and Communication:** Working with various departments to ensure that data is managed consistently across the organization and that stakeholders are informed about best practices for data usage.

OPUS Why data stewards

[data stewards]

1. **Ensuring Data Integrity:** Data is only valuable if it is accurate and reliable. Data stewardship ensures that data remains of high quality, free from errors, and fit for its intended purpose. This is especially important in industries like healthcare, finance, and scientific research, where decisions based on faulty data can have serious consequences.
2. **Protecting Privacy and Security:** With the rise in data breaches and cyberattacks, safeguarding sensitive information is a top priority for organizations. Data stewardship helps protect data from unauthorized access, ensuring that privacy and security protocols are strictly followed. This is vital for maintaining customer trust and avoiding legal repercussions.
3. **Compliance with Regulations:** Many industries are subject to strict data regulations that govern how data must be collected, stored, and used. Data stewardship ensures that an organization remains compliant with these regulations, reducing the risk of fines, penalties, and reputational damage.
4. **Supporting Data-Driven Decision Making:** High-quality data enables organizations to make informed, data-driven decisions. Data stewardship ensures that decision-makers have access to the accurate and relevant data they need to guide strategy, optimize operations, and drive innovation.
5. **Fostering Ethical Data Use:** As organizations collect more data on individuals, ethical considerations become increasingly important. Data stewardship helps promote the responsible and ethical use of data, ensuring that it is used in ways that respect individual rights and do not cause harm.

OPUS Why data stewards

…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

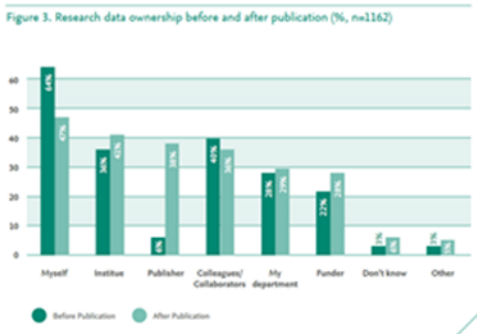
linkedin.com

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

[legal bo

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[1] can be protected by copyright if 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[2], 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

DATA Intelligence

< Previous Article Next Article >

Article Contents

- Abstract
- 1. THE PROTECTION OF DATA, DATA SETS AND DATABASES
- 2. SUITABLE OPTIONS FOR LICENSING DATA AND DATABASE RIGHTS

January 01 2020

Licensing FAIR Data for Reuse

Ignasi Labastida, Thomas Margoni

Author and Article Information

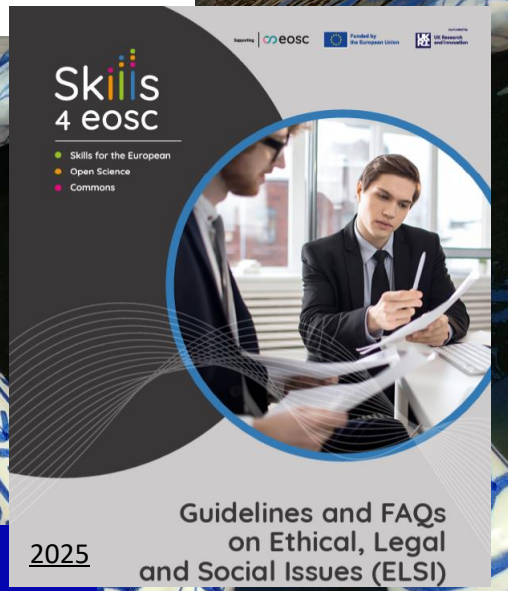
Data Intelligence (2020) 2 (1-2): 199–207.

https://doi.org/10.1162/dint_a_00042

Cite PDF Permissions Share

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



2025

- 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

The 3 steps

OPEN

FAIR

MANAGED

1. DATA MUST 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» [WITHIN EOSC, MANAGED/FAIR INCREASINGLY OVERLAPPING, «FAIR BY DESIGN»]

Article 3

General principle

1. Subject to paragraph 2 of this Article, Member States shall ensure that documents to which this Directive applies in accordance with Article 1 shall be re-usable for commercial or non-commercial purposes in accordance with Chapters III and IV.

Open data directive

(27) The volume of research data generated is growing exponentially and has potential for re-use beyond the scientific community. In order to be able to address mounting societal challenges efficiently and in a holistic manner, it has become crucial and urgent to be able to access, blend and re-use data from different sources, as well as across sectors and disciplines. Research data includes statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. It also includes meta-data, specifications and other digital objects. Research data is different from scientific articles reporting and commenting on findings resulting from their scientific research. For many years, the open availability and re-usability of scientific research data stemming from public funding has been subject to specific policy initiatives. Open access is understood as the practice of providing online access to research outputs free of charge for the end user and without restrictions on use and re-use beyond the possibility to require acknowledgement of authorship. Open access policies aim in particular to provide researchers and the public at large with access to research data as early as possible in the dissemination process and to facilitate its use and re-use. Open access helps enhance quality, reduce the need for unnecessary duplication of research, speed up scientific progress, combat scientific fraud, and it can overall favour economic growth and innovation. Beside open access, commendable efforts are being made to ensure that data management planning becomes a standard scientific practice and to support the dissemination of research data that are findable, accessible, interoperable and re-usable (the FAIR principle).

(28) For the reasons explained above, it is appropriate to set an obligation on Member States to adopt open access policies with respect to publicly funded research data and ensure that such policies are implemented by all research performing organisations and research funding organisations. Research performing organisations and research funding organisations could also be organised as

1. Data must be managed

DESCRIPTION FOR DISCOVERABILITY (metadata)

REPRODUCIBILITY

ORGANIZATION (file naming, folders, versioning…)

DOCUMENTATION

BACKUP AND STORAGE

LONG TIME PRESERVATION

LEGAL ASPECTS

ALONG THE ENTIRE LIFE CYCLE

2) Data must be FAIR

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 kn

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

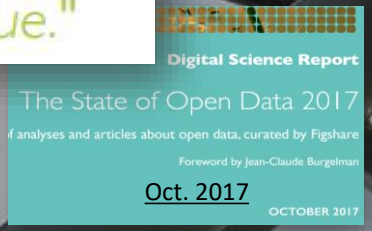
3. [WHENEVER POSSIBLE] Data should be Open

BECAUSE REUSE CREATES VALUE

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

«AS OPEN AS POSSIBLE»

REVOLUTION:
OPEN BY DEFAULT,
CLOSED ONLY UPON SOUND REASONS
(NOW: CLOSED BY DEFAULT, A FEW OPEN)



Oct. 2017

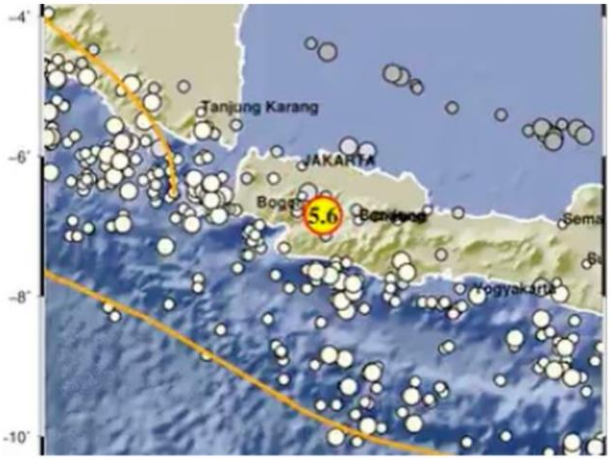
…the selfie…

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.

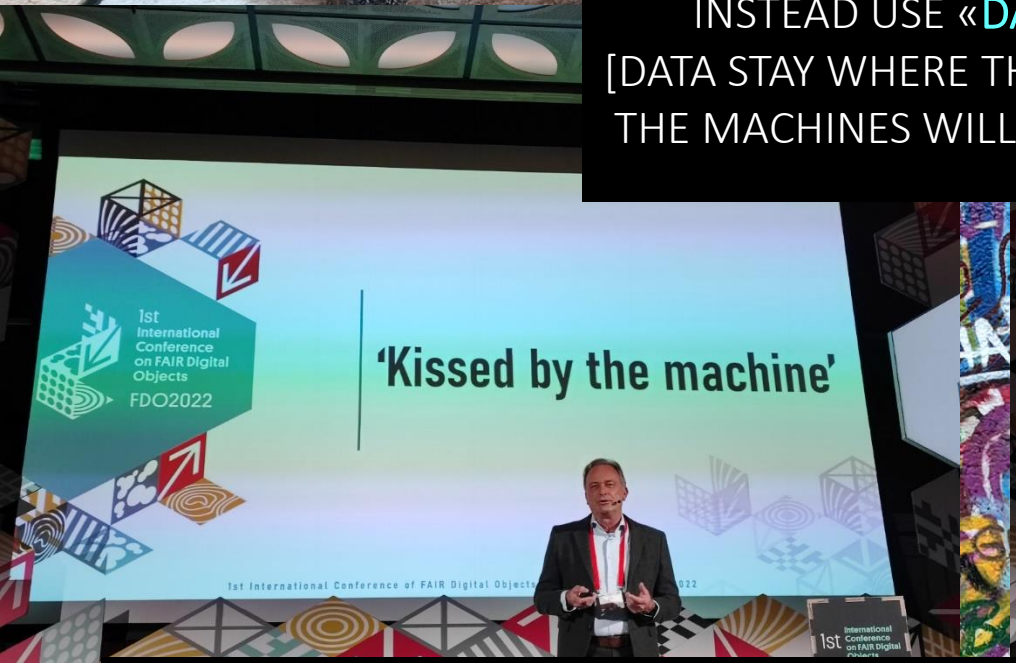


IN «FAIR» THE STRESS IS ON «R»

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

Kissed or missed?

STOP SAYING «DATA SHARING» -
INSTEAD USE «DATA VISITING»
[DATA STAY WHERE THEY ARE, IF «FAIR»
THE MACHINES WILL FIND/USE THEM]



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

[FAIR/AI]

FUNCTIONALLY AI READY

Functionally AI Ready: A Practice, Not an End State

- AI Readiness: a continuous practice rather than a final destination; an ongoing commitment rather than a checkbox
- "Fully" AI Ready is an aspiration
- What machine learning algorithms need may differ dramatically from what foundation models or LLMs require from the same dataset. The methods continue to evolve, and the applications multiply
- Sustained investment is required in
  - people
  - infrastructure
  - collaborative practices

Redefining FAIR

In the spirit of continuous improvement, FAIR in the AI context should mean **Functionally AI Ready** rather than "Fully AI Ready", acknowledging the dynamic nature of both AI technology and data readiness

Christine Kirkpatrick, OAI14

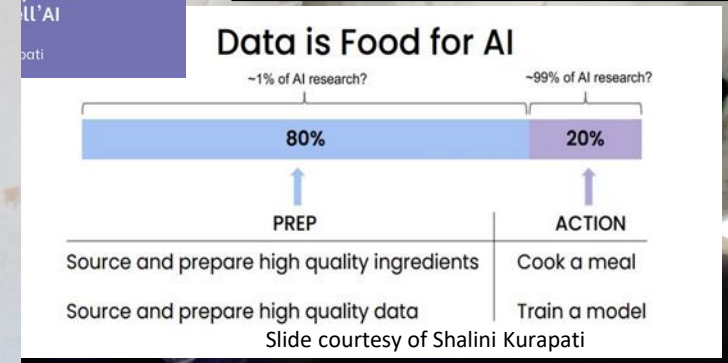
...not only for humans

At the end of the day, Artificial Intelligence is a powerful combination of *data* and *algorithms*. These AI algorithms are data-hungry. They require massive amounts of data to train themselves to do their intended job. And if they get bad data, the results are poor, too. *Garbage in, garbage out.*

GARBAGE IN, GARBAGE OUT: THAT'S WHY WE NEED MACHINE-ACTIONABLE FAIR DATA!

clearbox AI

presente e
ll'AI
pati



Slide courtesy of Shalini Kurapati

sssWhat if all our stereotypes are systematically programmed into the AI we are developing and using, and somehow, we attribute rationality to them? You guessed right. DALL-E also thinks a CEO can only be a man and a nurse only a woman. And not just in images but even in textual outputs.

BEWARE OF BIASES

FAIR/Open

AS OPEN AS POSSIBILE
AS EARLY AS POSSIBLE
AS FAIR AS POSSIBLE

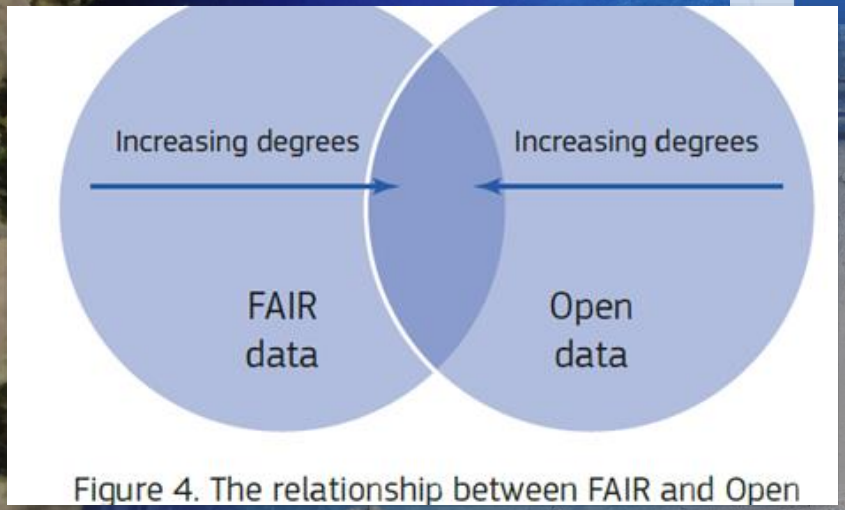
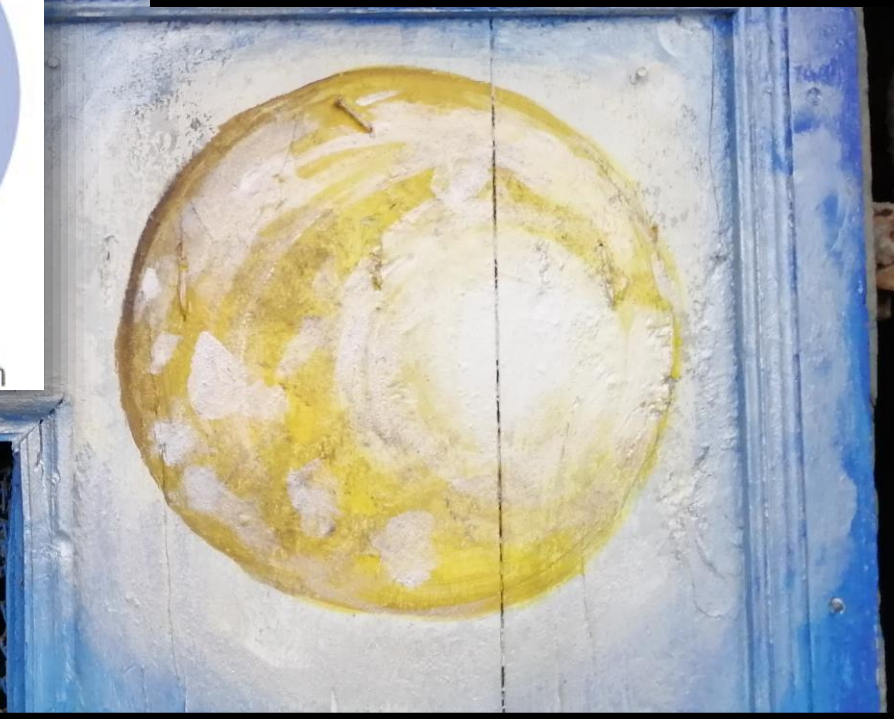
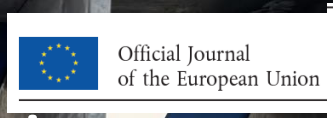


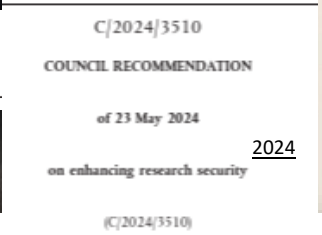
Figure 4. The relationship between FAIR and Open



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

Open Science and security




C/2024/3510

COUNCIL RECOMMENDATION

of 23 May 2024

on enhancing research security

(C/2024/3510)

2024

(1) Openness, international cooperation, and academic freedom are at the core of world-class research and innovation. Yet, with growing international tensions and the increasing geopolitical relevance of research and innovation, the Union's researchers and academics are increasingly exposed to risks to research security when cooperating internationally, resulting in European research and innovation being confronted with malign influence and being

(3) Open science ensures that scientific research is made as accessible as possible for the benefit of science, the economy and society as a whole. International cooperation in research and innovation is vital for finding solutions to pressing global challenges for the benefit of our societies and drives scientific excellence, while international mobility of research talent enriches scientific enquiry and is essential for fostering innovation and achieving scientific breakthroughs. Academic freedom implies that researchers are free to conduct their research and choose the research methods as well as their research partners from around the globe, taking into account that with academic freedom comes academic responsibility.

(b) continue to promote and encourage international cooperation in research and innovation that is both open and secure, in line with the principle 'as open as possible, as closed as necessary', ensuring that research outputs are findable, accessible, interoperable and reusable (FAIR), with due consideration to applicable restrictions, including security concerns;

«AS OPEN AS POSSIBLE» IS THE SOLUTION

(c) ensure proportionality of measures: where safeguards are introduced, these should not go beyond what is necessary to mitigate the risks at stake and avoid any unnecessary administrative burden. The objective is to manage rather than to avoid risk;

«MANAGING» RISK (NOT «AVOIDING»)

(d) steer research security measures to safeguard economic security, as well as Union and national security, and defending and promoting Union values and fundamental rights, academic freedom and research integrity, while avoiding protectionism and political instrumentalisation of research and innovation;

[Open/security are not in conflict]

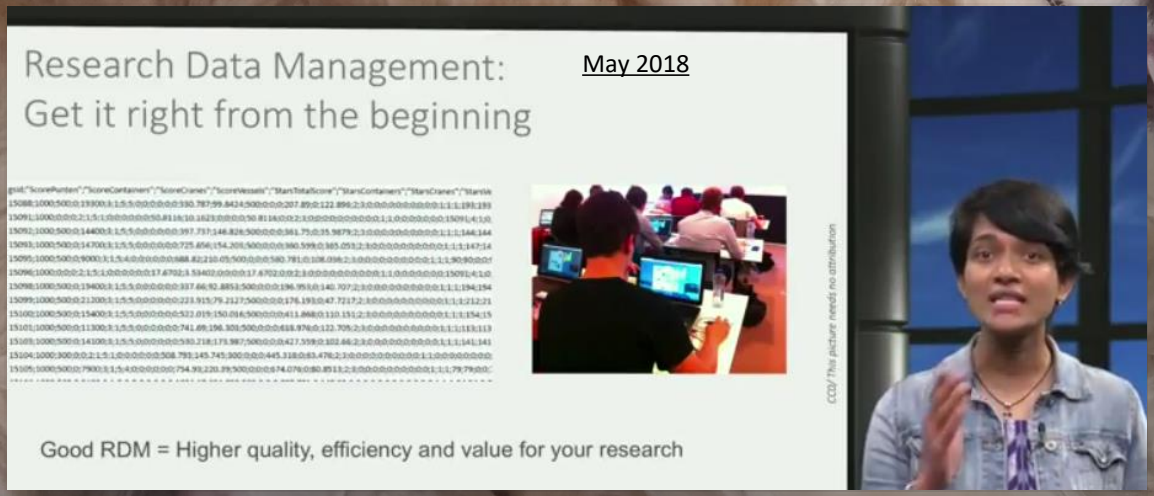
FAIR DATA INCLUDE SECURITY ANALYSIS/MEASURES (AND GOOD DATA STEWARDSHIP)

«OPEN» AND «SECURITY» ARE NOT IN CONFLICT
THEY REINFORCE EACH OTHER

LET’S GO WORKING

STEP #1
DATA
MANAGEMENT
proximus
proximus

… geti it right from the beginning



GET IT RIGHT FROM THE BEGINNING

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

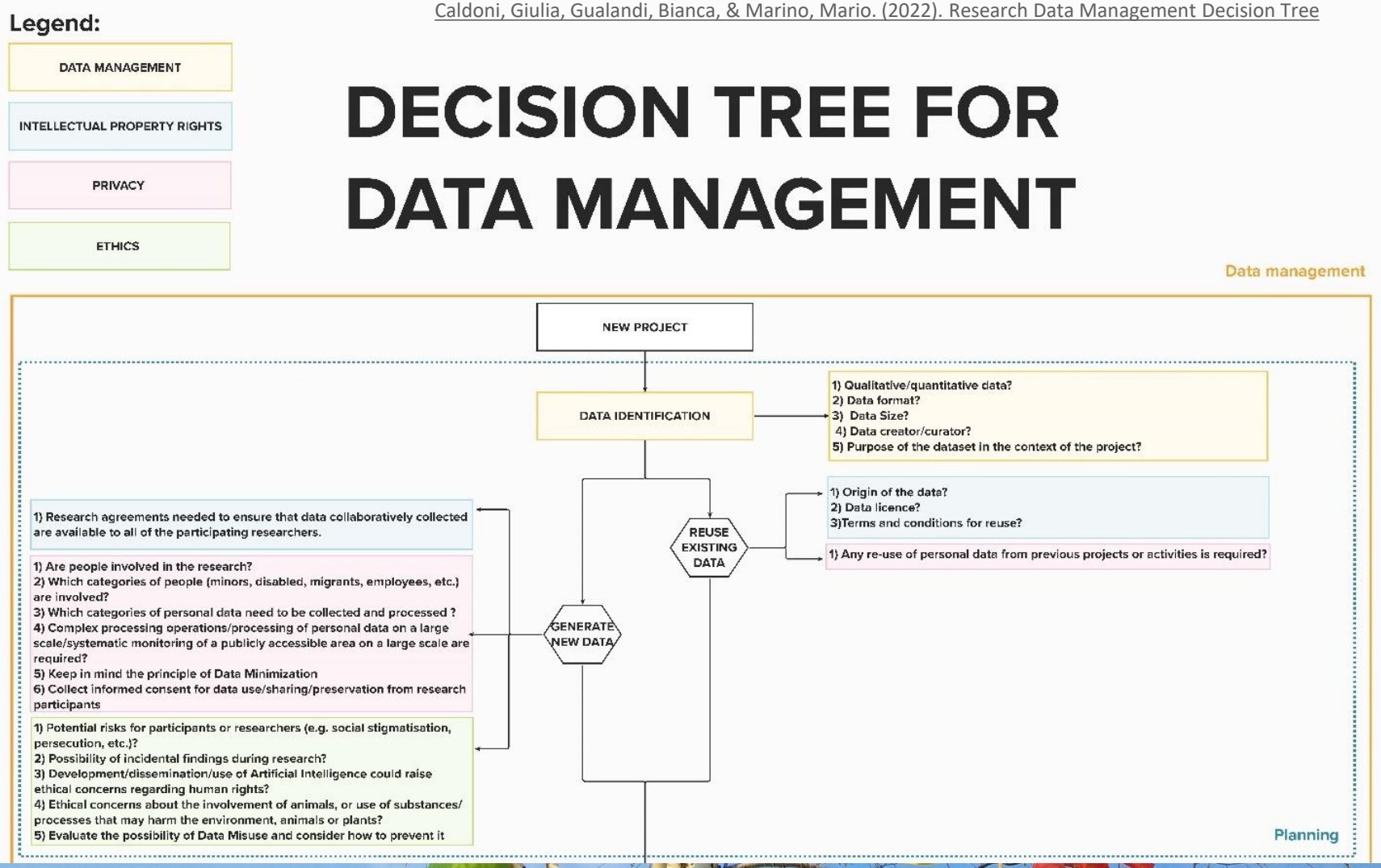
In addition to the above mentioned actions, there are also a few things to consider to make data management a standard practice in daily operations.

ASK THE RIGHT QUESTIONS

Data decision tree

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 a … open/close at the … security ongoing]

Tag Type	Description	Security Features	Access Credentials
Blue 2015	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

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
	ormation 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 HUIDs associated with names or any other information that could identify individuals Harvard personnel records (employees may discuss terms and conditions of employment with each other and third parties) Level 4 including non-directory student information and directory information about students who have requested a FERPA block HUIDs associated with names or any other information that could identify individuals Harvard personnel records (employees may discuss terms and conditions of employment with each other and third parties) Institutional financial records Individual donor information • Other personal information protected under state, federal and foreign privacy laws not classified as Level 4 or 5
	e of which would not cause material harm, but which the University has chosen to ork and intellectual property not in Level 3 or 4 assified as Level 2 by an IRB work papers, drafts of research papers mation about the University physical plant
	been de-identified in accordance with applicable rules bout the University bout students who have not requested a FERPA block ry information

…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?

Harvard security

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

Form

DISCIPLINE

TASKS

Search Type here...

https://rdmkit.elixir-europe.org/your_domain

Your domain
Bioimaging data
Data management solutions for bioimaging data.

Your domain
Biomolecular simulation data
Data management solutions for biomolecular simulation data.

Your domain
Epitranscriptome data
Data management solutions for epitranscriptome data.

Your domain
Health data
Data management solutions for human health data.

Your domain
Human data
Data management solutions for human data.

Your domain
Human pathogen genomics
Data management solutions for Human pathogen genomics

Your domain
Intrinsically disordered proteins
Data management solutions for

Your domain
Machine learning
This page is about setting best practices for data towards enabling

Your domain
Marine metagenomics
Data management solutions for marine metagenomics data.

martedì 4 giugn

Your domain
Microbial biotechnology
Data management solutions for microbial biotechnology data.

Your domain
Plant sciences
Data management solutions for plant sciences data.

Your domain
Proteomics
Data management solutions for proteomics data.

Your domain
Rare disease data
Data management solutions for rare disease data.

Your domain
Single-cell sequencing
Managing data generated from single-cell sequencing experimen

Your domain
Toxicology data
Data management solutions for toxicology data.

Your tasks
Compliance monitoring & measurement
How to measure compliance to data management regulations and standards.

Your tasks
Costs of data management
Budgeting and costing for data management

Your tasks
Data analysis
How to make data analysis FAIR.

Your tasks
Data brokering
Information on brokering data to data repositories on behalf of data producers.

Your tasks
Data management plan
How to write a Data Management Plan (DMP).

Your tasks
Data organisation
Best practices to name and organi
research data.

Your tasks
Data provenance
How to record information about data provenance.

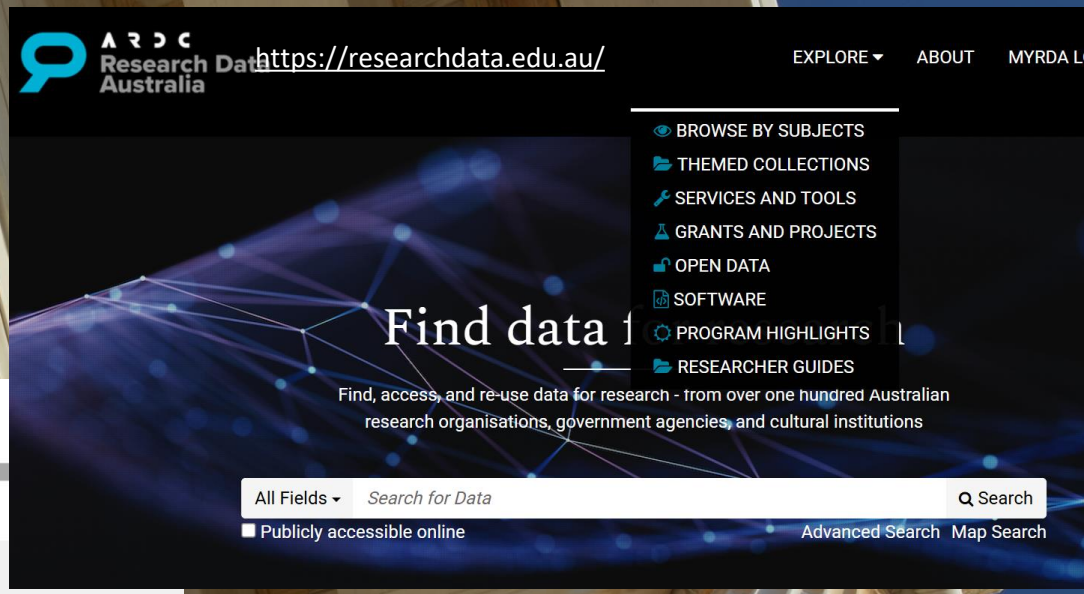
Your tasks
Data publication
How to prepare data and find repositories for publication.

Your tasks
Data quality
How to ensure high quality of research data.

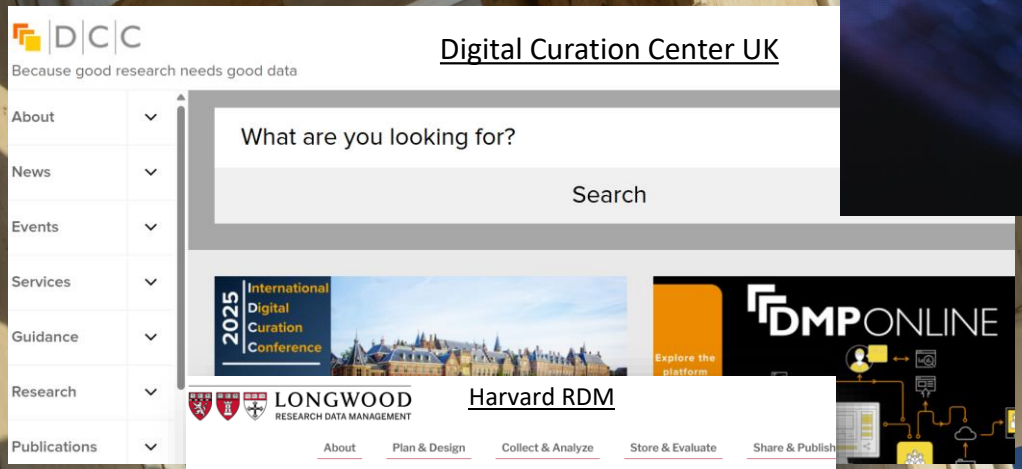
[DMP]

4 pillars

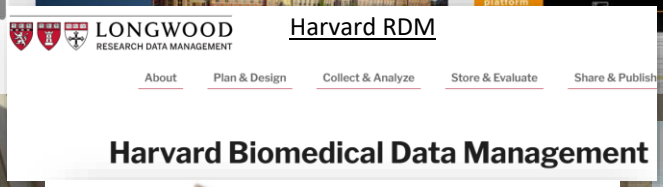
https://researchdata.edu.au/



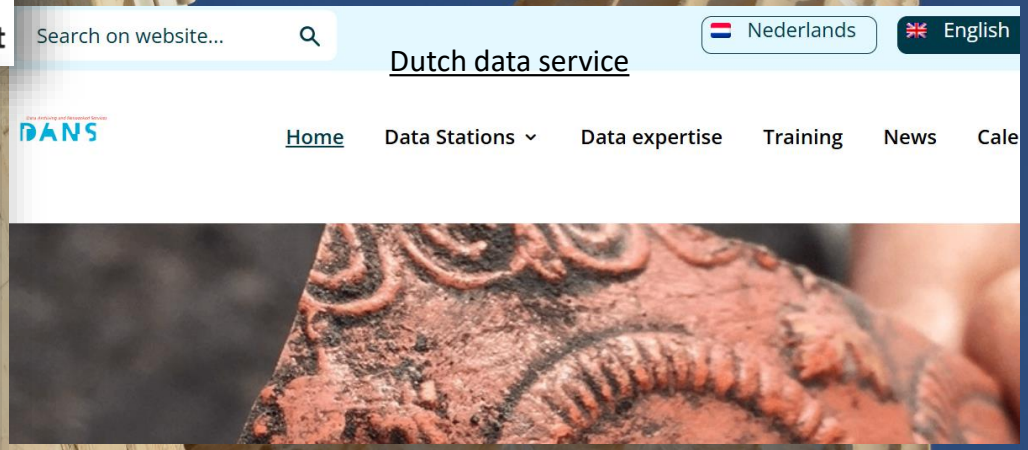
Digital Curation Center UK



Harvard RDM



Dutch data service



D

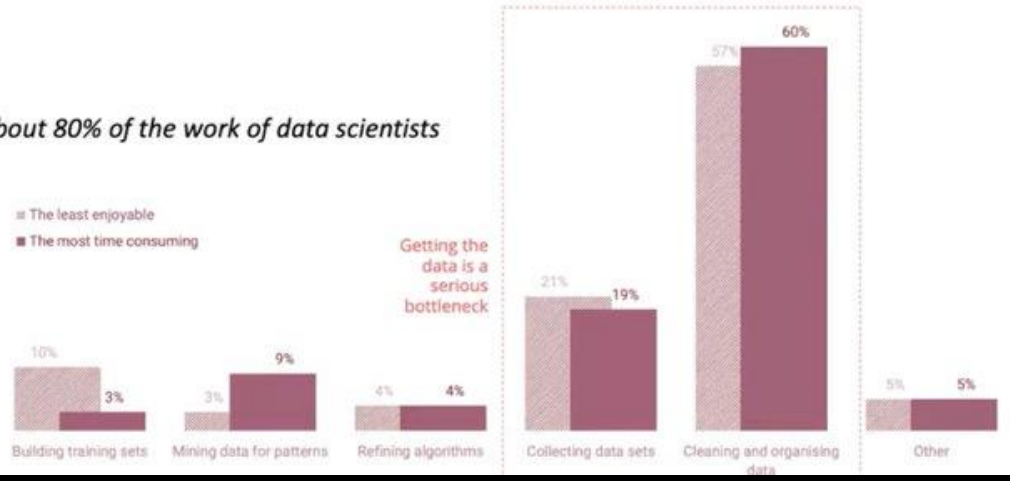
The problem

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

D ent ABC - Costing

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.		

Data management A

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

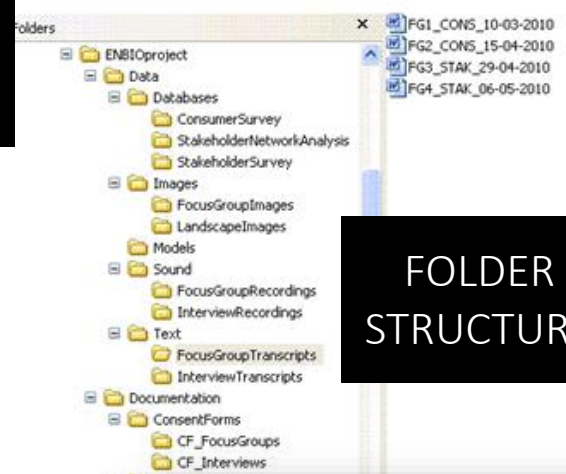
A readme file provides information about a data file and is intended to help ensure that the data can be correctly interpreted, by yourself at a later date or by others when sharing or publishing data. Standards-based metadata is generally preferable, but where no appropriate standard exists, for internal use, writing "readme" style metadata is an appropriate strategy.

Want a template? **Download one** and adapt it for your own data!

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

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



FOLDER STRUCTURE

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

File naming conventions

File naming

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 th
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: 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 c 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 correspon whether it is incoming or outgoing correspondence, except where the inclusion of any of these elements would be incompatible with rule
11. The file name of an email attachment should include the name of the correspondent, an indication of the subject, the date of the corresp 'attch', and an indication of the number of attachments sent with the covering email, except where the inclusion of any of these elements 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.

Example version control table:

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:		25/1	
Based on:		Visi	
Version	Responsible	Note	
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

VERSIONING

git Git

local-branching-on-the-cheap

Git is a free and open source distributed version control sy
designed to handle everything from small to very large pro
speed and efficiency.

D

OpenRefine Download Documentation Community Blog

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

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 knowledge base anyone can edit, or other Wikibase instances.

CESSDA Guide

cessda TRAINING

Data Management Expert Guide

1. Plan
2. Organise & Document
3. Process
   - Data entry and integrity
   - Quantitative coding
   - Qualitative coding
   - Weights of survey data
   - File formats and data conversion
   - Data authenticity
   - Wrap up: Data quality
   - Adapt your DMP: part 3
   - Sources and further reading
4. Store
5. Protect
6. Archive & Publish

- Check the completeness of records
- Reduce burden at manual data entry
- Minimise the number of steps
- Conduct data entry twice
- Perform in-depth checks for selected records
- Perform logical and consistency checks
- Automate checks whenever possible

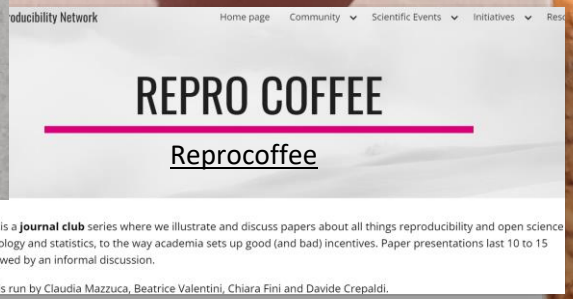
Reproducible coding

V. Pasquale, 2020

DO's

- Comment your code.
- Use versioning systems (such as Git, Mercurial, Subversion, etc.) to keep track of versions, collaborate, and perform automatic backups.
- Use code notebooks (e.g. Jupyter notebooks) for literate programming.
- Consider using emulation and virtualisation tools, such as Docker, to help recreating the original computing environment for reproducibility.
- Use workflow tools (Snakemake, Nextflow, Galaxy, DataJoint) to make your computational methods portable, maintainable, reproducible and shareable, especially if analytical pipelines can involve up to hundreds of successive steps (e.g. bioinformatics).
- Use software licenses (e.g., GNU GPL license)

DONT's



PAST

15 May 2024, Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. Review of General Psychology, 13(2), 90–100. https://doi.org/10.1037/a0015108

Hedges, L. V., & Schauer, J. M. (2019). Statistical analyses … erspectives. Psychological Methods, 24(5), 557–570. https://doi.org/10.1037/met000

ITRN REPRO COFFEE

28 November 2023, Better Accuracy for Better Science … -Storber et al., 2023.

24 October 2023, Reproducible brain-wide association st… by Marek et al., 2022

27 June 2023, Single case studies are a powerful tool for developing, testing and extending theories. by Nickels, L., Fischer-Baum, S. & Best, W. 2022.

28 March 2023, Increasing Transparency Through a Multiverse Analysis by Steegen et al., 2016 & Ethics and Statistics: Honesty and Transparency are not Enough by Gelman, 2017.

Data management ABC / documentation

Organising data

Electronic lab notebooks

MANTRA EDINA

Other tools, such as electronic lab notebooks (ELNs) can support the organisation of research data.

ELNs perform the same function as paper lab notebooks, and also enable better sharing and searching for data.

Some electronic lab notebooks also support integration with other tools used in the lab, and with research infrastructure.

Click on the headings to see the advantages of using an electronic lab notebook.

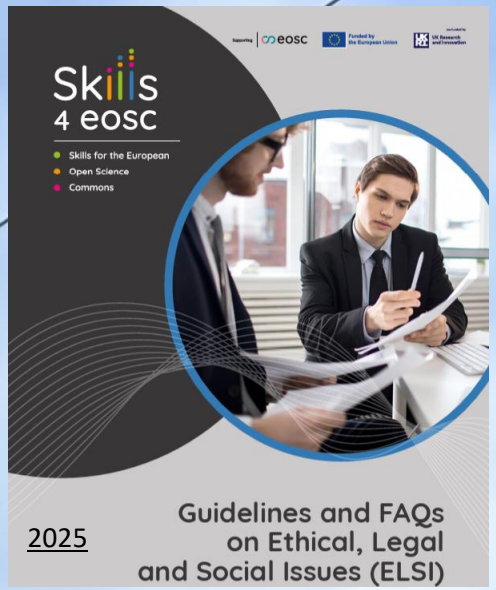
E-LAB NOTEBOOK TO DOCUMENT YOUR RESEARCH STEPS

Organising your data | Sharing data with others | Integration with other tools

- Data and documents have a unique ID assigned to them at time of creation which stays with them regardless of changes to the filename, preventing mix-ups with similarly named files.
- All files have appropriate metadata including creator, creation date, last modification, file type etc assigned to them at time of creation.
- Files may be tagged according to their content, grouping them together and facilitating their rapid identification and retrieval during searches.
- Powerful searching capabilities allow files to be quickly identified via filters, tags, unique IDs, modification dates, creator etc.
- Files may be organised into notebooks and folders making them easy to locate and browse.
- Files and datasets may be linked within documents allowing results and methodologies for different experiments to be kept together.
- Automatic recording of document versions helps ensure that research is recorded in a way that supports protection of intellectual property.

RSpace COMMUNITY | SciNote | <LAB FOLDER> | labguru

Data manageme



COMPLETE GUIDE ON LEGAL ASPECTS

Table of Contents

Skills 4 eosc

1. Introduction › 5

1.1 Background: Challenges facing Policy Makers and Civil Servants in the connected age › 5

1.2 Links to relevant Minimum Viable Skillset career profiles › 7

1.2.1 Interlinks with Policy Makers MVS (Research Policy, Evidence-based Policy Maker Roles) › 7

1.2.2 Interlinks with Civil Servants MVS › 8

1.3 Scope of the document › 9

2. Personal Data Protection › 12

2.1 Fact sheets › 12

2.2 FAQ › 19

2.3 Useful terms › 23

3. Intellectual Property Rights and other non-personal data › 26

3.1 Fact sheets › 26

3.2 FAQ › 32

3.2.1 Intellectual Property (general) › 32

3.2.2 Copyright › 34

3.2.2.1 Copyright and Open Science › 41

3.2.2.2 Licensing › 45

3.2.3 Sui Generis Database Rights (SGDR) › 49

3.2.4 Trade Secrets › 52

3.2.5 Non-personal data and public sector bodies › 56

3.2.5.1 Open Data Directive › 56

3.2.5.2 Digital Services Act › 59

3.2.5.3 Data Act › 61

3.2.5.4 Data Governance Act › 63

3.2.5.5 AI Act › 64

3.2.5.6 Data under EU-funded projects › 67

4. Ethics in research and open science › 69

4.1 FAIR Principles › 70

4.2 Responsible Research and Innovation (RRI) › 72

4.3 European Code of Conduct for Research Integrity › 73

4.4 Responsible Open Science for Europe (ROSiE) › 75

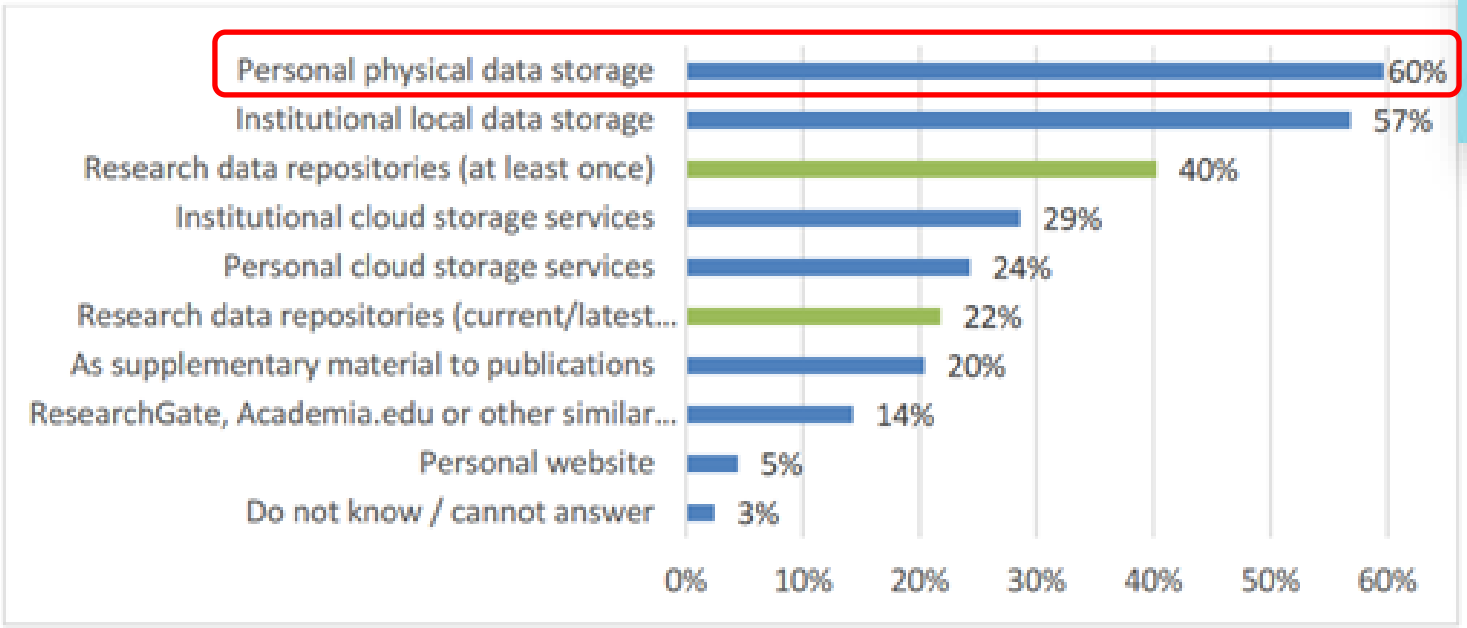
Data management ABC / storage

WHERE DO YOU STORE YOUR DATA?

2022

European Research Data Landscape

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- backup and storage

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

[persona

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

'Processing shall be lawful only if and to the extent that at least one of the following applies' (art 6(1)):

Legal Bases

Consent

- 'the data subject has given consent to the processing of his or her personal data for one or more specific purposes'

Performance of a contract

- 'processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract'

Compliance with a legal obligation

- 'processing is necessary for compliance with a legal obligation to which the controller is subject'

Protection of vital interests

- 'processing is necessary in order to protect the vital interests of the data subject or of another natural person'

Public interest

- 'processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested

or the purposes of the legitimate interests pursued by the controller or by a third party, except where such
y the interests or fundamental rights and freedoms of the data subject which require protection of
r where the data subject is a child.'

Author: Luca Schirru

2.2 FAQ

What is the GDPR?

The General Data Protection Regulation is a personal data protection law that safeguards the privacy of European residents with respect to how their digital data is collected, stored, processed, shared, and used. In its own wording, it "lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data.)" (Art. 1(1) GDPR).

What does "processing" encompass?

"Processing", under the GDPR, encompasses "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;" (Art. 4(2) GDPR).

What is "personal data"? (for GDPR purposes)

Personal data is understood as "any information relating to an identified or identifiable natural person ('data subject')" (Art. 4(1) GDPR).

What does "identifiable" means?

In order to be considered "identifiable" for the purposes of the regulation, a natural person may be "identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical,

2023

Supporting eosc

Skills 4 eosc

Deliverable 3.6.

Coordinated set of guides, fact-sheets and FAQs on ELSI aspects for Civil Servants and Policy Makers

[pers

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

cessda TRAINING

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,

[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

Amnesia OpenAIRE

amnesia

High accuracy Data Anonymization.

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

GET STARTED

Anonymising **quantitative data** may involve r reducing the precision or detailed textual mea

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

... Thanks!