

# Data sharing

Dr Cyril Pernet, PhD

11 December 2023

<https://github.com/CPernet/ReproducibleQuantitativeDataScience>

## As usual ...

- This is required that you engage with the lecture; group exercises for the class are in the

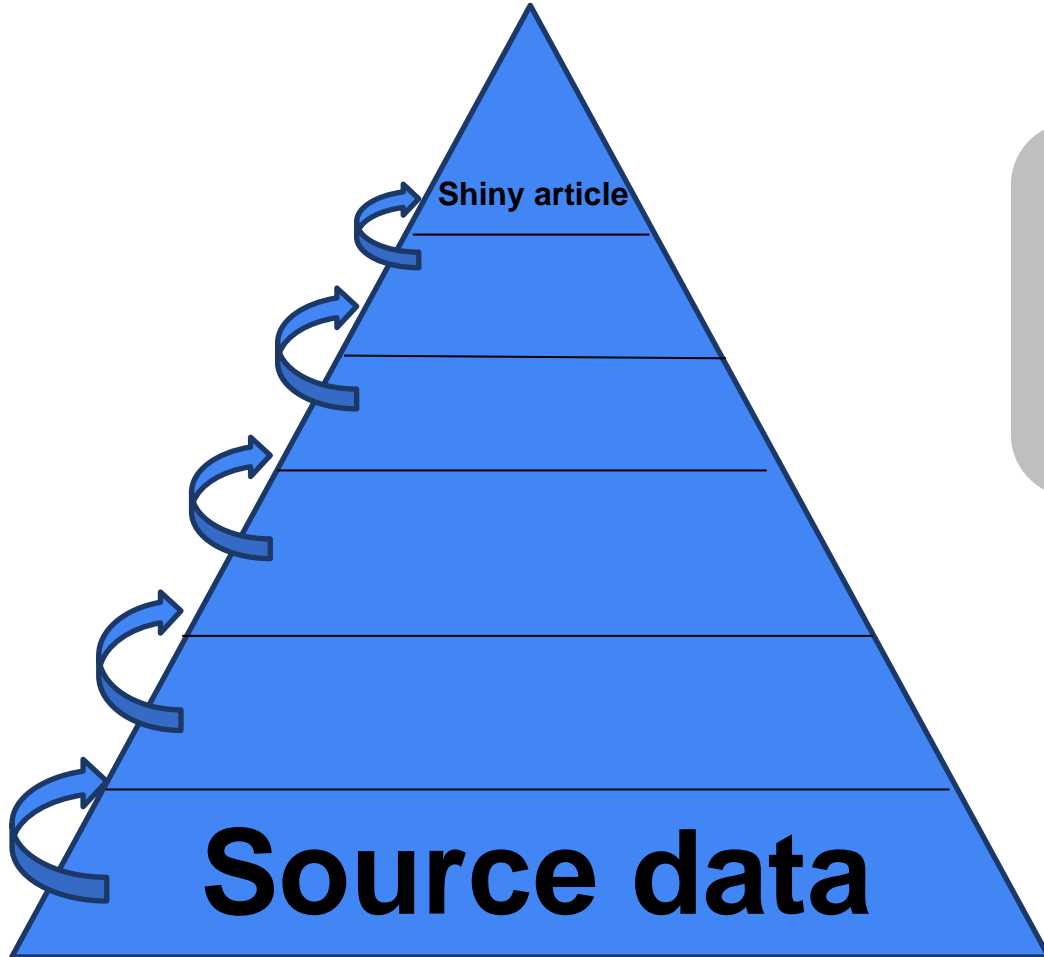
**Question boxes**

- Of course, stop me at any time if you have questions

# Scientific Data Management

- What are scientific data?
- For whom is that useful?
- How to share?

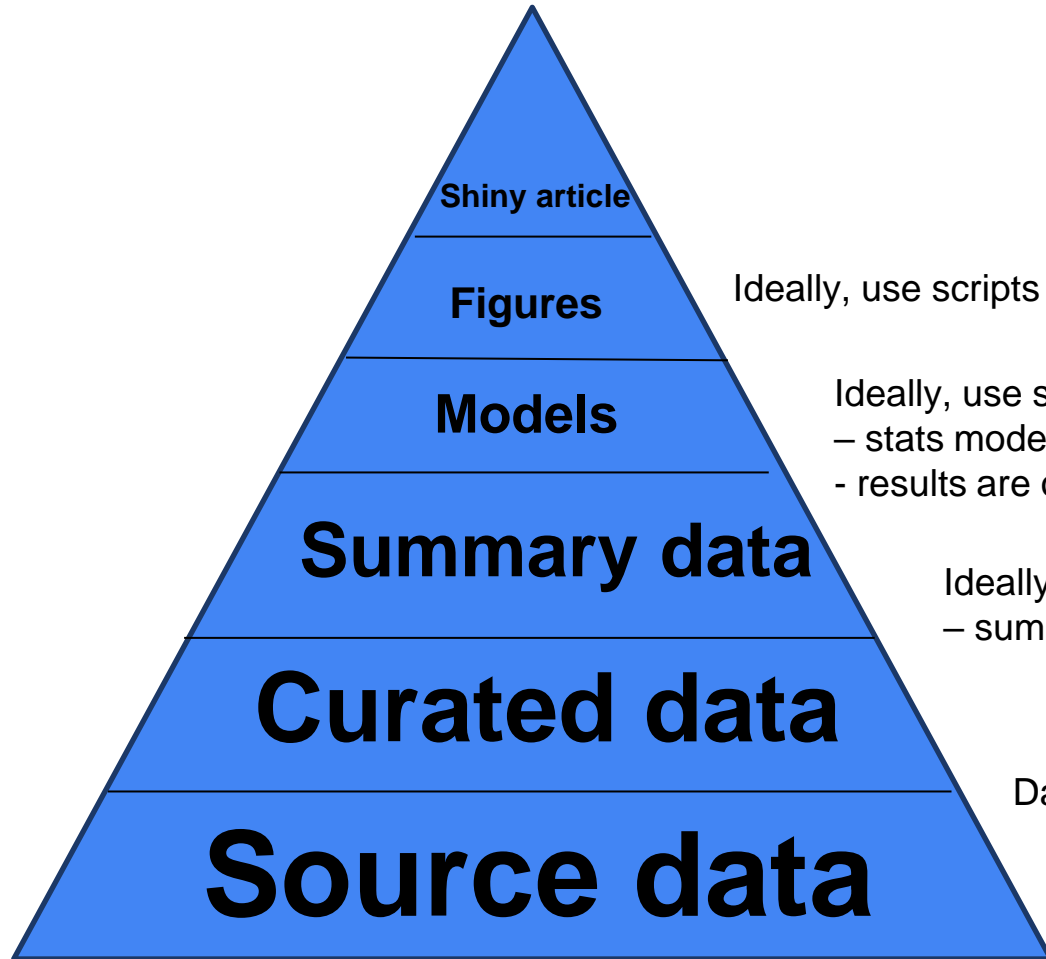




### What data?

- Levels/types of data
- how it is made
- usefulness to share

(make groups – let's draw on the board)



Ideally, use scripts (often post edited)

Ideally, use scripts

- stats model and data models

- results are often presented as tables and figures

Ideally, use scripts

- summary data are often presented as tables

Data curation – manual/scripts

# Scientific Data Management

- Experimental Data are organized in a consistent way
  - i) for you in 6 months
  - ii) for your collaborators (and PI)
  - iii) for easy data sharing
- Analysis data (code and results) are stored and shared allowing reproducibility using the well-organized experimental data (lecture 1.02)
- Documentation is systematic being in the procedures (lecture 1.03) or the analysis (lectures 2.02, 2.03)
- Open science encompasses a variety of tools and practices that allow good data management and thereby reproducible science

# **Data Management Plan**

# Definition

Google Data Management Plan

About 2,520,000,000 results (0,53 seconds) « Add Grepper Answer (a) | Add Writeup

A data management plan (DMP) is a written document that describes the data you expect to acquire or generate during the course of a research project, how you will manage, describe, analyze, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data.

<https://library.stanford.edu/data-management-services> :  
**Data management plans | Stanford Libraries**

People also ask :


How do you write a good data management plan? ▾


What is data management examples? ▾

Why do you need a data management plan? ▴

By laying out the blueprint for lifecycle management of data, a DMP provides valuable details, such as how the data will be preserved for the long term, how and where the researcher will make the data available for sharing, and whether reuse of the data, including derivatives, will be allowed.

<https://www.e-education.psu.edu/dmpt/node>  
**Why Do You Need a Data Management Plan?**





**Data management plan**

A data management plan or DMP is a formal document that outlines how data are to be handled both during a research project, and after the project is completed. [Wikipedia](#)

**Purpose** ▾

**Importance** ▴

Importance of a Data Management Plan

A data management plan helps you: **Increase impact and visibility of your research with data citation.** document and provide evidence for your research in conjunction with published results. comply with funding mandates, and meet copyright and ethical compliance requirements. 27 Sept 2021

Acquired data  
Generated data  
Management

- Analyze
- Store
- Share
- Preserve



# The basics

## 1) Data

- what data will be collected (type and size)
- what type of analyses will be done
- what derived data will be generated (type and size) vs. need to be archived (might be just summary statistics as csv, new images, a mathematical model)

2) How will you document the data? i.e. make metadata describing what this is, how it was collected, analysed and generated - is there a data curation standard?

3) What will you share and how?

4) How will the data be preserved?

# Definition

Google Data Management Plan

About 2,520,000,000 results (0,53 seconds) « Add Grepper Answer (a) | Add Writeup

A data management plan (DMP) is a written document that describes the data you expect to acquire or generate during the course of a research project, how you will manage, describe, analyze, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data.

<https://library.stanford.edu/data-management-services> : Data management plans | Stanford Libraries

People also ask :


How do you write a good data management plan?


What is data management examples?

Why do you need a data management plan?

By laying out the blueprint for lifecycle management of data, a DMP provides valuable details, such as how the data will be preserved for the long term, how and where the researcher will make the data available for sharing, and whether reuse of the data, including derivatives, will be allowed.

<https://www.e-education.psu.edu/dmpt/node> Why Do You Need a Data Management Plan?





**Data management plan**

A data management plan or DMP is a formal document that outlines how data are to be handled both during a research project, and after the project is completed. [Wikipedia](#)

**Purpose**

**Importance**

Importance of a Data Management Plan

A data management plan helps you: **Increase impact and visibility of your research with data citation.** document and provide evidence for your research in compliance requirements.

Acquired data  
Generated data  
Management

- Analyze
- Store
- Share
- Preserve

# Don't ask Why

Yes, it is required by funders, admin, etc ... but see this as **an opportunity**

- 1 - to request/ask for enough computing power and space
- 2 - improve your science
- 3 - make science better by making your data FAIR



# FAIR Principles

F  
A  
I  
R



## The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons  - [Show fewer authors](#)

# FAIR Principles

Findable  
Accessible  
Interoperable  
Reusable

## The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons  - [Show fewer authors](#)

# **FAIR DMP**

Findable and Accessible

**How do you achieve that?**

# FAIR DMP

Findable and Accessible = must be in an indexed repository (your university library? Other repository?)

- Preferably searchable
- Preferably with keywords and other metadata
- Open Access make things easier
- Access controlled is fine too - but by who?  
Under which conditions access is granted?

**You must have  
your data open  
access to be FAIR**



# GDPR Reminder

- When processing Personal Identifiable Information – Art 4(1): data relating to a person leading to direct or indirect identification for instance using physical, physiological, genetic, mental, economic, cultural or social data.

- Animal studies??
- Dead individuals??



# GDPR Reminder

- When processing Personal Identifiable Information – Art 4(1): data relating to a person leading to direct or indirect identification for instance using physical, physiological, genetic, mental, economic, cultural or social data.
- Animal studies, studies from dead individuals, fully anonymous tissue data are all out of scope (no GDPR = no excuse not to share).
- Data from living humans, with no associated metadata (e.g. demographics and health data) and no ID (that is the data have a random ID and you can never go back to figure who this is from) are anonymous.
- Anything else, including pseudonymized data (the *process* of removing identifiers) should be seen as PII.

# Infrastructure Requirements

- To be FAIR, sharing PII requires consent to share but also an infrastructure allowing public data sharing (metadata are findable ≠ open data) with access control to identified users (accessible) using legal agreement(s) such as a DUA and SCC of non EU users.
- A DMP doesn't have to be FAIR, and many cases you will not have the means to make PII easily findable (need a platform). Using data publication and a dedicated repository you can make data accessible and more easily findable.
- Identify if that is possible early on! Check with your library/institution what is planned, this is their job to provide infrastructure to support research/funders requirement.

# FAIR DMP

Interoperability = commonly used and open data formats

Reusable = metadata using international standards and ontologies



Your file format, structure, organisation, naming and documentation (if any) might not be understandable to anyone (lectures 1.02 and 1.03)

# Resources

<https://dmponline.dcc.ac.uk/> (there are some country specific versions, google it)

<https://www.openaire.eu/how-to-create-a-data-management-plan>

<https://library.stanford.edu/research/data-management-services/data-management-plans>

<https://howtofair.dk/how-to-fair/metadata/>

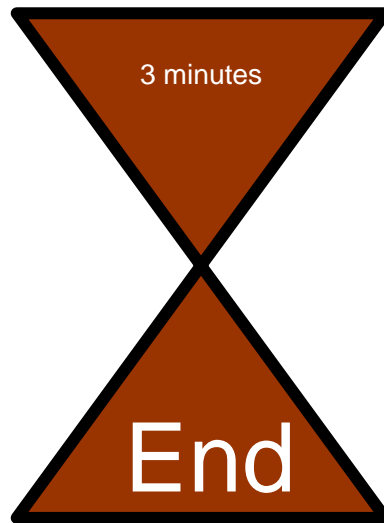
# Licensing

# What licence?

- Licence versus patent (creator/user agreement vs. prevent usage)
- What do you expect from a licence?

Let see if we can find  
something that match your  
expectations:

<https://choosealicense.com/>



# Some Caveats

- Give rights to *use, copy, modify*, and possibly *redistribute*
- MIT (most permissive), Apache, BSD (grant patent to user), GNU GPL (enforce reuse to also be open, not quite like CCNC)
- Your institution will have its own policies and procedures that lay out how to obtain permission to open source your software, you cannot decide – but funders requirements make the difference
- The reuse, bundling or redistribution of code snippets, libraries and other assets such as images as part of your software can only be done if the licences are *compatible*.
- Biomedical data are ‘personal’ = no open licence but Data User Agreement

<https://www.software.ac.uk/resources/guides/adopting-open-source-licence>

# Publications

Just a word, before the next lecture on that





## Conventional

Subscription-based/  
Pay wall Protected

Reader-pay model

No free access

© transfer agreement



GREEN

## Open Access

Self-archiving after  
embargo

Hybrid model

Free access embargoed

© transfer agreement



GOLD

## Open Access

Immediate Availability

Author-pay model

Free access at all times

CC license agreement

+ updated preprint  
= gold for way cheaper



By Anke van Eekelen

# Publish and own your figures !! Example:



NeuroImage



Volume 170, 15 April 2018, Pages 348-364



## A critical analysis of neuroanatomical software protocols reveals clinically relevant differences in parcellation schemes

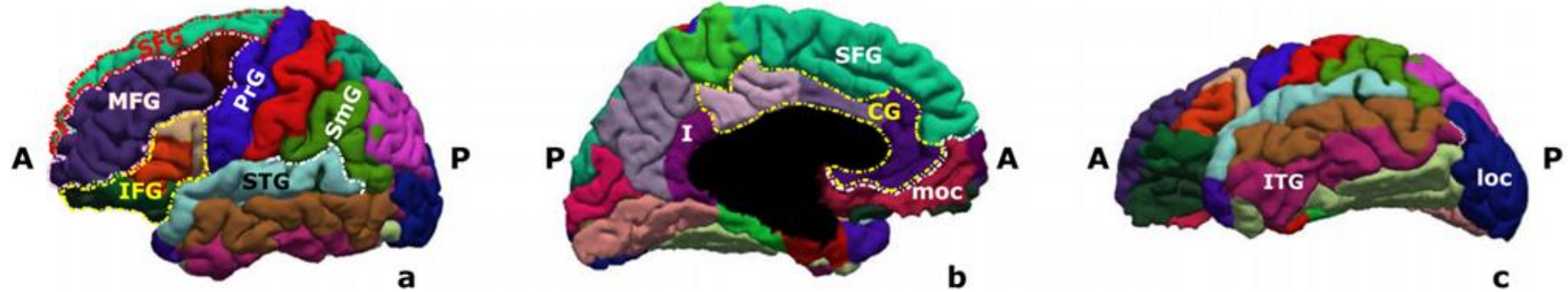
Shadia Mikhael<sup>a</sup>  , Corné Hoogendoorn<sup>b</sup>, Maria Valdes-Hernandez<sup>a</sup>, Cyril Pernet<sup>a</sup>

Show more 

+ Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.neuroimage.2017.02.082> 

[Get rights and content](#) 



**Fig. 4.** Lateral (a), medial (b), and inferior (c) cortical surfaces of a 30–35 year-old right-handed male subject, as per FreeSurfer's Desikan-Killiany parcellation protocol. Border precision lacked for the ROIs we investigated, particularly at (1) the PrG's medial border, (2) the SFG and CG's anterior border, (3) the SmG-STG border, (4) the CG-isthmus border, and (5) the ITG-loc border. The raw volume was downloaded from <http://psydata.ovgu.de/studyforrest/structural/sub-01/>. I: isthmus, moc: medial orbitofrontal cortex, loc: lateral occipital cortex (Mikhael, 2017) .

**Publish your figures before review (and update as needed)**  
**- get an open licence and DOI (ideally have csv or related data associated) and include DOIs of your figures in the article**

**The figures belong to you! An open licence like CCBY says the shared object can be used for profit, so the publisher does not mind using it, but now you can also reuse your figures in other articles, books, for on-line teaching, etc. +**



**INFORMATION SERVICES**





Edinburgh DataShare / College of Medicine & Veterinary Medicine / Edinburgh Medical School / Clinical Sciences / Edinburgh / View Item

**Examples of Automated Cortical Parcellation**

No Thumbnail	<p><b>Citation</b>  Mikhael, Shadia. (2017). Examples of Automated Cortical Parcellation. [image]  University of Edinburgh. Centre for Clinical Brain Sciences.  http://dx.doi.org/10.7488/doi.1963.</p> <p><b>Description</b>  The figures presented show examples of anatomical variability and the resulting automated parcellation from different software (also included a diagram of our search for such software). Software packages for cortical parcellation generally combine tools which segment the brain into various regions, or volumes, and provide corresponding measurements, or morphometrics. The results are used for volumetric, region-of-interest (ROI) and connectivity analysis, offering a better understanding of these volumes in the population(s) under investigation. It is therefore crucial to understand how these regions are defined, or the segmentation protocol of a package, as this has significant implications on the analysis that follows. These images reflect the variability in anatomy and in the way software handle this.</p> <p><b>Date Available</b>  2017-03-09</p> <p><b>Type</b>  image</p> <p><b>Data Creator</b>  Mikhael, Shadia</p> <p><b>Publisher</b>  University of Edinburgh. Centre for Clinical Brain Sciences</p> <p><b>Relation (is Referenced By)</b>  http://dx.doi.org/10.1016/j.neuroimage.2017.03.039; agnemo_jmri (17.00KB)</p> <p><b>Metadata</b>  Show full item record</p> <p><b>Download all files</b> (17.00KB)</p> <ul style="list-style-type: none"> <li>Cortical parcels as seen in the coronal view (97.75KB)</li> <li>Interpretation of the MarsAtlas for 5 regions of interest (ROIs) (1.549MB)</li> <li>Cortical surfaces parcellated by FreeSurfer (3.152MB)</li> <li>Flow diagram identifying the software packages of interest (2.062MB)</li> <li>An example of a double cingulate gyrus and sulcus and its interpretation by various software packages (11.81MB)</li> <li>Examples of sulcal patterns in a middle-aged male (1.234MB)</li> <li>README file (1.769KB)</li> </ul>
--------------	--

# Christmas share

F<sub>indable</sub> A<sub>ccessible</sub> I<sub>nteroperable</sub> R<sub>eusable</sub>

## License Types

<b>Copyleft</b> "GPL"	<b>permissive</b> MIT, BSD, ...	<b>Creative Commons</b>	<b>Public Domain</b>
viral	attribution	building blocks;	"Don't care"
attribution	protection against liability	→ attribution → virality → no commercial use → no derivatives	liability? 🟡
protection against liability	patents, trademark rules in details		

## Self-archiving!

## Code Sharing (GitHub)

## Data Repositories

