Research Data Management: A short course for PoliTo Researchers

Sala Maxwell, DET

18 Feb 2020

About us

Dr. ir. Shalini Kurapati is Open Science fellow at PoliTo (Adjunct) with Prof. Federica Cappelluti, OS advisor to rector to:

- **Provide awareness and training on Research Data Management**
- **Design of a policy roadmap (e.g. IPR and RDM)**
- **Advice (as much as we can) on all data related matters of Open Science**
- **Fully researcher oriented**

Together with us today:

Important RDM stakeholder reps from PoliTo

Mr. Enrico Venuto, Area IT

Ms. Nicoletta Roz, DPO
Affari Legali

Shiva Loccisano, head of TRIN

And you?

Go to

www.menti.com

Code:

145285

Today's programme

Part 1:

- Introducing the rationale of Open Science and the role of RDM: Why should you care
- What should I do and where do I start: Intro to Data Management Plans
- Storage and computing infrastructure at PoliTo by Mr. Enrico Venuto, Area IT

Break

Part 2:

- Introduction to GDPR and PoliTo processes, by Nicoletta Roz, DPO
- Managing and sharing personal data during a research project
- IPR and Tech transfer considerations, by Shiva Loccisano, head of TRIN
- Wrap up and intro to the Feb 20 session

This workshop is for you- Make it yours!

- Ask any questions you might have, no question is trivial
- Add your suggestions / solutions - we learn from each other
- We co-create this workshop

Introduction to the rationale of Open Science and the role of RDM

Definition of open science

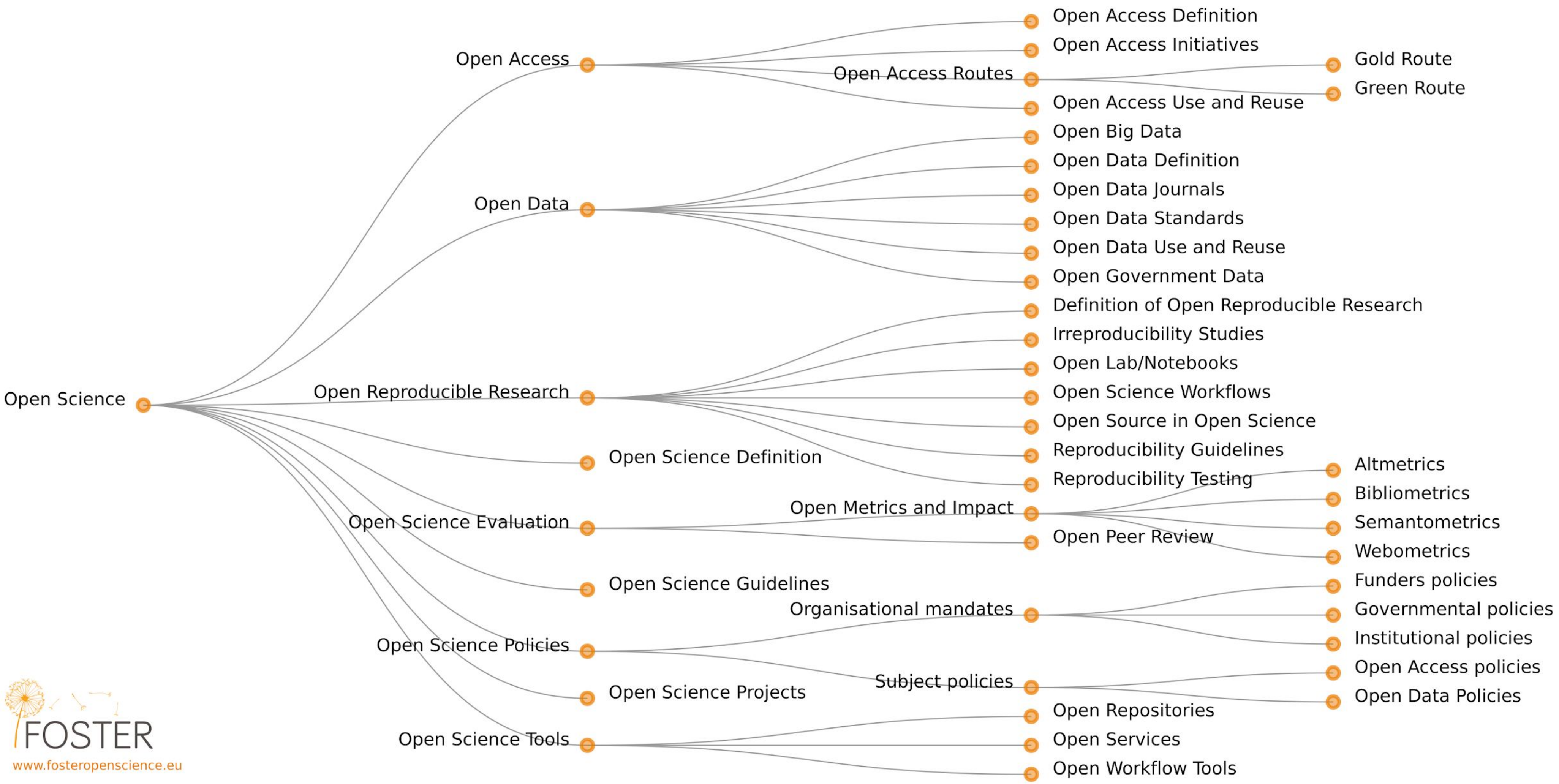
Definition of open science

There is no single doctrine or paper that definitively captures open science. Rather, open science can be defined as a **set of practices** that increase the **transparency** and **accessibility of scientific research** (van der Zee & Reich, 2018).

https://www.apa.org/science/about/psa/2019/02/open-science

What is open science

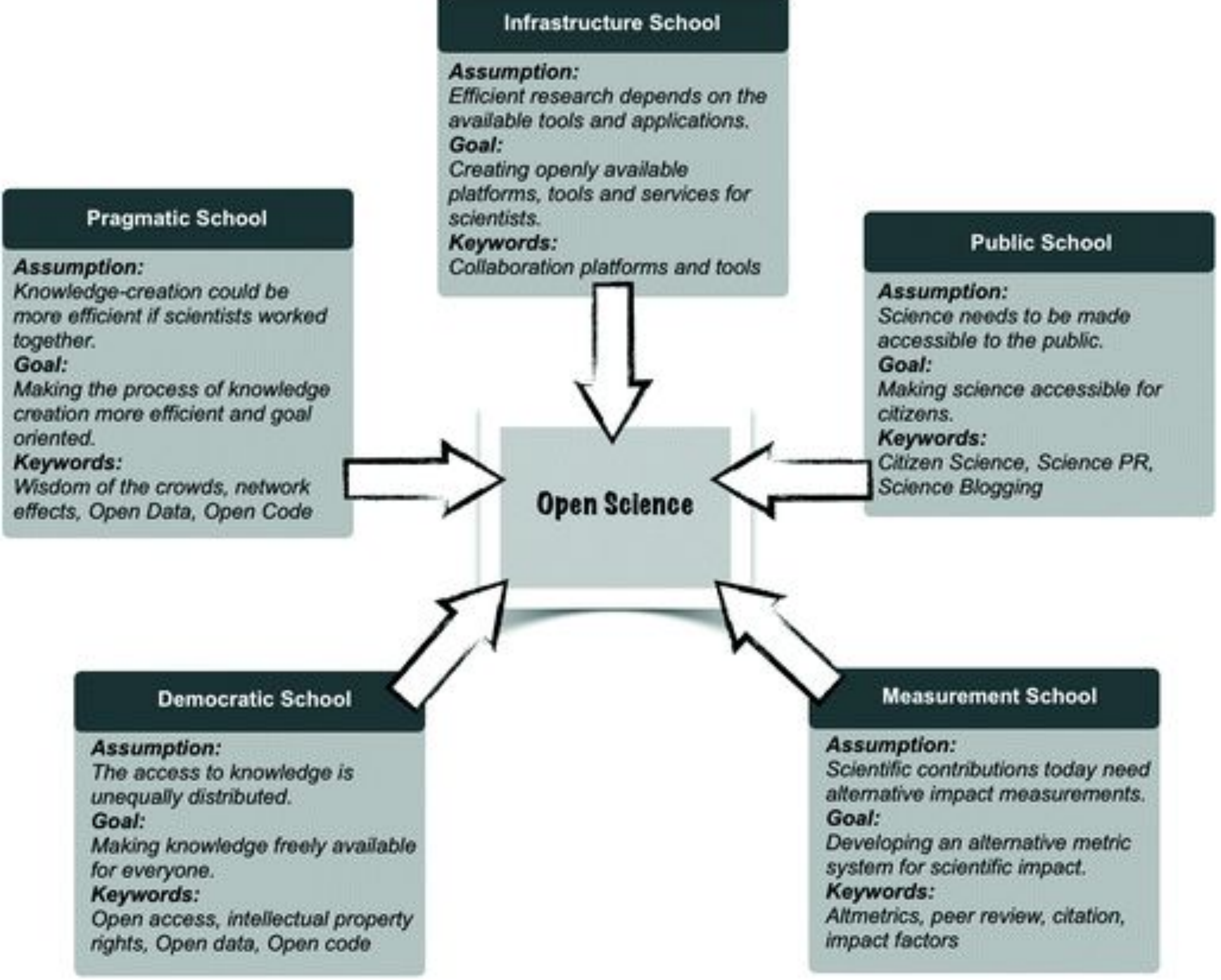
Open Science Taxonomy



- Open Science
  - Open Access
    - Open Access Definition
    - Open Access Initiatives
    - Open Access Routes
      - Gold Route
      - Green Route
    - Open Access Use and Reuse
  - Open Data
    - Open Big Data
    - Open Data Definition
    - Open Data Journals
    - Open Data Standards
    - Open Data Use and Reuse
    - Open Government Data
  - Open Reproducible Research
    - Definition of Open Reproducible Research
    - Irreproducibility Studies
    - Open Lab/Notebooks
    - Open Science Workflows
    - Open Source in Open Science
    - Reproducibility Guidelines
    - Reproducibility Testing
  - Open Science Definition
  - Open Science Evaluation
    - Open Metrics and Impact
      - Altmetrics
      - Bibliometrics
      - Semantometrics
      - Webometrics
    - Open Peer Review
  - Open Science Guidelines
  - Open Science Policies
    - Organisational mandates
      - Funders policies
      - Governmental policies
      - Institutional policies
    - Subject policies
      - Open Access policies
      - Open Data Policies
  - Open Science Projects
  - Open Science Tools
    - Open Repositories
    - Open Services
    - Open Workflow Tools

FOSTER
www.fosteropenscience.eu

POLITECNICO DI TORINO



Infrastructure School

Assumption:
Efficient research depends on the available tools and applications.
Goal:
Creating openly available platforms, tools and services for scientists.
Keywords:
Collaboration platforms and tools

Pragmatic School

Assumption:
Knowledge-creation could be more efficient if scientists worked together.
Goal:
Making the process of knowledge creation more efficient and goal oriented.
Keywords:
Wisdom of the crowds, network effects, Open Data, Open Code

Public School

Assumption:
Science needs to be made accessible to the public.
Goal:
Making science accessible for citizens.
Keywords:
Citizen Science, Science PR, Science Blogging

Open Science

Democratic School

Assumption:
The access to knowledge is unequally distributed.
Goal:
Making knowledge freely available for everyone.
Keywords:
Open access, intellectual property rights, Open data, Open code

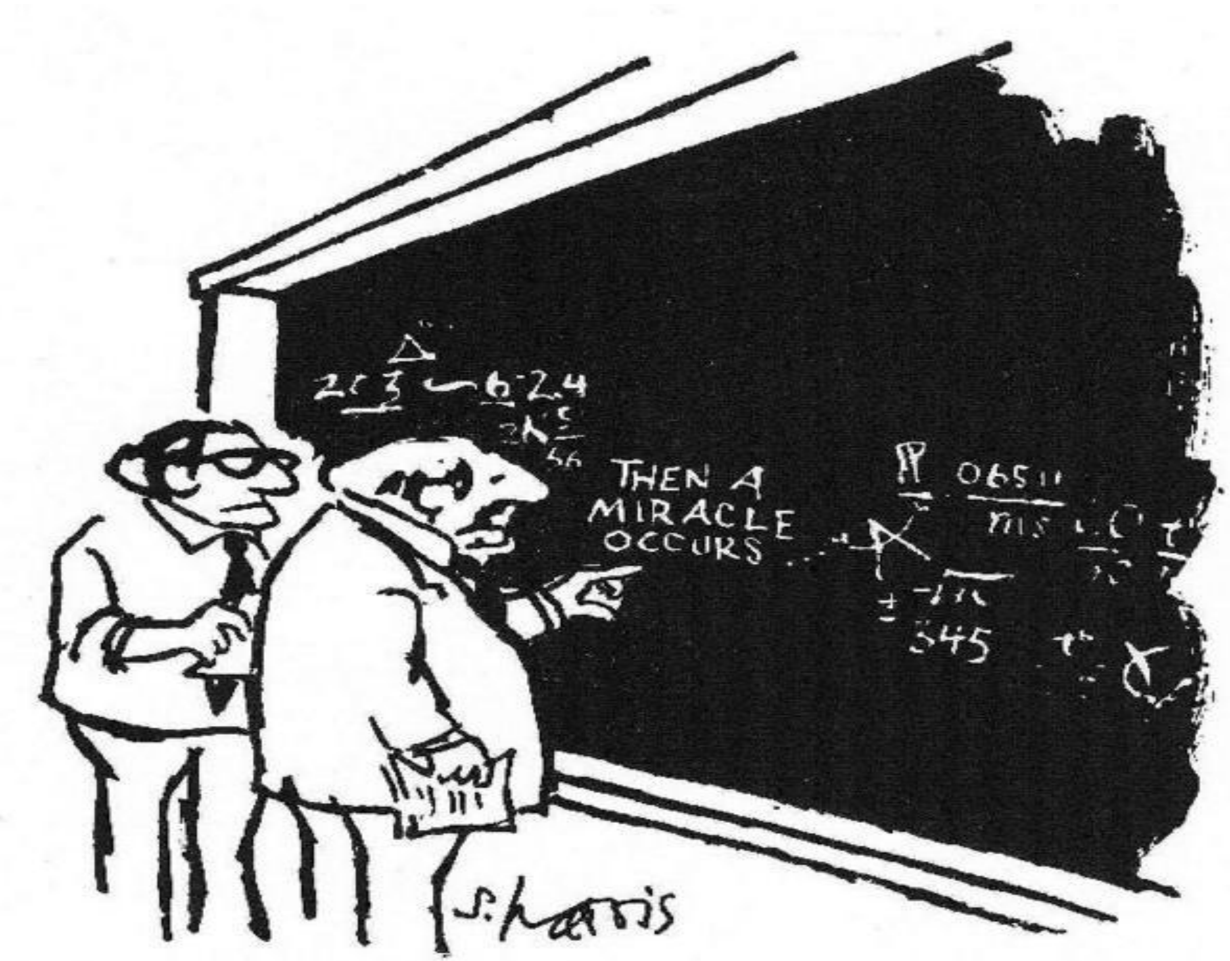
Measurement School

Assumption:
Scientific contributions today need alternative impact measurements.
Goal:
Developing an alternative metric system for scientific impact.
Keywords:
Altmetrics, peer review, citation, impact factors

Open Science: The five schools of thought

Fecher & Friesike, 2014
https://link.springer.com/chapter/10.1007/978-3-319-00026-8_2

Science ≠ Miracles



"I think you should be more explicit here in step two."

Open Science means:

- Evidence based results,
- Transparency, reproducibility, research rigour
- Validation and verification
- Dissemination and access
- And all other things that basically define science.

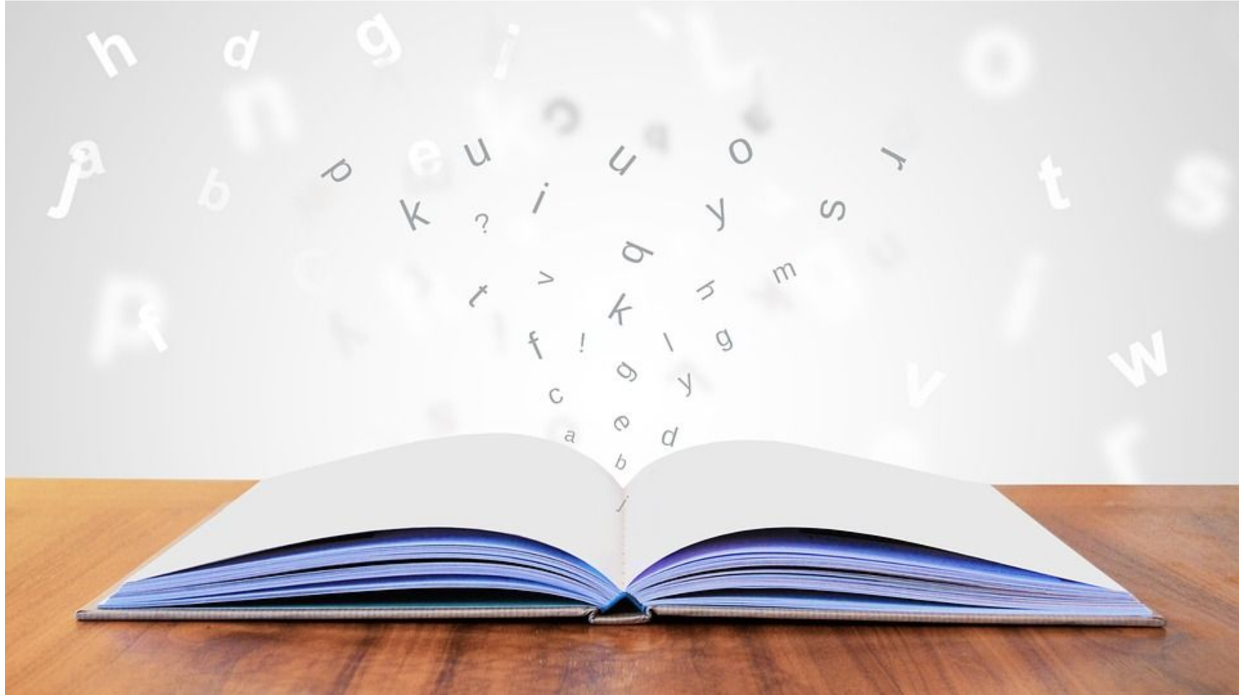
10.5281/zenodo.3472587

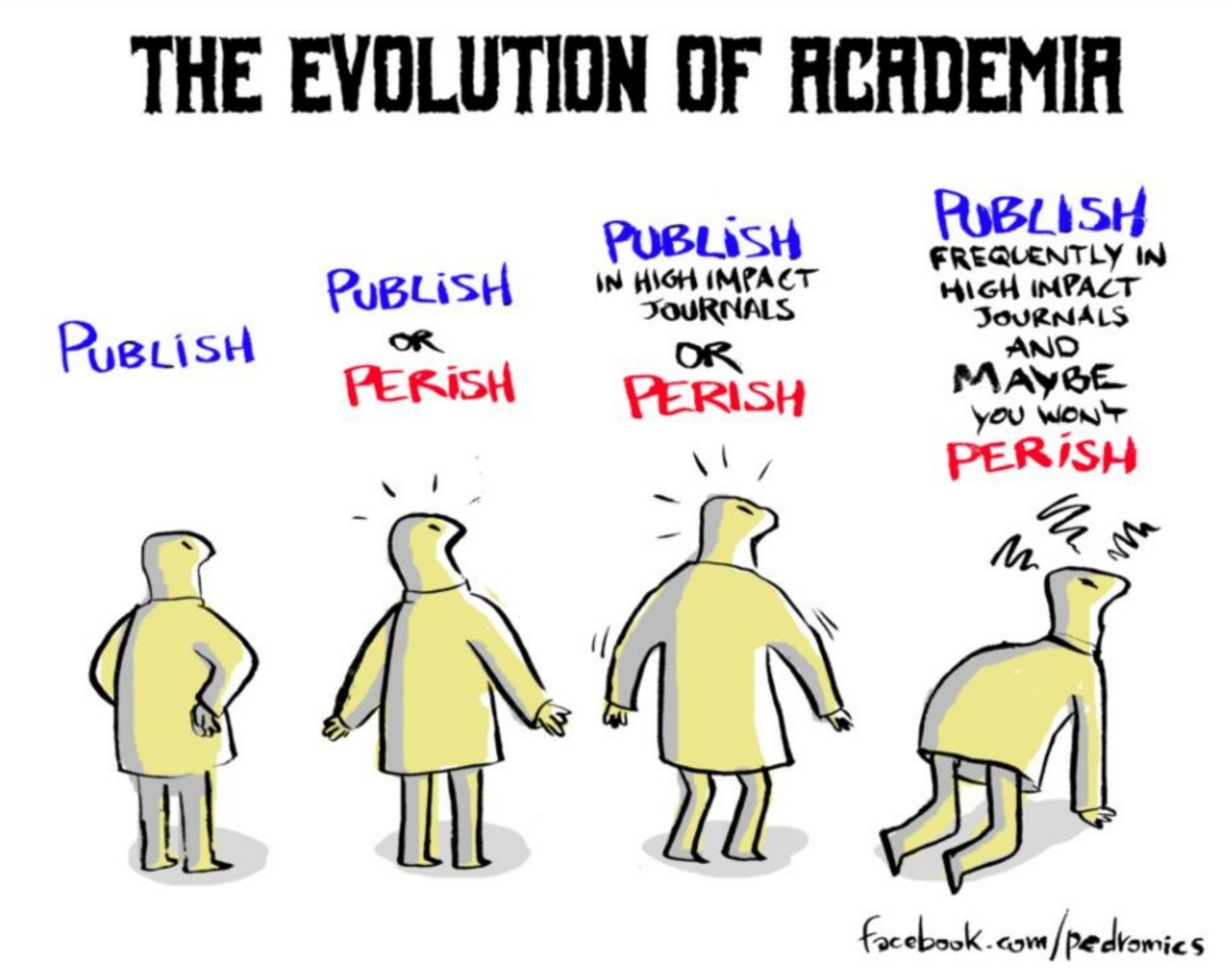
Open science is nothing new, it’s just science

https://zenodo.org/record/1285575

If open science is just science

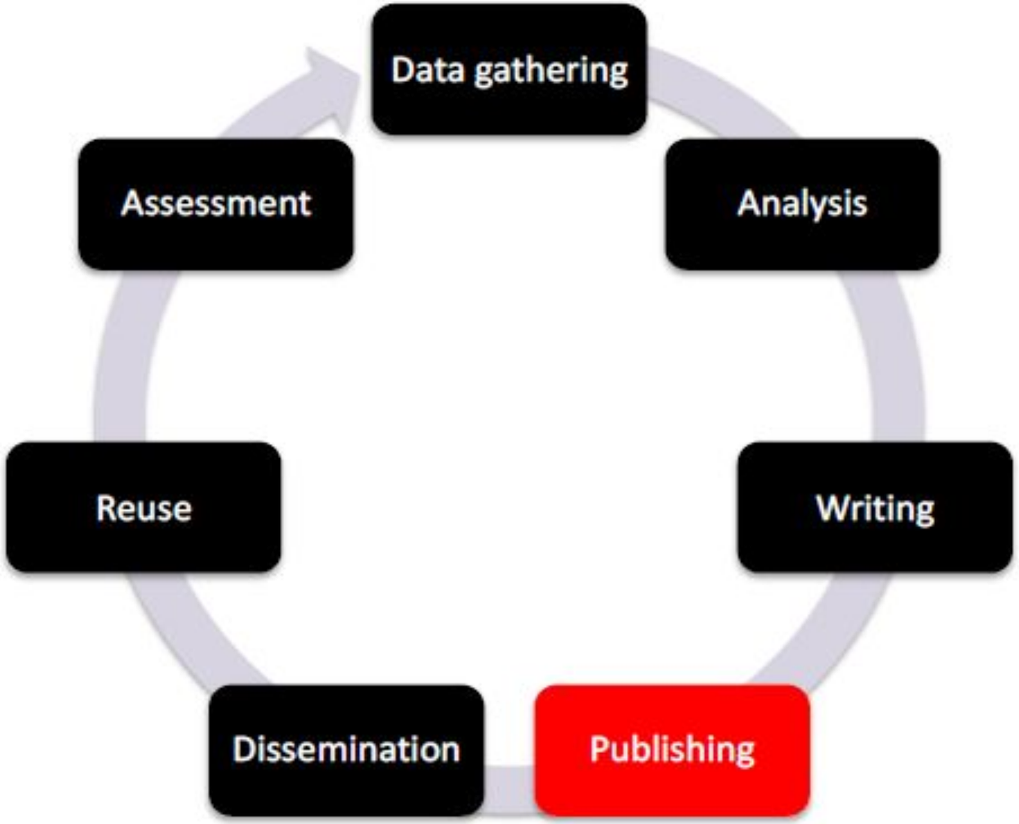
Why is everyone talking about it now!




THE EVOLUTION OF ACADEMIA
PUBLISH
PUBLISH OR PERISH
PUBLISH IN HIGH IMPACT JOURNALS OR PERISH
PUBLISH FREQUENTLY IN HIGH IMPACT JOURNALS AND MAYBE YOU WON'T PERISH
facebook.com/pedromics

Copyright:
https://velica.deviantart.com/art/Publish-or-Perish-645355248

How did we get there?



The only thing that counts in academia is
publication of novel results in high impact
journals

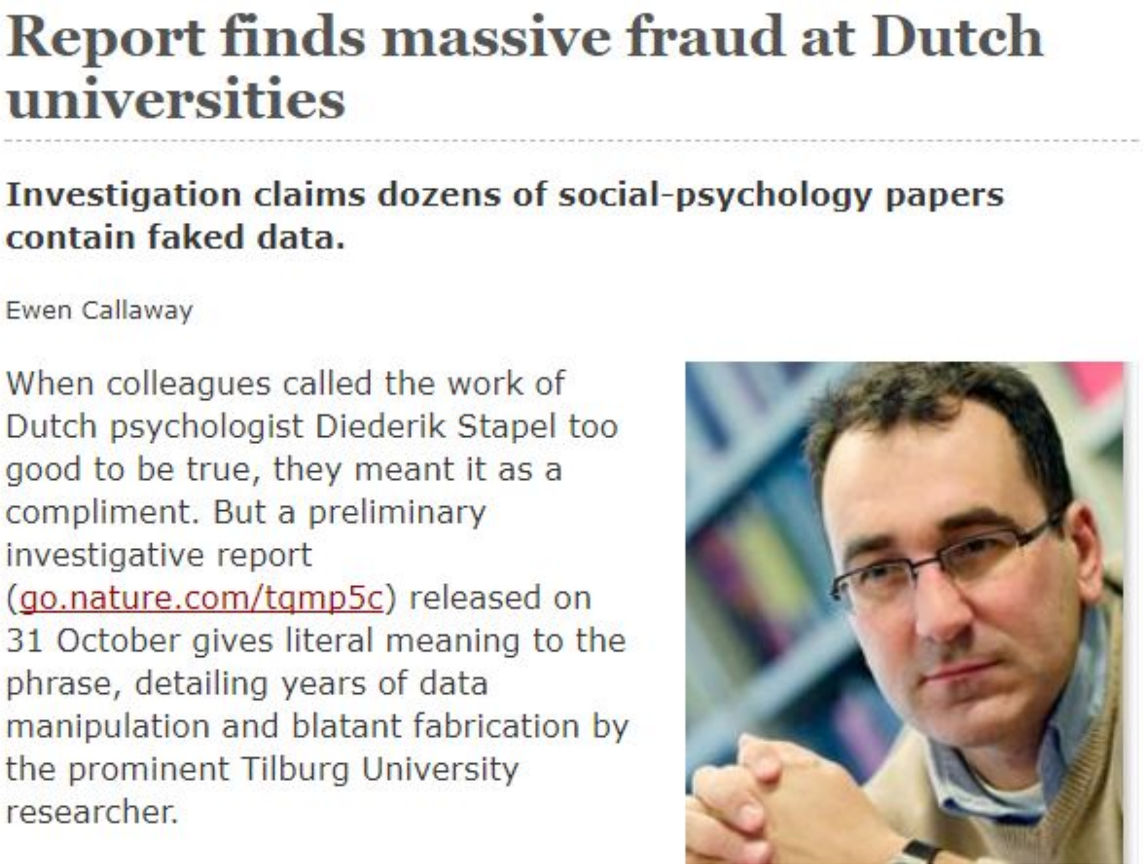
https://www.repository.cam.ac.uk/handle/1810/276106

How did we get here?

- Design the study
- Collect data
- Analyze data as prespecified
- Oops! P> 0.05?
- Torture data until it confesses
- Then, and only then… write the manuscript

Adapted from:
Turner, Erick (2016):
https://doi.org/10.6084/m9.figshare.3381379.v1 c

Some consequences: In extreme cases



Report finds massive fraud at Dutch universities

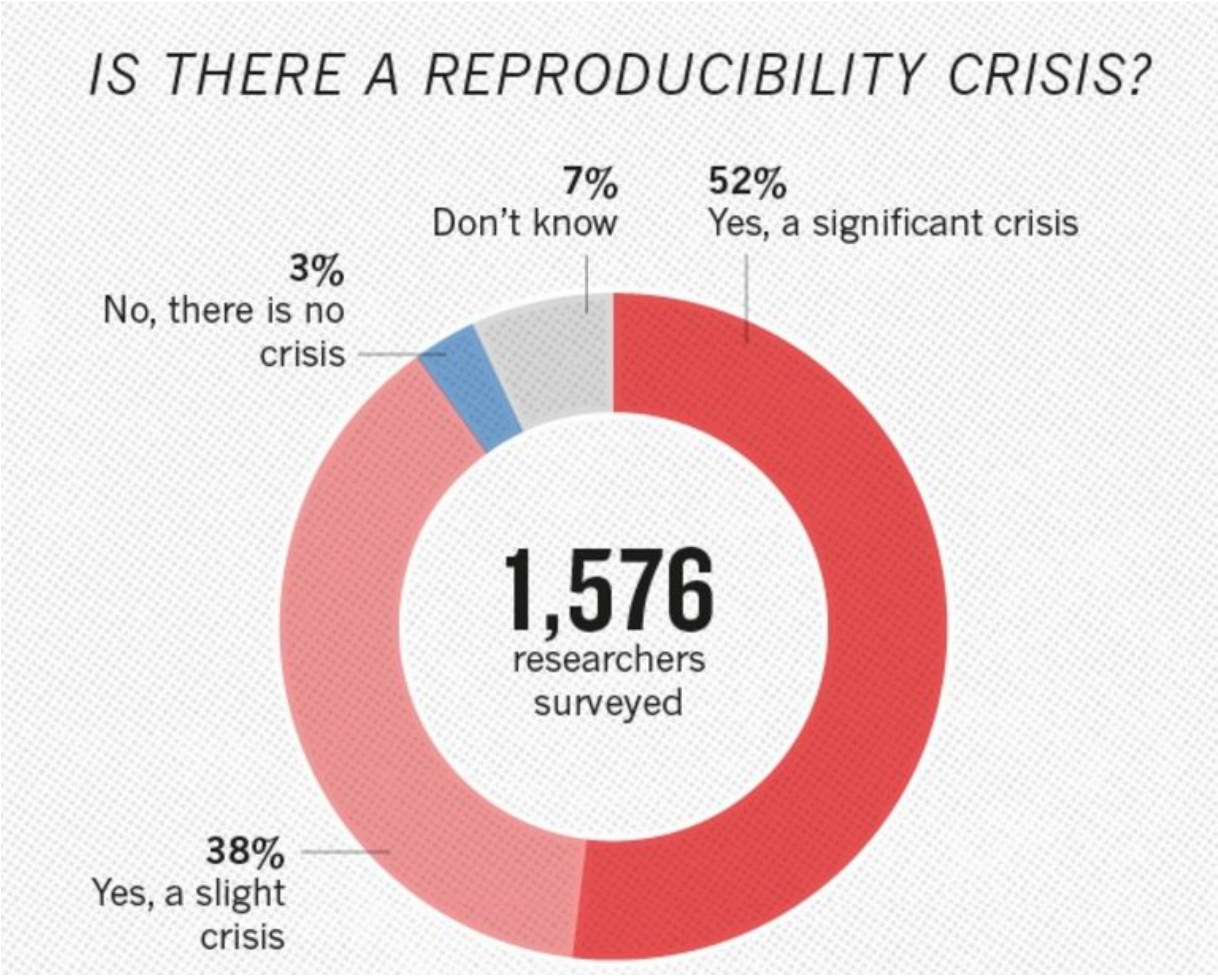
Investigation claims dozens of social-psychology papers contain faked data.

Ewen Callaway

When colleagues called the work of Dutch psychologist Diederik Stapel too good to be true, they meant it as a compliment. But a preliminary investigative report (go.nature.com/tqmp5c) released on 31 October gives literal meaning to the phrase, detailing years of data manipulation and blatant fabrication by the prominent Tilburg University researcher.

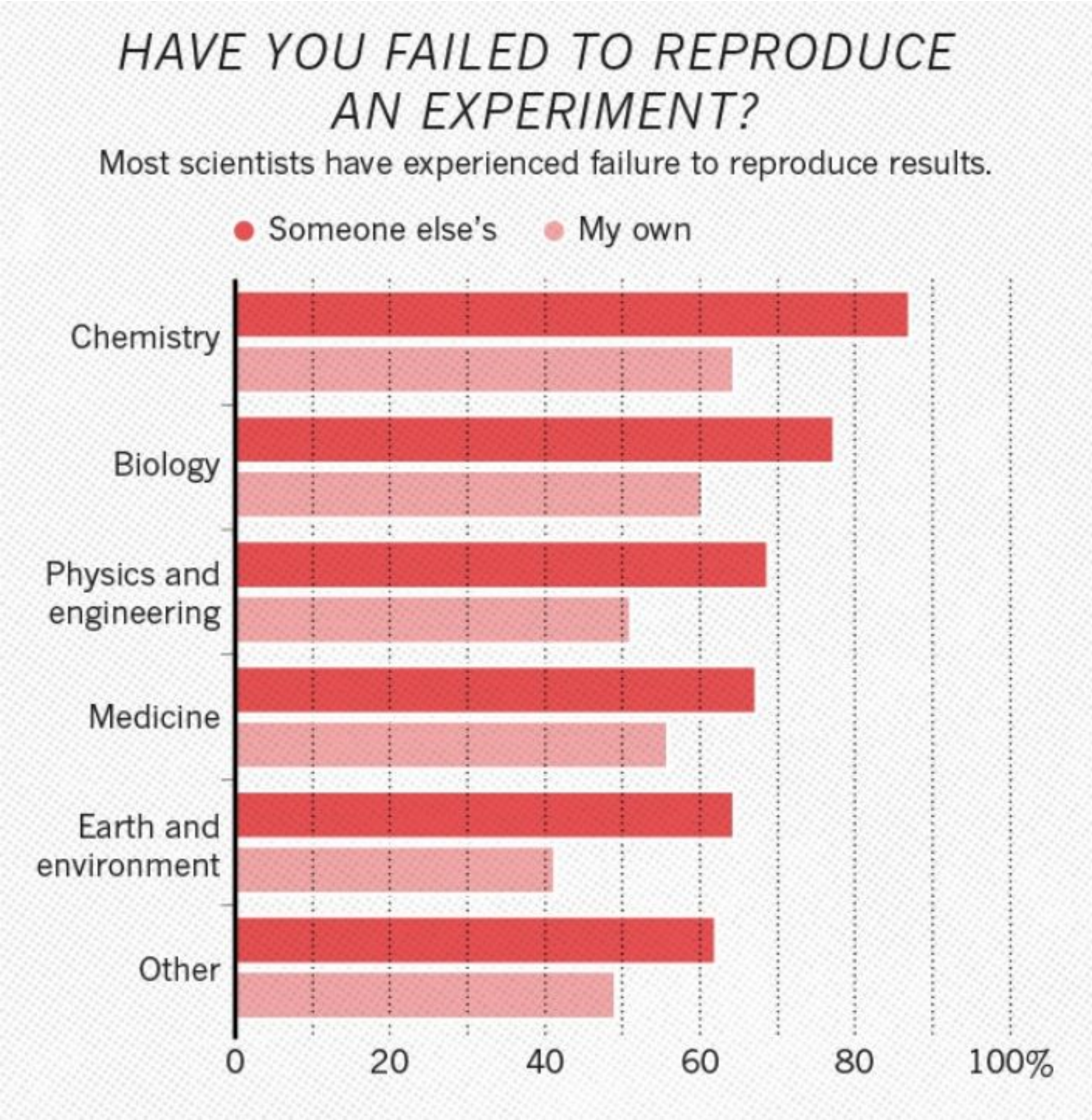
In everyday scientific practice, fraud is minimal, but the main issue is the reproducibility

The reproducibility crisis



https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

In the engineering and physics disciplines

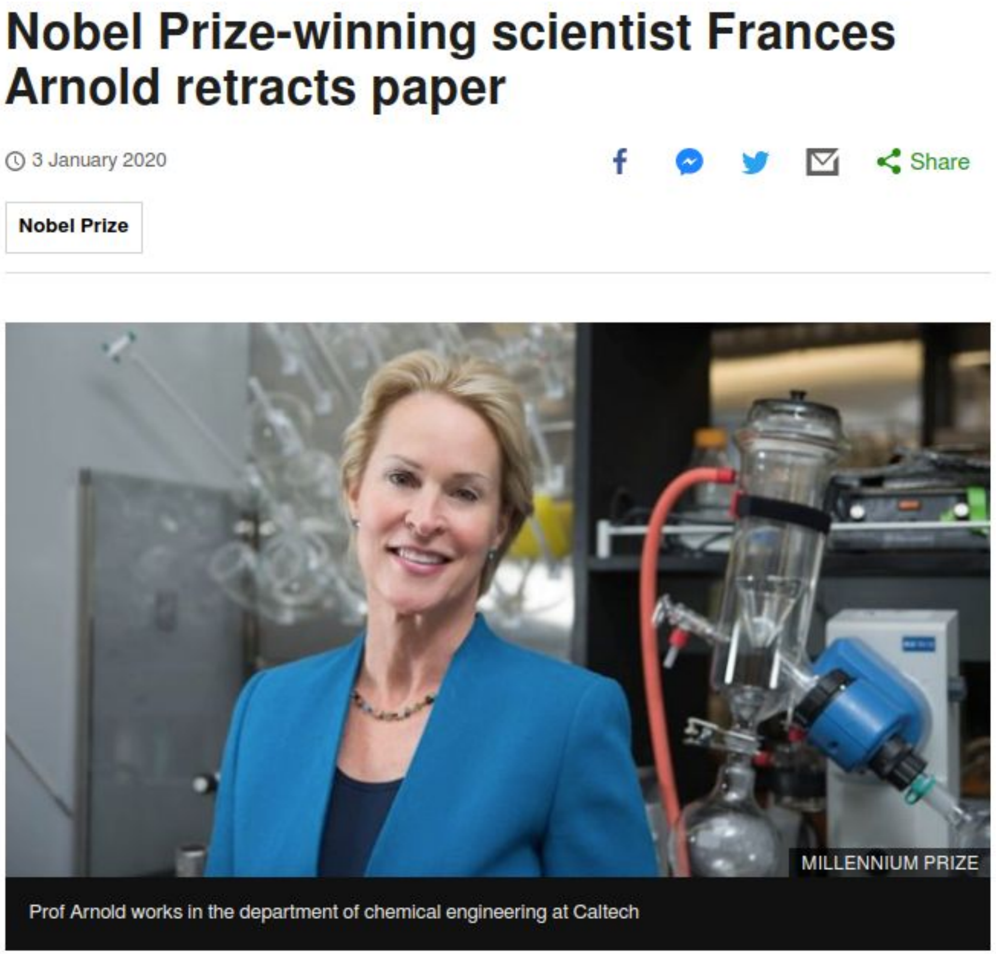


About 70% cannot reproduce others' experiments

and more than 50% cannot reproduce their own experiments!

Nature 533, 452–454 doi:10.1038/533452a

Happens even to the best of scientists



Nobel Prize-winning scientist Frances Arnold retracts paper

3 January 2020

Share

Nobel Prize

MILLENNIUM PRIZE

Prof Arnold works in the department of chemical engineering at Caltech

“It has been **retracted because the results were not reproducible,** and the authors found data missing from a lab notebook.

Reproduction is an essential part of validating scientific experiments. If an experiment is a success, one would expect to get the same results every time it was conducted.”

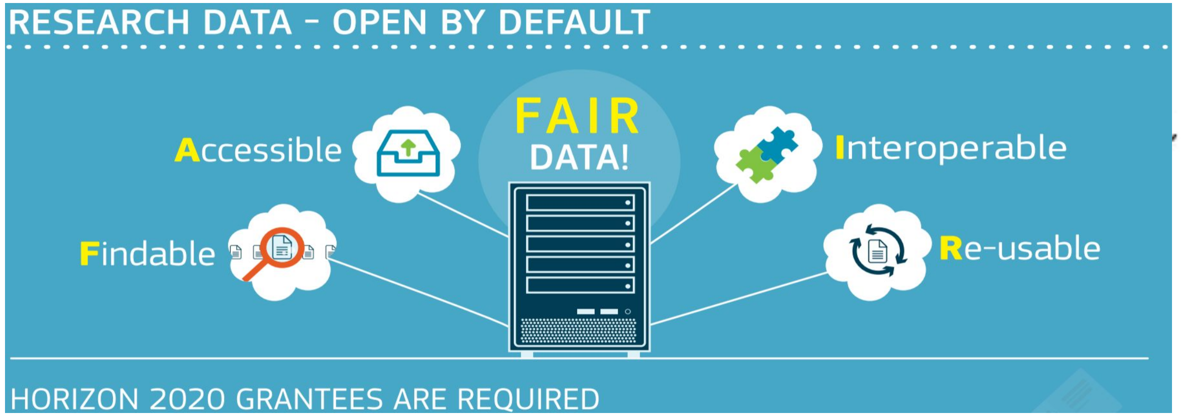
https://www.bbc.com/news/world-us-canada-50989423

Reasons for the crisis

- Selective reporting
- Pressure to publish
- Insufficient supervision and training
- **Supporting data / methods / code not available**

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature, [online] 533(7604), pp.452-454. Available at: https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970 [Accessed 26 Apr. 2018].

Funding bodies are pushing for open science, focus on FAIR data



RESEARCH DATA - OPEN BY DEFAULT

FAIR DATA!

Accessible

Findable

Interoperable

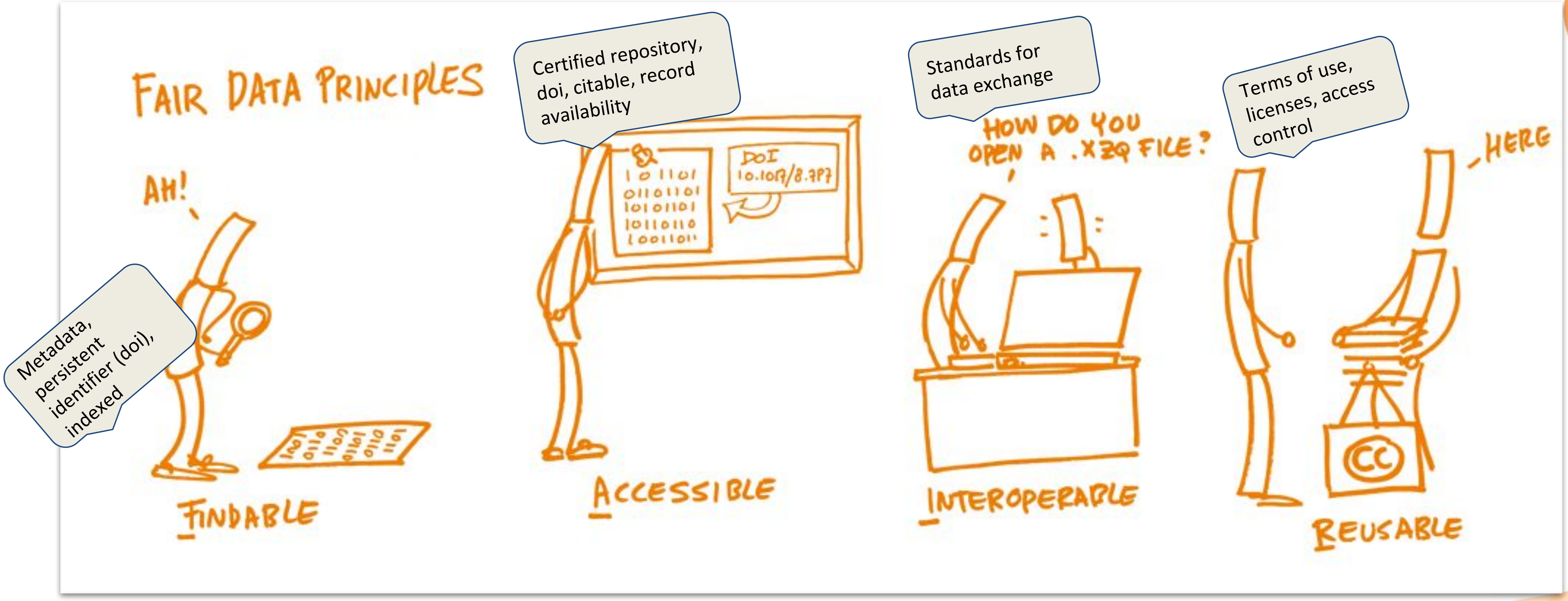
Re-usable

HORIZON 2020 GRANTEES ARE REQUIRED

More are following, nationally and regionally

https://www.ukri.org/funding/information-for-award-holders/data-policy/common-principles-on-data-policy/

What is FAIR data?



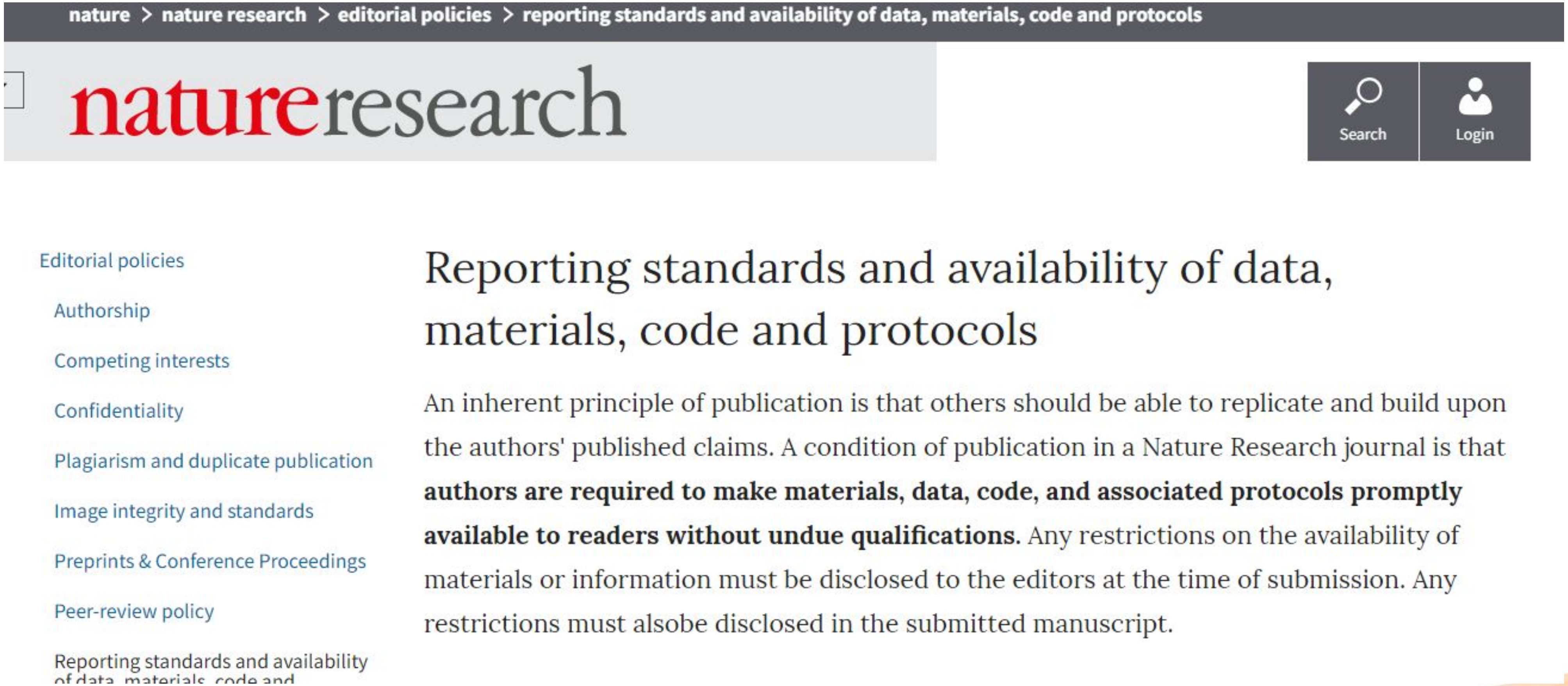
FAIR DATA PRINCIPLES
Certified repository, doi, citable, record availability
Standards for data exchange
Terms of use, licenses, access control
AH!
DOI 10.1017/8.787
HOW DO YOU OPEN A .XZQ FILE?
HERE
Metadata, persistent identifier (doi), indexed
FINDABLE
ACCESSIBLE
INTEROPERABLE
REUSABLE

Hochstenbach, P. (2018). *Open Research Data Material - FAIR data principles*. [image] Available at: https://hochstenbach.wordpress.com/ [Accessed 26 Apr. 2018].

You can have a closed/restricted access and still be FAIR

More on that on feb 20 :)

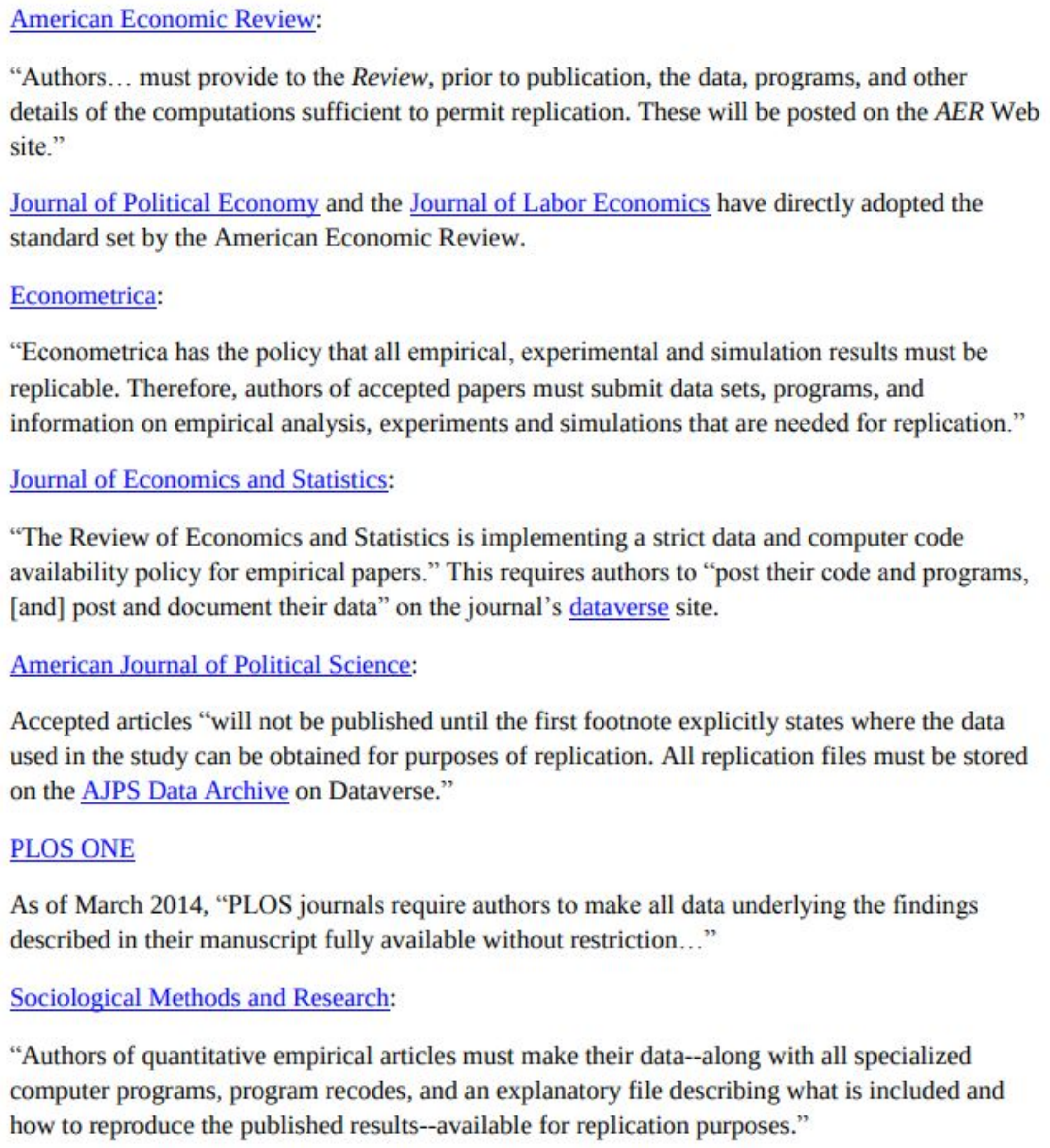
Top journals need it already!


nature > nature research > editorial policies > reporting standards and availability of data, materials, code and protocols
natureresearch
Search
Login
Editorial policies
Authorship
Competing interests
Confidentiality
Plagiarism and duplicate publication
Image integrity and standards
Preprints & Conference Proceedings
Peer-review policy
Reporting standards and availability of data, materials, code and

Reporting standards and availability of data, materials, code and protocols

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. A condition of publication in a Nature Research journal is that **authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications.** Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission. Any restrictions must alsobe disclosed in the submitted manuscript.

Top journals in all fields need them!



American Economic Review:

"Authors… must provide to the *Review*, prior to publication, the data, programs, and other details of the computations sufficient to permit replication. These will be posted on the *AER* Web site."

Journal of Political Economy and the Journal of Labor Economics have directly adopted the standard set by the American Economic Review.

Econometrica:

"Econometrica has the policy that all empirical, experimental and simulation results must be replicable. Therefore, authors of accepted papers must submit data sets, programs, and information on empirical analysis, experiments and simulations that are needed for replication."

Journal of Economics and Statistics:

"The Review of Economics and Statistics is implementing a strict data and computer code availability policy for empirical papers." This requires authors to "post their code and programs, [and] post and document their data" on the journal's dataverse site.

American Journal of Political Science:

Accepted articles "will not be published until the first footnote explicitly states where the data used in the study can be obtained for purposes of replication. All replication files must be stored on the AJPS Data Archive on Dataverse."

PLOS ONE

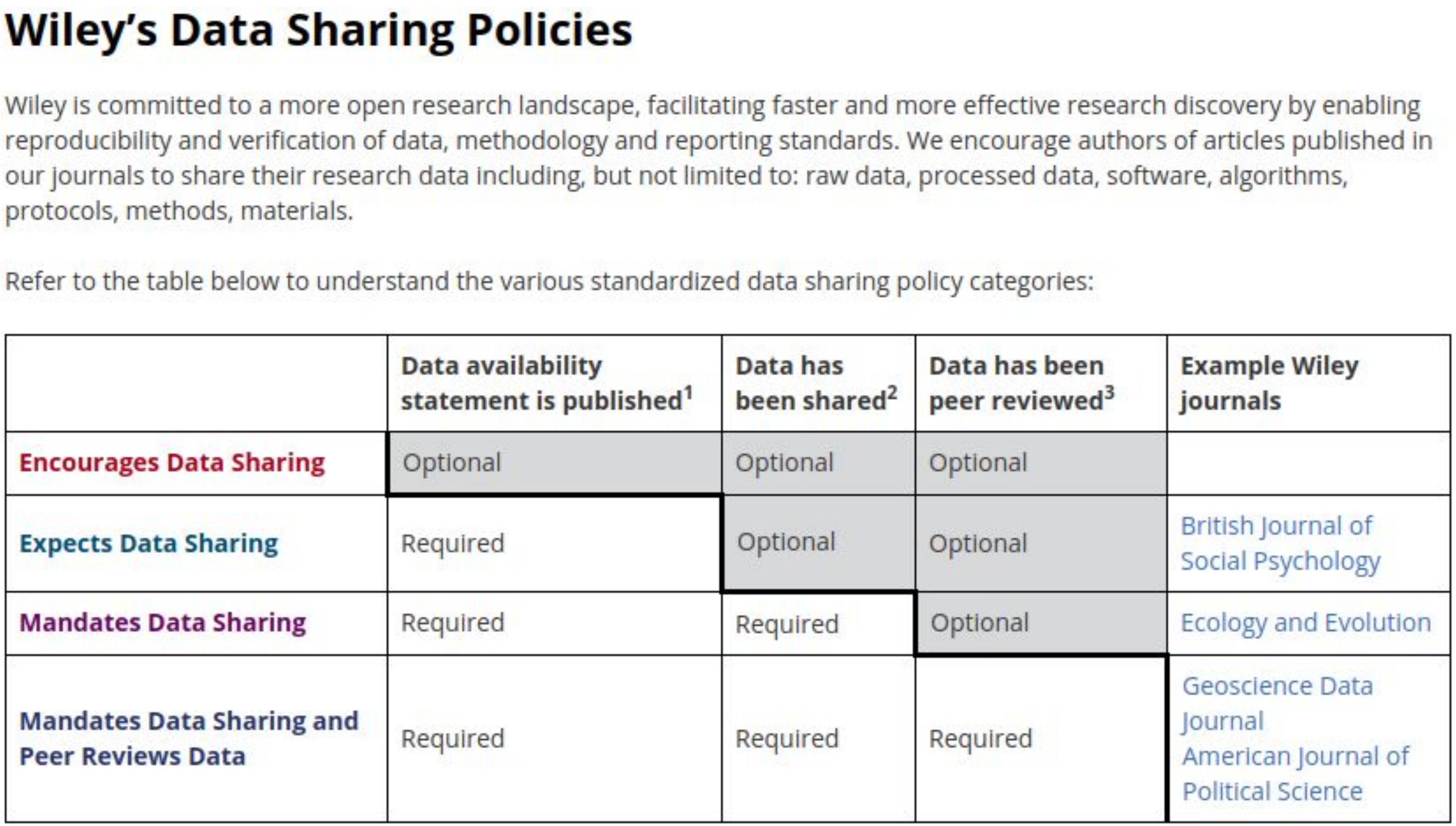
As of March 2014, "PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction…"

Sociological Methods and Research:

"Authors of quantitative empirical articles must make their data--along with all specialized computer programs, program recodes, and an explanatory file describing what is included and how to reproduce the published results--available for replication purposes."

https://web.stanford.edu/~cy10/public/data/Data_Availability_Policies.pdf

The big publishers have data policies



Wiley's Data Sharing Policies

Wiley is committed to a more open research landscape, facilitating faster and more effective research discovery by enabling reproducibility and verification of data, methodology and reporting standards. We encourage authors of articles published in our journals to share their research data including, but not limited to: raw data, processed data, software, algorithms, protocols, methods, materials.

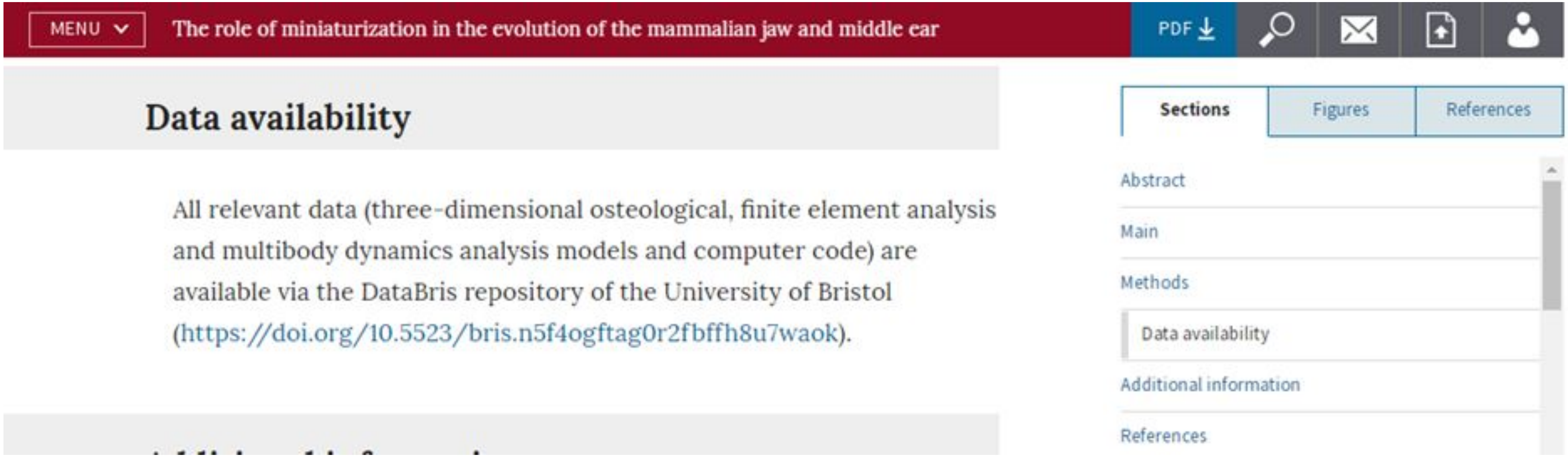
Refer to the table below to understand the various standardized data sharing policy categories:

	Data availability statement is published[1]	Data has been shared[2]	Data has been peer reviewed[3]	Example Wiley journals
Encourages Data Sharing	Optional	Optional	Optional	
Expects Data Sharing	Required	Optional	Optional	British Journal of Social Psychology
Mandates Data Sharing	Required	Required	Optional	Ecology and Evolution
Mandates Data Sharing and Peer Reviews Data	Required	Required	Required	Geoscience Data Journal American Journal of Political Science

https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-policy.html

https://www.springernature.com/it/authors/research-data-policy/data-availability-statements/12330880 (springer nature)

Data availability statement?



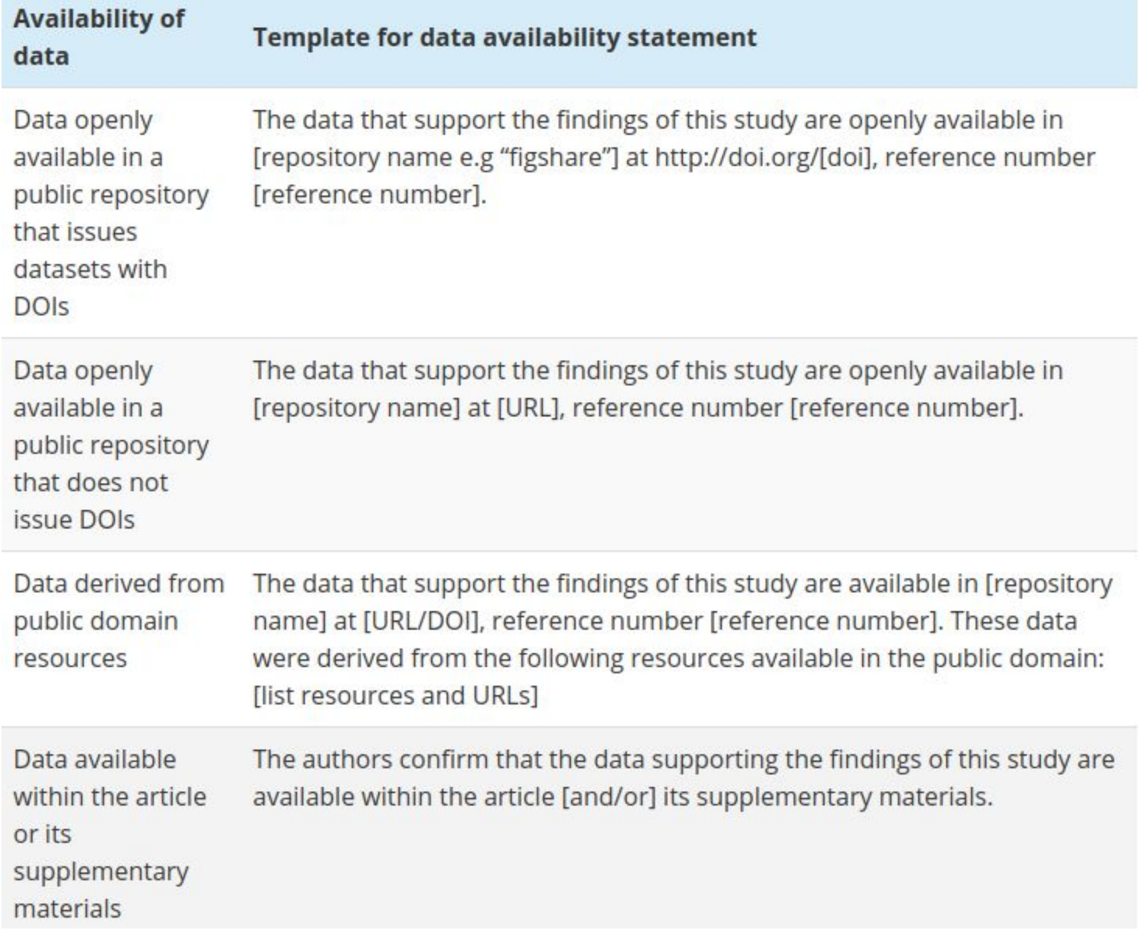
The role of miniaturization in the evolution of the mammalian jaw and middle ear

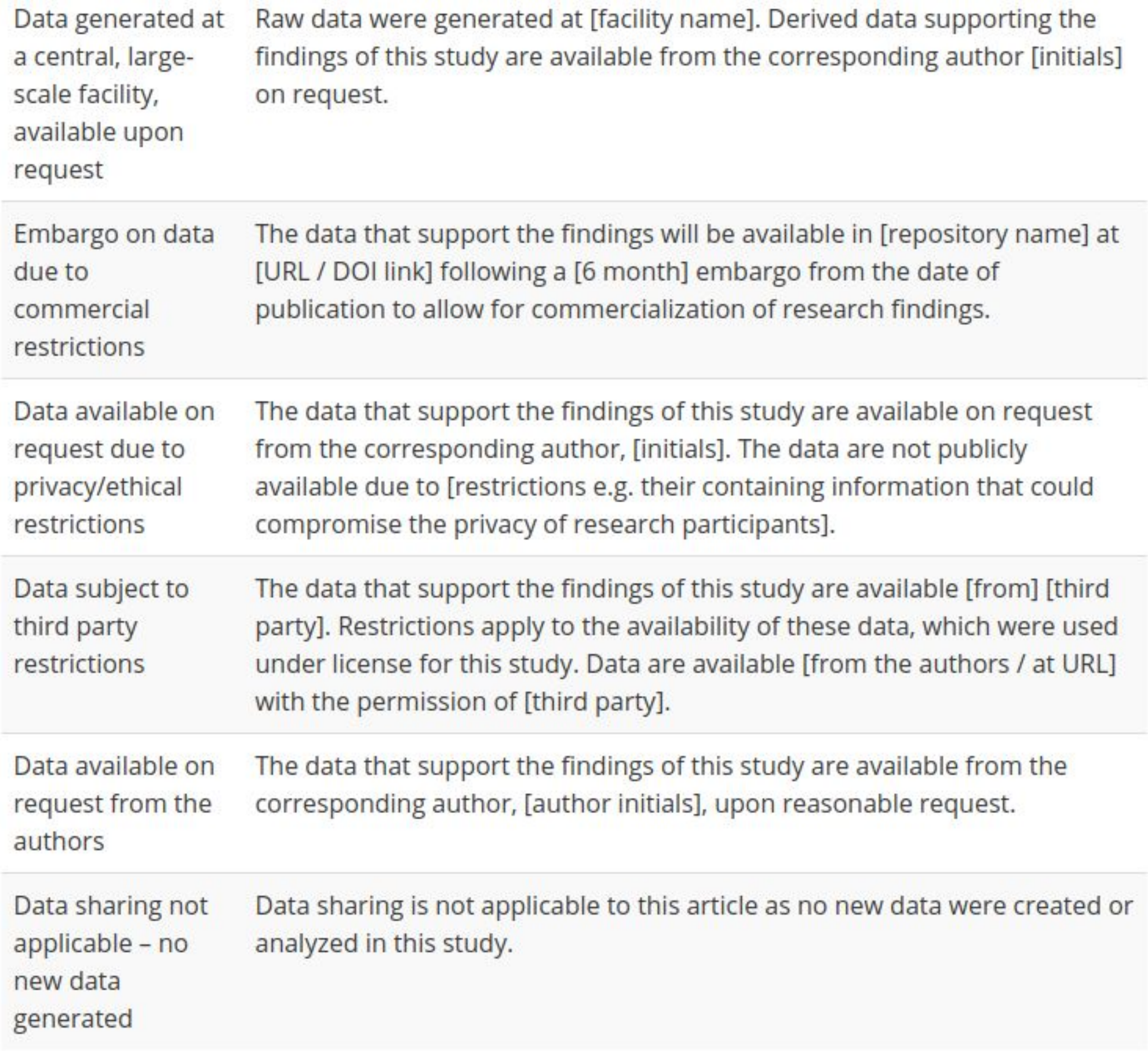
Data availability

All relevant data (three-dimensional osteological, finite element analysis and multibody dynamics analysis models and computer code) are available via the DataBris repository of the University of Bristol (https://doi.org/10.5523/bris.n5f4ogftag0r2fbffh8u7waok).

https://www.nature.com/articles/s41586-018-0521-4

Other examples of data availability





Availability of data	Template for data availability statement
Data openly available in a public repository that issues datasets with DOIs	The data that support the findings of this study are openly available in [repository name e.g "figshare"] at http://doi.org/[doi], reference number [reference number].
Data openly available in a public repository that does not issue DOIs	The data that support the findings of this study are openly available in [repository name] at [URL], reference number [reference number].
Data derived from public domain resources	The data that support the findings of this study are available in [repository name] at [URL/DOI], reference number [reference number]. These data were derived from the following resources available in the public domain: [list resources and URLs]
Data available within the article or its supplementary materials	The authors confirm that the data supporting the findings of this study are available within the article [and/or] its supplementary materials.
Data generated at a central, large-scale facility, available upon request	Raw data were generated at [facility name]. Derived data supporting the findings of this study are available from the corresponding author [initials] on request.
Embargo on data due to commercial restrictions	The data that support the findings will be available in [repository name] at [URL / DOI link] following a [6 month] embargo from the date of publication to allow for commercialization of research findings.
Data available on request due to privacy/ethical restrictions	The data that support the findings of this study are available on request from the corresponding author, [initials]. The data are not publicly available due to [restrictions e.g. their containing information that could compromise the privacy of research participants].
Data subject to third party restrictions	The data that support the findings of this study are available [from] [third party]. Restrictions apply to the availability of these data, which were used under license for this study. Data are available [from the authors / at URL] with the permission of [third party].
Data available on request from the authors	The data that support the findings of this study are available from the corresponding author, [author initials], upon reasonable request.
Data sharing not applicable – no new data generated	Data sharing is not applicable to this article as no new data were created or analyzed in this study.

https://authorservices.taylorandfrancis.com/data-sharing-policies/data-availability-statements/

In any context, RDM is key!

We cannot achieve open science or FAIR data without RDM.

It is not simple but crucial!

Solutions at the intersection of technology, culture and awareness and policy

Now you know the why you need to care about open science and RDM, let’s see what you can do!

Immediate reactions to “data” requirements:

- It would take me 5 years to find all my data!
- The PhD/postdoc who had the data left the lab
- Should we write down all protocols?
- Data management is a waste of time
- Nobody will understand my data
- People can just ask me for it when they need it

What if

- What if someone asks you for data supporting your publication?

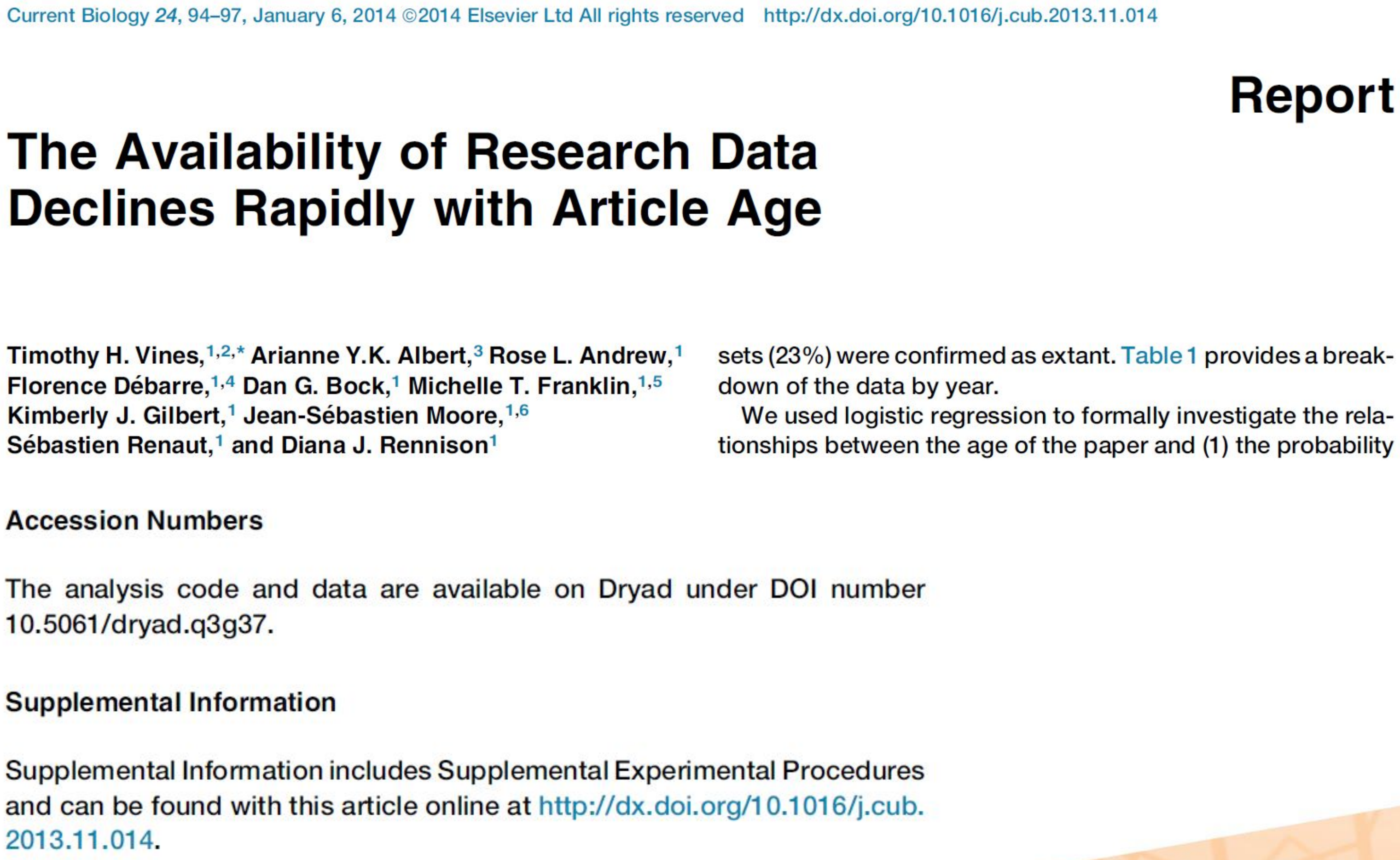
What if

- What if someone asks you for data supporting your publication?
- What if someone asks you for data supporting your publication, 5 year after publication?

What if

- What if someone asks you for data supporting your publication?
- What if someone asks you for data supporting your publication, 5 year after publication?
- What if the request comes 10 years later?

Infact..



Current Biology *24*, 94–97, January 6, 2014 ©2014 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2013.11.014

Report

The Availability of Research Data Declines Rapidly with Article Age

Timothy H. Vines,[1,2,*] Arianne Y.K. Albert,[3] Rose L. Andrew,[1] Florence Débarre,[1,4] Dan G. Bock,[1] Michelle T. Franklin,[1,5] Kimberly J. Gilbert,[1] Jean-Sébastien Moore,[1,6] Sébastien Renaut,[1] and Diana J. Rennison[1]

sets (23%) were confirmed as extant. Table 1 provides a breakdown of the data by year.

We used logistic regression to formally investigate the relationships between the age of the paper and (1) the probability

Accession Numbers

The analysis code and data are available on Dryad under DOI number 10.5061/dryad.q3g37.

Supplemental Information

Supplemental Information includes Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2013.11.014.

Research Data Policies - reactions

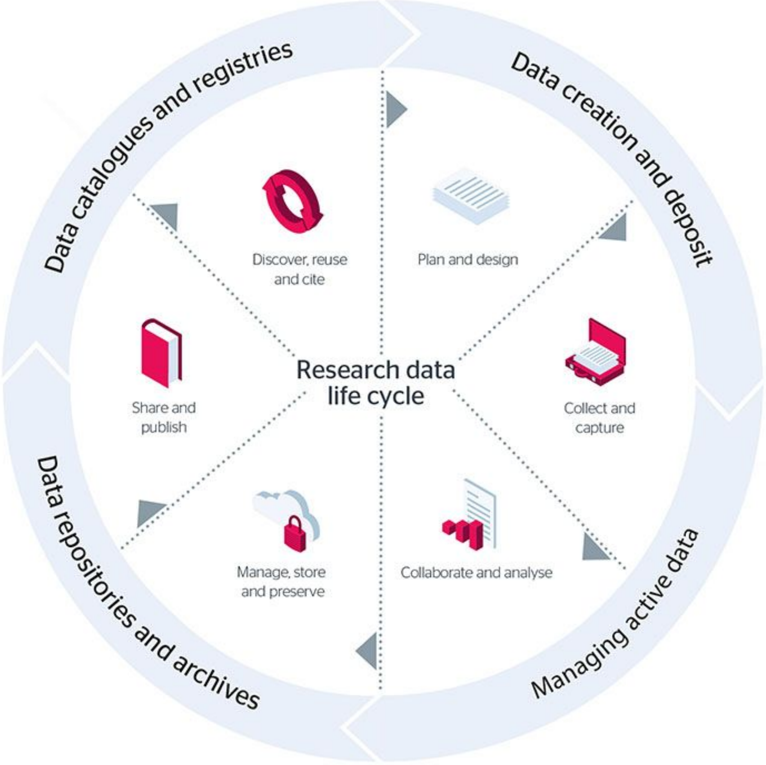
Immediate reactions to “sharing” requirements:

- It would take me 5 years to find all my data!
- The PhD/postdoc who had th
- Sh
- Dat
- Nob y data
- People can just ask me for it when they need it

Good planning is needed from the start

Data Management Plans

A written plan on how you plan and execute your research life cycle



https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data

DMP = Assurance to the funder

- You are aware of their data management and sharing expectations
- You will manage your data well
- You will be prepared to share your data (if and when)
- You will make appropriate resource allocation for this

- **Most importantly, you and your research group will benefit from good data management and sharing**
- **It's a live document throughout your research process**

What does a DMP cover: A checklist

1. Administrative Data
2. Data Collection & Organisation
3. Storage and Backup
4. Documentation and Metadata
5. Ethics and Legal compliance
6. Selection and Preservation
7. Data Sharing
8. Responsibilities and Resources

http://www.dcc.ac.uk/resources/data-management-plans/checklist

What does a DMP cover: A checklist

1. Administrative Data
2. Data Collection & Organisation
3. Storage and Backup
4. Documentation and Metadata
5. Ethics and Legal compliance
6. Selection and Preservation
7. Data Sharing
8. Responsibilities and Resources

We will cover these today
and the rest of the topics
on Feb 20

1. Administrative data

Here you should record basic information to identify and contextualise your plan.

- Basic information e.g. project title, your name, contact details, reference numbers
- A summary of the research to explain the purpose for which data are being collected.
- Details of related policies and procedures e.g. institutional data polices

2. Data Collection & Organisation

- Are there any existing data that you can reuse?
- What standards or methodologies will you use to create data?
- Do your chosen formats and software enable sharing and long-term access to the data?
- How will you structure and name your folders and files?
- What quality assurance processes will you adopt?

2.1 Data collection & organisation: data types

- Raw instrument readings
  - proprietary data (consider converting into common file types)
- Images
- Tabular data (Excel, txt, csv…)
- Genomic data
- Proteomic data
- Patient data
- Documentation in lab notebooks
- Protocols
- Code / software

- **Will you generate your own data?**
- **Will you re-use somebody else’s data? Do you have permission to do this?**

2.2 Data collection & organisation: Format

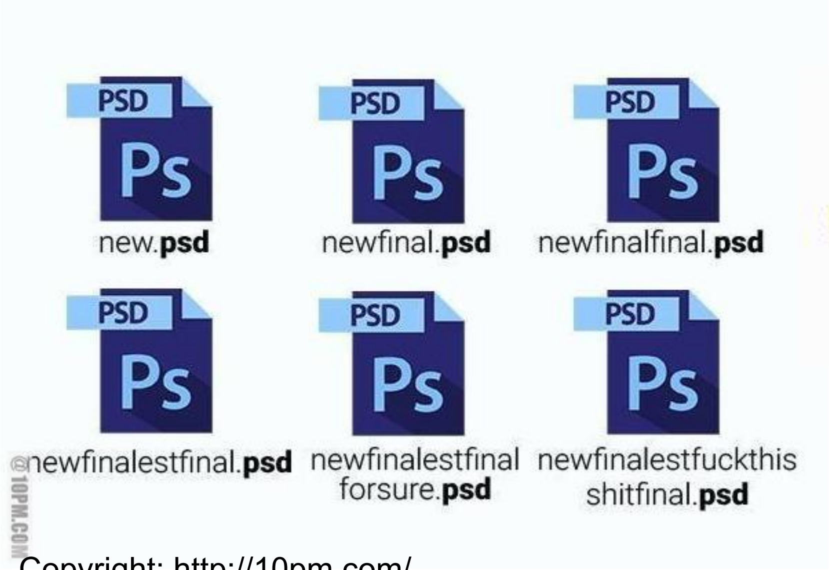
- Non-proprietary, Unencrypted, Uncompressed
- In common usage by the research community, adherent to an open, documented standards
- Containers: TAR, GZIP, ZIP
- Databases: XML, CSV
- Geospatial: SHP, DBF, GeoTIFF, NetCDF
- Moving images: MOV, MPEG, AVI, MXF
- Sounds: WAVE, AIFF, MP3, MXF
- Statistics: ASCII, DTA, POR, SAS, SAV
- Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Tabular data: CSV
- Text: XML, PDF/A, HTML, ASCII, UTF-8
- Web archive: WARC

https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats

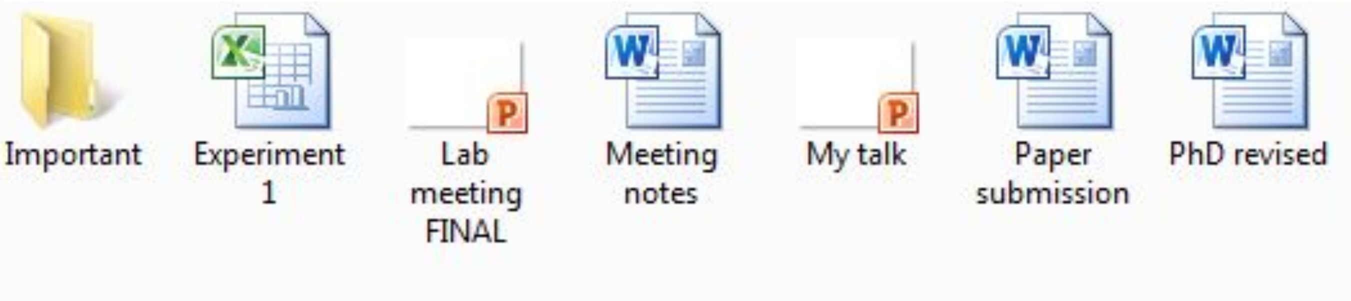
2.3. Data collection & organisation: File & folder structure 1/4

- Consistent
- Meaningful to you and your colleagues
- Think about the continuity of your research
- Would you be able to easily find your own data files?
- Would your colleague be able to easily find and understand your data files?
- Critically look at your files (and clean them if needed!) once a week

2.3. Data collection & organisation: File & folder structure 2/4



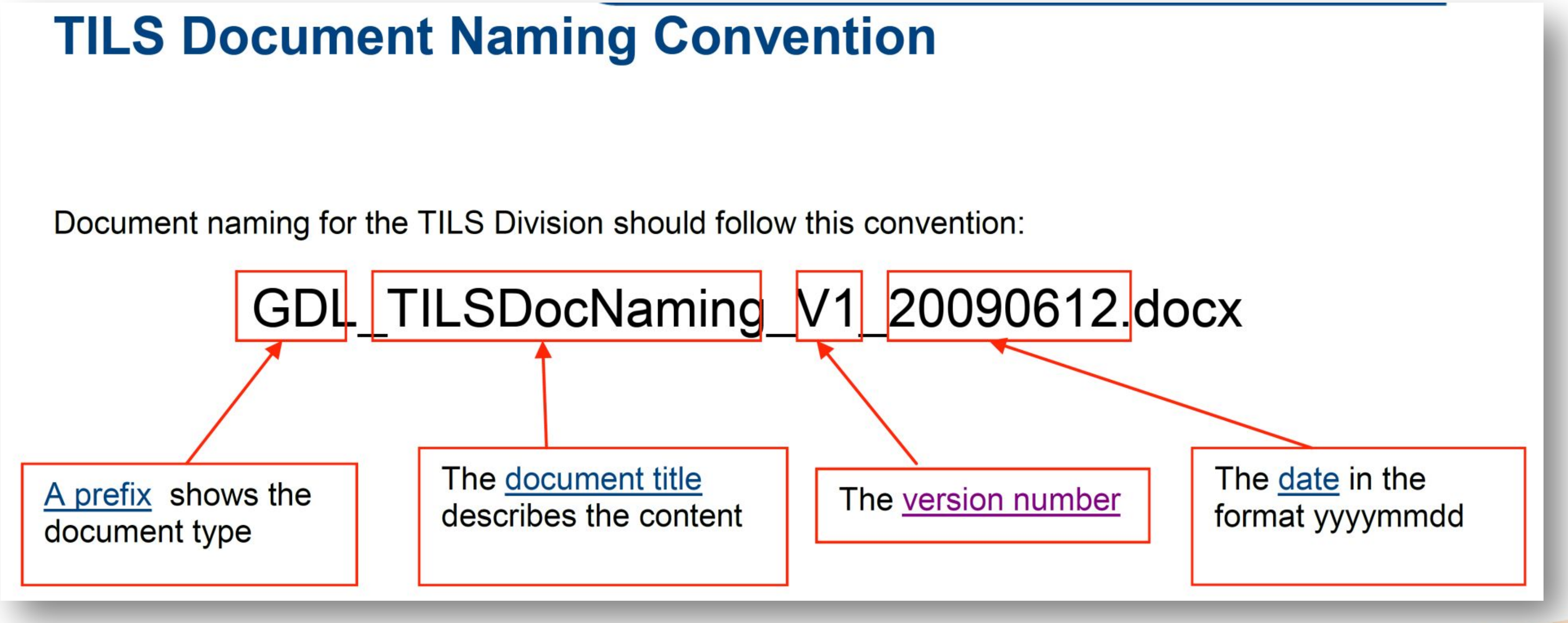
Copyright: http://10pm.com/



Would you recognise these in three years?

2.3. Data collection & organisation: File & folder structure 3/4

File naming convention e.g.



TILS Document Naming Convention

Document naming for the TILS Division should follow this convention:

GDL_TILSDocNaming_V1_20090612.docx

A prefix shows the document type

The document title describes the content

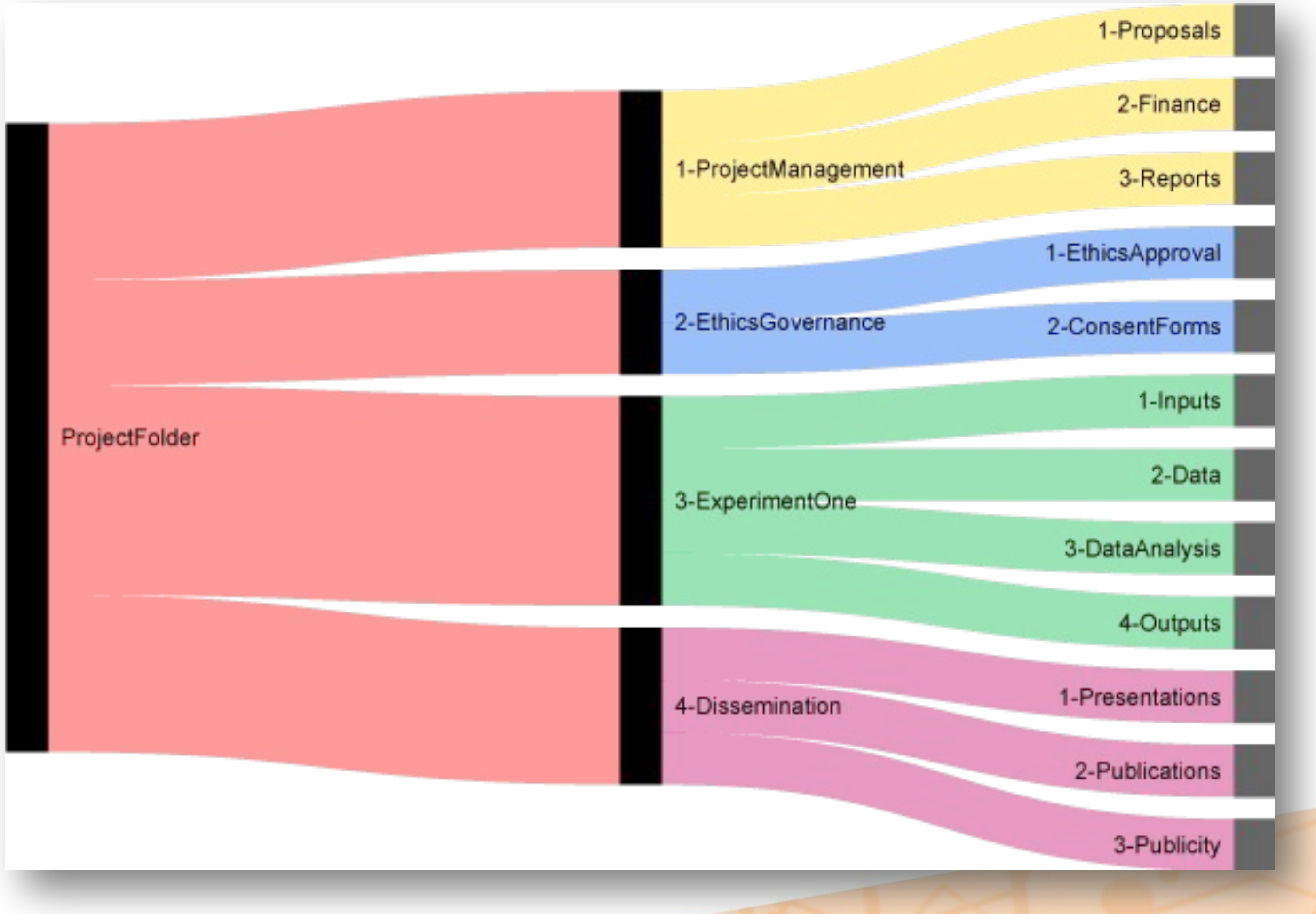
The version number

The date in the format yyyymmdd

http://www.data.cam.ac.uk/files/gdl_tilsdocnaming_v1_20090612.pdf

2.3. Data collection & organisation: File & folder structure 4/4

Folder structures:



Copyright: http://nikola.me/folder_structure.html

Other ways of organising: Electronic Lab Notebooks

- Digital documentation, categorization and linking of
  - Raw, intermediate and final data
  - Experimental and measurement parameters
  - Samples
- Searchable
- Traceable (version control)

https://datamanagement.hms.harvard.edu/electronic-lab-notebooks
https://www.gurdon.cam.ac.uk/institute-life/computing/elnguidance

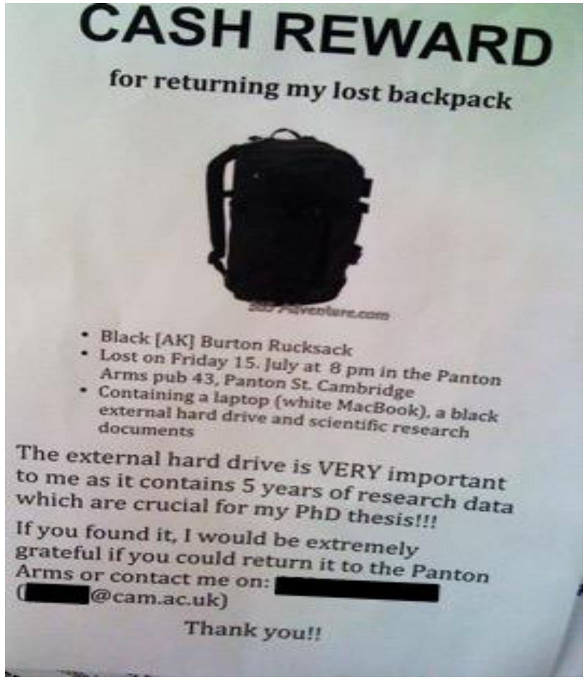
3. Storage and Backup

During your research:

- Where will you store your data?
- Is your data going to be safe?
- How will you share your data with your collaborators?
- Will you use cloud solutions?
- Will you backup your data?
- Is your backup safe?

3.1 Storage and Backup

Data loss is real!



CASH REWARD

for returning my lost backpack

- Black [AK] Burton Rucksack
- Lost on Friday 15. July at 8 pm in the Panton Arms pub 43, Panton St. Cambridge
- Containing a laptop (white MacBook), a black external hard drive and scientific research documents

The external hard drive is VERY important to me as it contains 5 years of research data which are crucial for my PhD thesis!!!

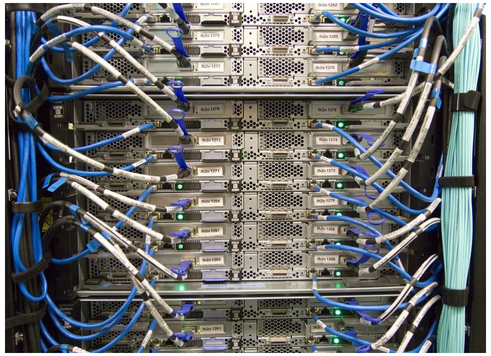
If you found it, I would be extremely grateful if you could return it to the Panton Arms or contact me on: [redacted]
([redacted]@cam.ac.uk)

Thank you!!

3.2 Storage and Data

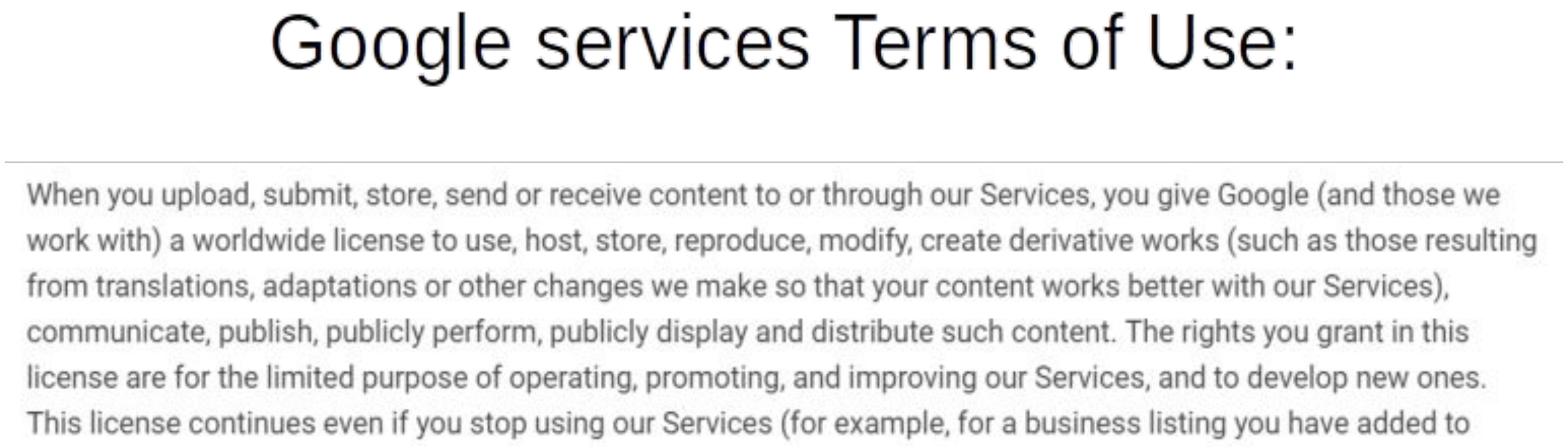
Backup!

- Departmental/institutional backup system
- External drives
- Online backups
- At least two backups, at two different locations



3.3 Storage and Backup: Cloud services

Very convenient, but read the small print



Google services Terms of Use:

When you upload, submit, store, send or receive content to or through our Services, you give Google (and those we work with) a worldwide license to use, host, store, reproduce, modify, create derivative works (such as those resulting from translations, adaptations or other changes we make so that your content works better with our Services), communicate, publish, publicly perform, publicly display and distribute such content. The rights you grant in this license are for the limited purpose of operating, promoting, and improving our Services, and to develop new ones. This license continues even if you stop using our Services (for example, for a business listing you have added to

https://policies.google.com/terms?hl=en

3.4 Storage and Backup: Cloud act & data sovereignty

“The Clarifying Lawful Overseas Use of Data Act or CLOUD Act (H.R. 4943) is a United States federal law enacted in 2018 by the passing of the Consolidated Appropriations Act, 2018, PL 115-141, section 105 executive agreements on access to data by foreign governments. Primarily the CLOUD Act amends the Stored Communications Act (SCA) of 1986 to allow federal law enforcement to **compel U.S.-based technology companies** via **warrant or subpoena to provide requested data stored on servers** regardless of whether the data are stored in **the U.S. or on foreign soil**.”

https://en.wikipedia.org/wiki/CLOUD_Act

PoliTo a lot to offer

- Take a close look at the IT services and offerings of PoliTo (Thanks Mr. Venuto for presenting today)

- It is not only more safe and secure, it is much more professional, many options to explore.

- It reassures funders that the institution professionally supports data needs of researchers

4. Documentation and Metadata

- What documentation and metadata will accompany the data?
- How will you capture / create this documentation and metadata?
- What metadata standards will you use and why?

4.1 Metadata

- Have you ever heard of the term ‘metadata’?

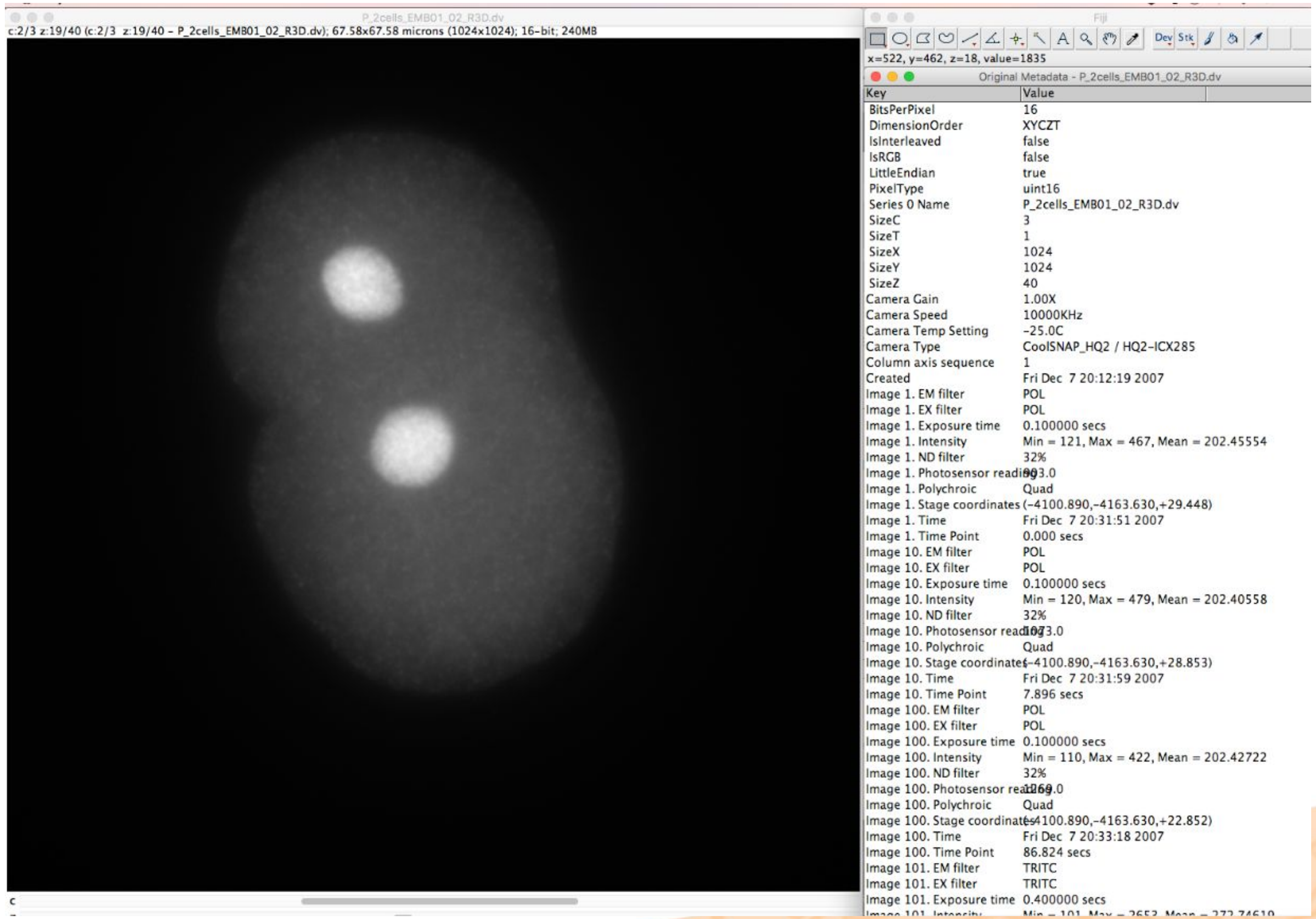
4.1.1 Metadata

- Have you ever heard of the term ‘metadata’?

  - Metadata = information about data
    - Contextual information about your data collection
  - Is it important?
    - Yes, if you want your research to be **reproducible**

4.2 Many ways of describing data

- Automated description added by software



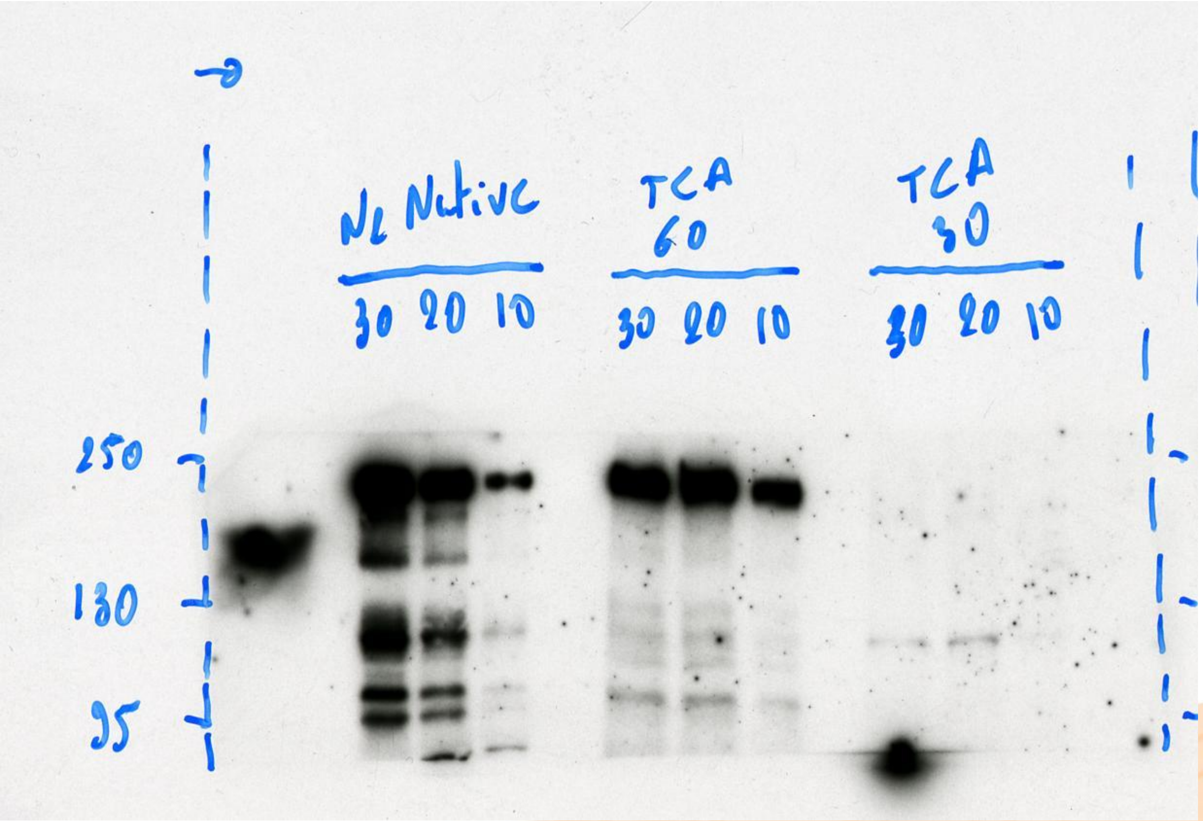
c:2/3 z:19/40 (c:2/3 z:19/40 – P_2cells_EMB01_02_R3D.dv); 67.58x67.58 microns (1024x1024); 16-bit; 240MB

x=522, y=462, z=18, value=1835

Key	Value
BitsPerPixel	16
DimensionOrder	XYCZT
IsInterleaved	false
IsRGB	false
LittleEndian	true
PixelType	uint16
Series 0 Name	P_2cells_EMB01_02_R3D.dv
SizeC	3
SizeT	1
SizeX	1024
SizeY	1024
SizeZ	40
Camera Gain	1.00X
Camera Speed	10000KHz
Camera Temp Setting	-25.0C
Camera Type	CoolSNAP_HQ2 / HQ2-ICX285
Column axis sequence	1
Created	Fri Dec 7 20:12:19 2007
Image 1. EM filter	POL
Image 1. EX filter	POL
Image 1. Exposure time	0.100000 secs
Image 1. Intensity	Min = 121, Max = 467, Mean = 202.45554
Image 1. ND filter	32%
Image 1. Photosensor reading	993.0
Image 1. Polychroic	Quad
Image 1. Stage coordinates	(-4100.890,-4163.630,+29.448)
Image 1. Time	Fri Dec 7 20:31:51 2007
Image 1. Time Point	0.000 secs
Image 10. EM filter	POL
Image 10. EX filter	POL
Image 10. Exposure time	0.100000 secs
Image 10. Intensity	Min = 120, Max = 479, Mean = 202.40558
Image 10. ND filter	32%
Image 10. Photosensor reading	1073.0
Image 10. Polychroic	Quad
Image 10. Stage coordinates	(-4100.890,-4163.630,+28.853)
Image 10. Time	Fri Dec 7 20:31:59 2007
Image 10. Time Point	7.896 secs
Image 100. EM filter	POL
Image 100. EX filter	POL
Image 100. Exposure time	0.100000 secs
Image 100. Intensity	Min = 110, Max = 422, Mean = 202.42722
Image 100. ND filter	32%
Image 100. Photosensor reading	1269.0
Image 100. Polychroic	Quad
Image 100. Stage coordinates	(-4100.890,-4163.630,+22.852)
Image 100. Time	Fri Dec 7 20:33:18 2007
Image 100. Time Point	86.824 secs
Image 101. EM filter	TRITC
Image 101. EX filter	TRITC
Image 101. Exposure time	0.400000 secs
Image 101. Intensity	Min = 101, Max = 2653, Mean = 272.74619

4.2.1 Many ways of describing data

- Notes added manually

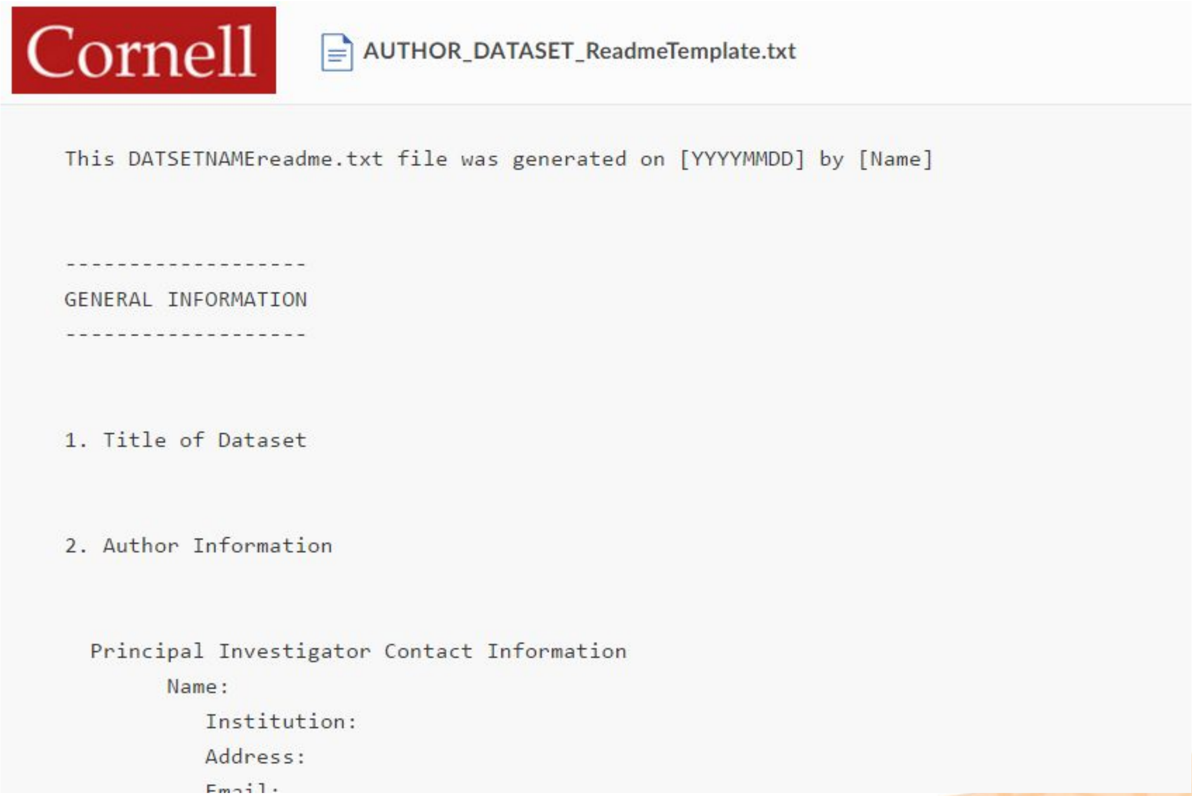


4.2.2 Many ways of describing data

- README files
  - Did you ever come across README files?
  - Have you ever created a README file?

4.2.3 Many ways of describing data

How to create useful README files : https://data.research.cornell.edu/content/readme



AUTHOR_DATASET_ReadmeTemplate.txt

```
This DATSETNAMEreadme.txt file was generated on [YYYYMMDD] by [Name]


-------------------
GENERAL INFORMATION
-------------------


1. Title of Dataset


2. Author Information


  Principal Investigator Contact Information
        Name:
           Institution:
           Address:
           Email:
```

README files template: https://cornell.app.box.com/v/ReadmeTemplate

4.3 How to know what’s best for my research?

4.4 Will you adhere to any discipline-specific metadata standards?

FAIRsharing.org
standards, databases, policies

https://fairsharing.org/standards/

5. Ethics and Legal compliance

Here you should consider any ethical or legal issues, particularly in terms of restrictions they may place on data sharing.

- Have you gained consent for data sharing and preservation? (Thank you Ms. Nicoletta Roz)
- How will you protect the identity of participants if required? e.g. via anonymisation
- Will data sharing be postponed / restricted? e.g. to publish or seek patents (Thank you Dr. Shiva Loccisano)
- How will the data be licensed for reuse?

PoliTo storage and computing services

Talk by Mr. Enrico Venuto, Area IT

Coffee break

Introduction to GDPR and PoliTo processes

Talk by Ms. Nicoletta Roz, DPO at PoliTo

Practical advice for managing and sharing personal data

- Relating GDPR concepts to Research Data
- Practical advice for management and sharing personal data

Thought exercise

What’s all this new stuff on privacy?

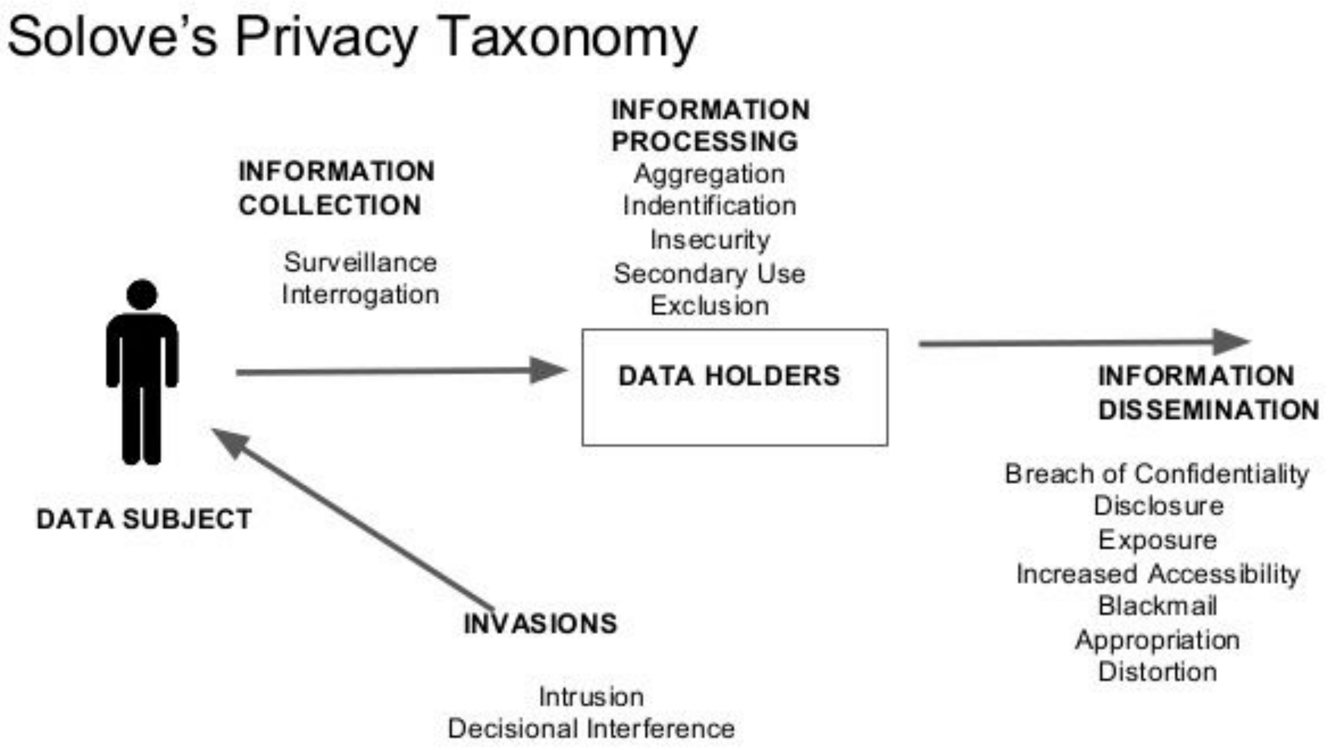
What is privacy

“Broadly speaking, privacy is the right to be let alone, or freedom from interference or intrusion. Information (in our case data) privacy is the right to have **some control** over how your personal information is collected and used.”

It is recognised as a human right in the UN declaration on human rights since 1948!

-IAPP, UN

Privacy breach taxonomy



Solove's Privacy Taxonomy

INFORMATION COLLECTION

Surveillance
Interrogation

INFORMATION PROCESSING

Aggregation
Indentification
Insecurity
Secondary Use
Exclusion

DATA HOLDERS

INFORMATION DISSEMINATION

Breach of Confidentiality
Disclosure
Exposure
Increased Accessibility
Blackmail
Appropriation
Distortion

DATA SUBJECT

INVASIONS

Intrusion
Decisional Interference

Solove, D. J. (2005). A taxonomy of privacy. *U. Pa. L. Rev.*, *154*, 477

Recap: What is personal data

Directive 95/46/EC definition of **Personal Data**

"any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity"

Recap: Personal data elements

- General: Name, Gender, Age, DOB, civil status, nationality, languages, IP addresses
- Organisational: work/home addresses, phone, e-mail, id number
- “Special”: Race, Religion, sexual orientation, health, sex life, criminal record, Biometric data
- Sometimes country specific categories

Thought exercise

Think of personal data elements related to research data based on your research topic

What does GDPR principles mean to me as a researcher: Short mapping

- Lawfulness, fairness and transparency: legal bases, ethics
- Purpose limitation: Research questions
- Data minimisation: What data should I collect?
- Accuracy: Data quality
- Storage limitation: How long?
- Integrity and confidentiality (security): Technical and policy measures
- Accountability: Who is responsible when unfortunate events happen

GDPR exemptions for Research

- Scientific or historical research purposes; or statistical purposes. Possible exemptions:
  a. the right of access;
  b. the right to rectification;
  c. the right to restrict processing; and
  d. the right to object.
- Archiving in the public interest: more exemptions

https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/exemptions/#ex18

GDPR for research: Special considerations

Further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes is not considered to be incompatible with the initial purposes

Appropriate safeguards, e.g.

- data minimisation
- pseudonymisation

Principles 2 and 5 less strict:

- Purpose: further processing of personal data allowed (2)
- Personal data may be stored for longer periods (5)

Most important consideration

Start early! Plan ahead

E.g. DMP

Most common legal base: Informed consent

- Consent needs to be **freely given**, **informed**, **unambiguous**, **specific** and by a **clear affirmative** action that signifies agreement to the processing of personal data

- Consent must be documented, e.g. consent form or audio-recorded verbal consent

Adapted from:
https://zenodo.org/record/1408108#.XGUQkTBKjIU, https://zenodo.org/record/1408579#.XGUP-TBKjIU

Information in consent form

- The contact details of the researcher, data controller and sometimes the Data Protection Officer
- Who will receive or have access to the personal data, information if the personal data is to be transferred outside the EU
- The right of the participant to request access to their personal data
- The period of retention for holding the data or the criteria used to determine this.
- Contact your HREC/DPO for further requirements

Adapted from Veerle Van den Eynden, Uk Data Service

https://zenodo.org/record/1408108#.XGUQkTBKjIU, https://zenodo.org/record/1408579#.XGUP-TBKjIU

Good practices to follow

- A statement that asks the participant to note their understanding of any procedures for handling any personal data collected (e.g. confidentiality, anonymisation, etc.);
- A statement or statements that asks the participant to consent to proposals for data sharing and re-use (whether in de-identified and/or identifiable form) for future research
- (If relevant and as appropriate) A statement that asks the participant to consent to the export of their personal data outside the EEA (e.g. to share it with another research institution or on an international database).
- (If relevant) A statement that asks the participant to consent to any planned audio or visual recording.

https://www.research-integrity.admin.cam.ac.uk/academic-research-involving-personal-data

Remember

Consent is not the only legal base for Research! It can sometimes be complex to manage

Consider others: Public interest etc. work with your Legal dept./DPO

Security of processing

- Confidentiality, integrity, resilience, availability
- Controller and process should have adequate tech and org. measures
- Specific requirements by certain states
- Policy, physical environment, IT, Incident detection and response

Cross-border transfers

- Adequacy decision: e.g. Canada, NZ, Australia etc
- Can be found in the EC website
- Appropriate safeguards: binding corporate rules, ad hoc contracts

De-identification

De-identification – refers to a process of removing or masking *direct identifiers* in personal data

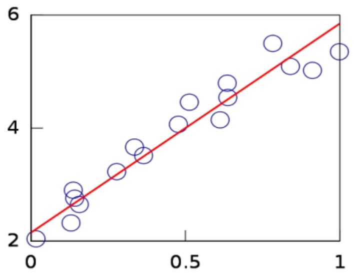
Anonymisation - refers to a process of ensuring that the risk of somebody being identified in the data is negligible. This invariably involves doing more than simply de-identifying the data, and often requires that data be further altered or masked. Anonymisation allows data to be shared ethically and legally while preserving confidentiality

But will it be useful for reuse?

Types of disclosure

Identity Disclosure – the intruder successfully associates an individual n with the released data – through direct identifiers, combinations of key identifying variables, linking with other available data

Attribute Disclosure – the intruder is able to determine unknown/sensitive features of an individual based on the info in the released data – high certainty



Inferential Disclosure - the intruder is able to determine more accurately sensitive features of an individual/organization with the use of the released data than it would be possible without

Adapted from Cristina Magder,
https://zenodo.org/record/1408579#.XGUP-TBKjIU

Some tools for de-identification

- Statistical disclosure control, k-anonymity to detect uniqueness
- **Amnesia**: Cloud based, easy to use https://amnesia.openaire.eu/
- **R tool - sdcMicro (scripting + GUI)**
  - R package (and free dependable software R and RStudio)
  - reference manual
  - new Shiny GUI – in detailed vignette
  - SDC methods and sdcMicro
- **µ-Argus**
  - standalone software recommended by Eurostat for government statisticians software and manual
- **ARX :** comprehensive open source software for anonymizing sensitive personal data
- software and documentation

Adapted from Cristina Magder,
https://zenodo.org/record/1408579#.XGUP-TBKjIU

Exercise to de-identify data: Home work

Find an open data set (e.g. Titanic dataset) and try out the tools.

De-identifying qualitative data

- plan or apply editing at time of transcription

  except: longitudinal studies – de identify when data collection complete (linkages)
- avoid blanking out; use pseudonyms or replacements
- avoid over-anonymising – removing / aggregating information in text can distort data, make them unusable, unreliable or misleading
- consistency within research team and throughout project
- show replacements, e.g. with [brackets]
- keep a log of all replacements, aggregations or removals made – keep separate from de-identified data files

https://zenodo.org/record/1408579#.XGUP-TBKjIU

Qualitative data considerations

- Before sharing data, check whether people can be identified from the data
- Check consent – if researching people, have they agreed to have their information shared for research?
- Regulate or restrict user access

Always consider risk vs. utility of anonymised data

Example

I was born in Philadelphia. My parents were both born and raised in Philadelphia. My father, Manuel Kaufman, was Jewish and my mother, Helen Carroll, was Irish Catholic. They both lived in South Philadelphia, on either side of Broad Street, and there was no chance that they would meet each other. Back in those days, and even when I was growing up, Philadelphia was a city of great ethnic divides, where the Italian, the Jewish, the Irish, the Polish, the black community, lived in their own neighborhood(s) with very little interaction.

They both went to the University of Pennsylvania, but didn’t meet there. They met later on. They were both working in public assistance as social workers when they got married. The biggest thing was that back in those days an Irish Catholic was not very welcome in a Jewish family, and a Jew was not very welcome in an Irish Catholic family, so it was interesting growing up with these two ethnic backgrounds.

At Penn, my mother was president of her sorority and was a big person on campus. Interesting point, at that point the Daily Pennsylvanian , even though women had been there for a number of years, never had a woman’s name in the newspaper. Even though they were students there, they were never mentioned. My mother went to John W. Hallahan Catholic Girls High School in South Philadelphia.

Example

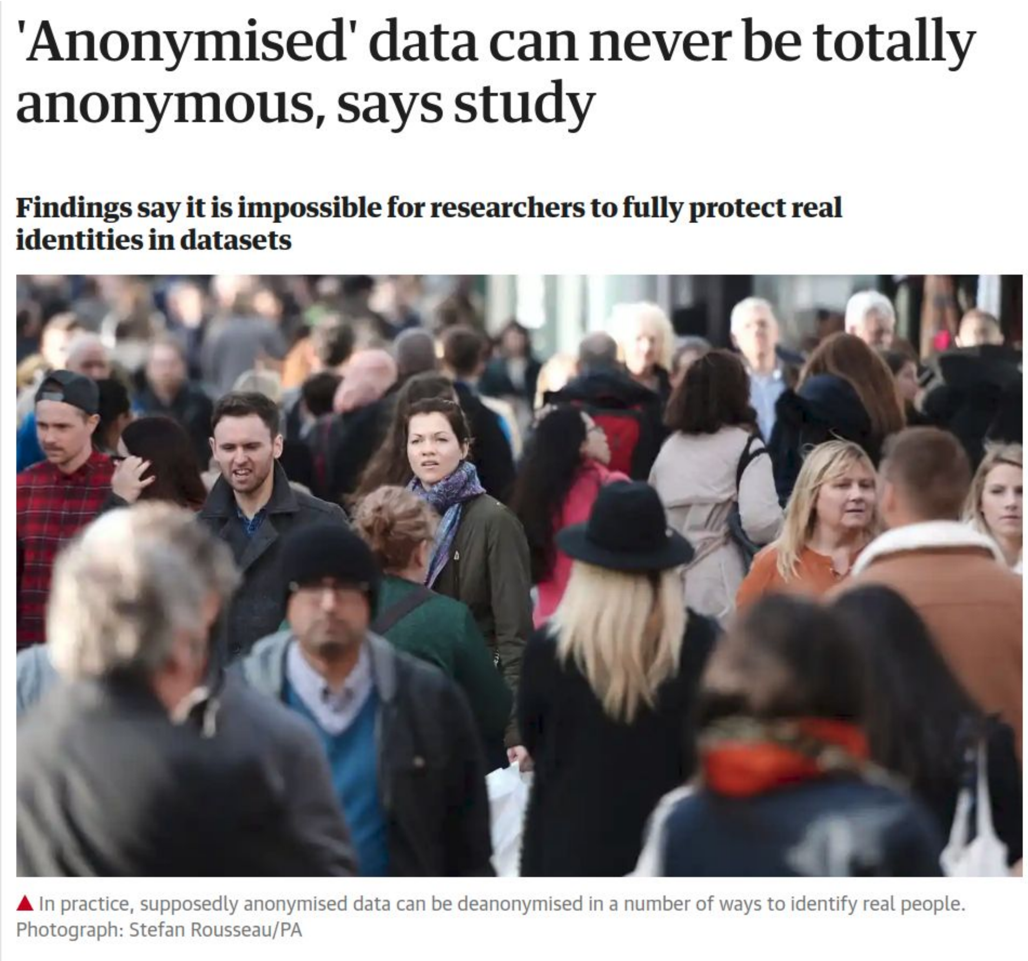
I was born in Philadelphia. My parents were both born and raised in Philadelphia. My father, **Manuel Kaufman**, was **Jewish** and my mother, **Helen Carroll,** was **Irish Catholic**. They both lived in **South Philadelphia, on either side of Broad Street**, and there was no chance that they would meet each other. Back in those days, and even when I was growing up, Philadelphia was a city of great ethnic divides, where the Italian, the Jewish, the Irish, the Polish, the black community, lived in their own neighborhood(s) with very little interaction.

They b**oth went to the University of Pennsylvania,** but didn't meet there. They met later on. They were both working in **public assistance as social workers** when they got married. The **biggest thing was that** back in those days an Irish Catholic was not very welcome in a Jewish family, and a Jew was not very welcome in an Irish Catholic family, so it was interesting growing up with these two ethnic backgrounds.

At Penn, my mother was **president of her sorority** and was a big person on campus. Interesting point, at that point the **Daily Pennsylvanian ,** even though women had been there for a number of years, never had a woman's name in the newspaper. Even though they were students there, they were never mentioned. My mother went to **John W. Hallahan Catholic Girls High School in South Philadelphia**.

https://www.senate.gov/artandhistory/history/resources/pdf/Kaufman_Oral_History.pdf

New technologies to consider



'Anonymised' data can never be totally anonymous, says study

Findings say it is impossible for researchers to fully protect real identities in datasets

▲ In practice, supposedly anonymised data can be deanonymised in a number of ways to identify real people. Photograph: Stefan Rousseau/PA

https://www.theguardian.com/technology/2019/jul/23/anonymised-data-never-be-anonymous-enough-study-finds

Differential Privacy

Federated Learning

Case studies: Open Discussion

Autonomous driving

AI and machine learning.. healthcare applications…

Social media data

Bio med, genetics engineering

Remember

It is all about proportionality

A comprehensive approach: A good example

Safe data - treat data to protect confidentiality

Safe people - educate researchers to use data safely

Safe projects - research projects for ‘public good’

Safe settings – Secure system for sensitive data

Safe outputs – Secure projects outputs screened

https://www.youtube.com/embed/Mln9T52mwj0, UK data service

Finally

- Don’t be scared of GDPR :) It’s about proportionality, ethics and common sense
- Investigate early which aspects apply to your data
- Seek advice from you research office/DPO if in doubt
- If you must collect / handle personal or sensitive data:
  - be transparent about processing personal data
  - follow best practices for processing

IPR and Technology Transfer considerations

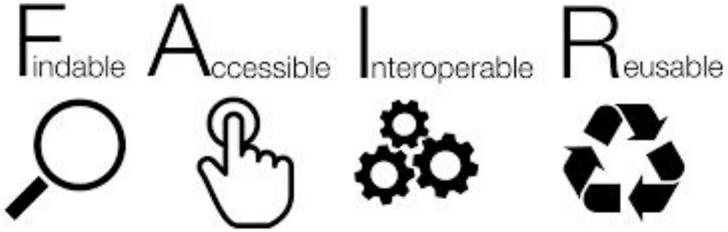
Talk by Dr. Shiva Loccisano, head of TRIN

Wrap-up and intro to the Feb 20 session

Today we covered

- Open Science rationale and role of RDM: the why and what
- Funder and journal requirements
- Introduction to DMPs
- PoliTo infrastructure and the various stakeholders related to RDM
- On Feb 20: From theory to practice

On Feb 20, more on FAIR data



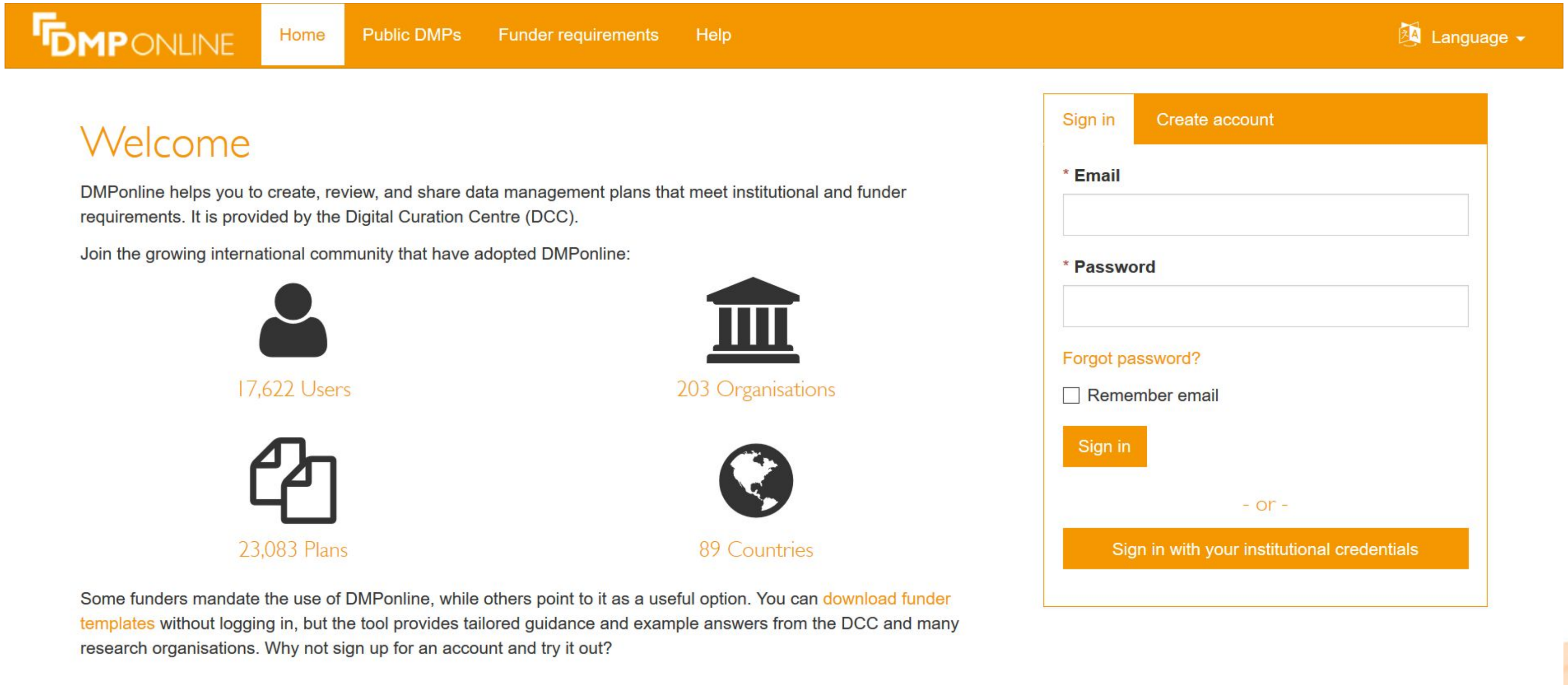
On Feb 20: The remaining DMP checklist

1. Administrative Data
2. Data Collection & Organisation
3. Storage and Backup
4. Documentation and Metadata
5. Ethics and Legal compliance
6. Selection and Preservation
7. Data Sharing
8. Responsibilities and Resources

We will cover these topics
on Feb 20 with a short
recap of the above

Writing a Data Management Plan

Using dmponline, please bring your laptop and charger!



Welcome

DMPonline helps you to create, review, and share data management plans that meet institutional and funder requirements. It is provided by the Digital Curation Centre (DCC).

Join the growing international community that have adopted DMPonline:

17,622 Users

203 Organisations

23,083 Plans

89 Countries

Some funders mandate the use of DMPonline, while others point to it as a useful option. You can download funder templates without logging in, but the tool provides tailored guidance and example answers from the DCC and many research organisations. Why not sign up for an account and try it out?

Sign in | Create account

* Email

* Password

Forgot password?

□ Remember email

Sign in

- or -

Sign in with your institutional credentials

Thank you! Contact us for OS questions

Contacts

Policy OA:
copyright@polito.it, open.science@polito.it

Gruppo di Lavoro OA:
Maria Girard – Monica Margara copyright@polito.it

RDM:
shalini.kurapati@polito.it, open.science@polito.it

Obblighi OA in H2020:
Area Ricerca, ari@polito.it