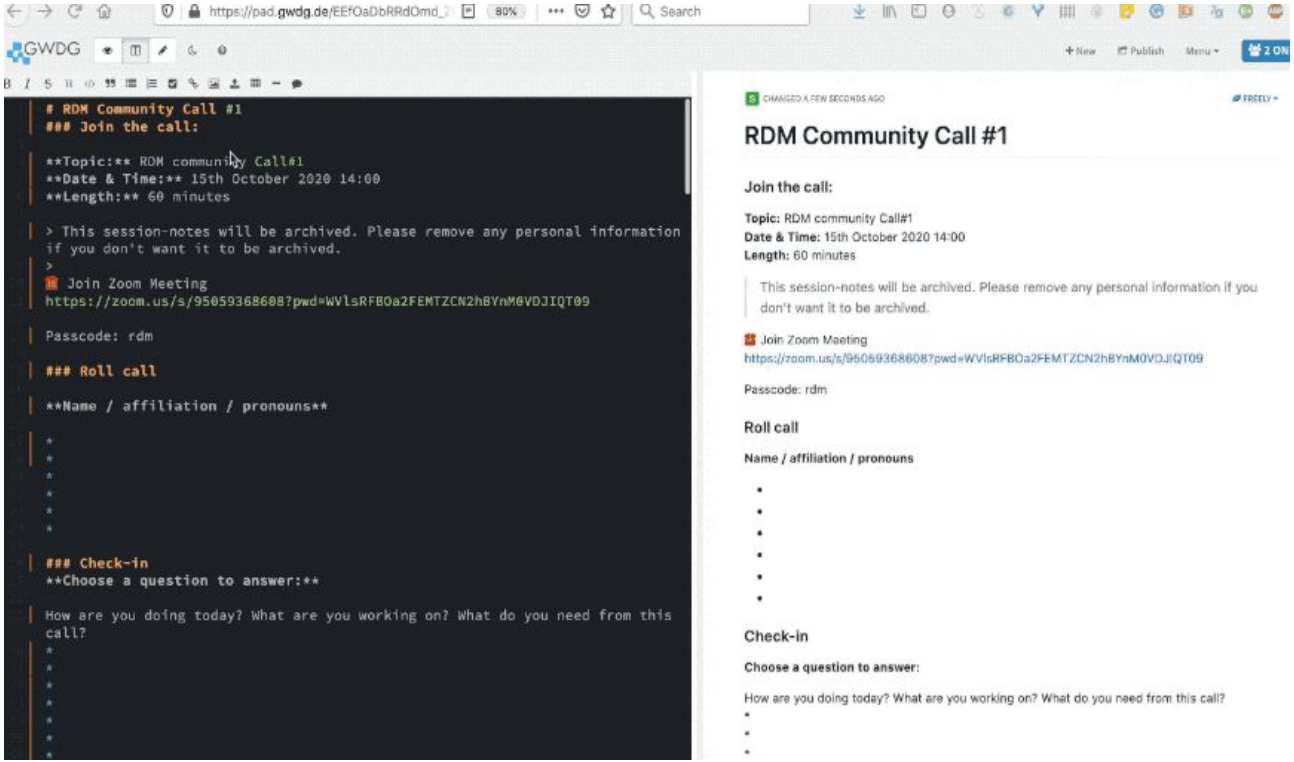
Research Data Management

Dr. Sara El-Gebali
https://orcid.org/0000-0003-1378-5495

Housekeeping

- Stay on mute! 🔇
- Raise your /hand ✋
- Notes are for you ✍️
- Code of conduct 📖
- Welcome!! 🎉



Round table introductions

🗣️ Name

□ Role

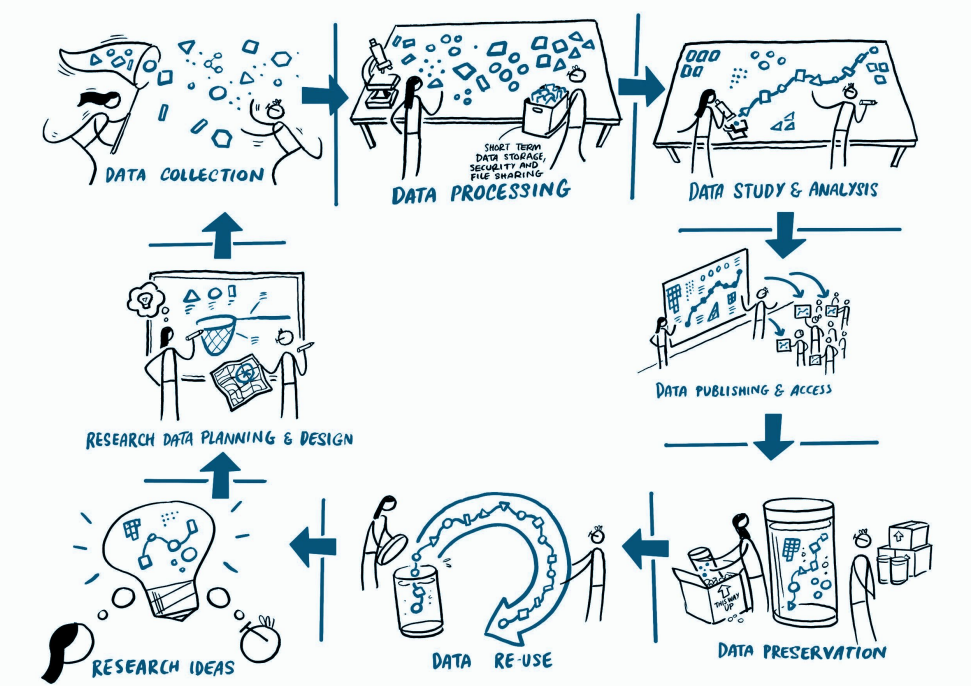
❓ Affiliation

🔎 What are you hoping to get out of this session?

Part 1: Overview of the RDM unit services and support

Our vision is to support researchers through all phases of the **research data life cycle** (Planning, Data collection, Management and Analysis, Preservation and Sharing).

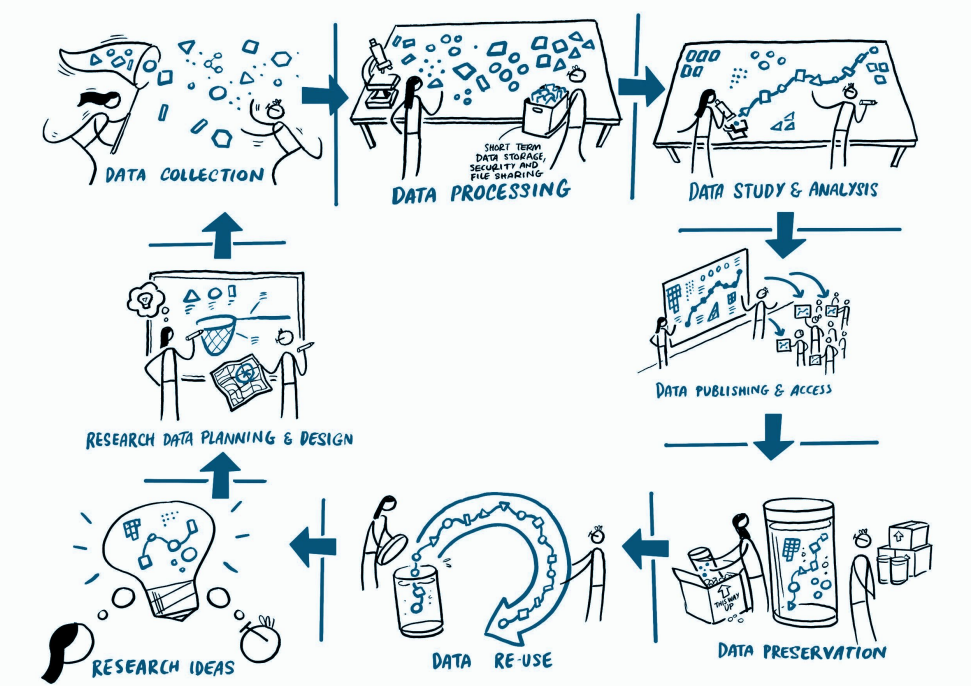
It involves the everyday management of research data during the lifetime of a research project and to preserve and share it beyond the project completion.



The Turing Way Community, & Scriberia. (2020, March 3). Illustrations from the Turing Way book dashes. Zenodo. http://doi.org/10.5281/zenodo.3695300

Our Role

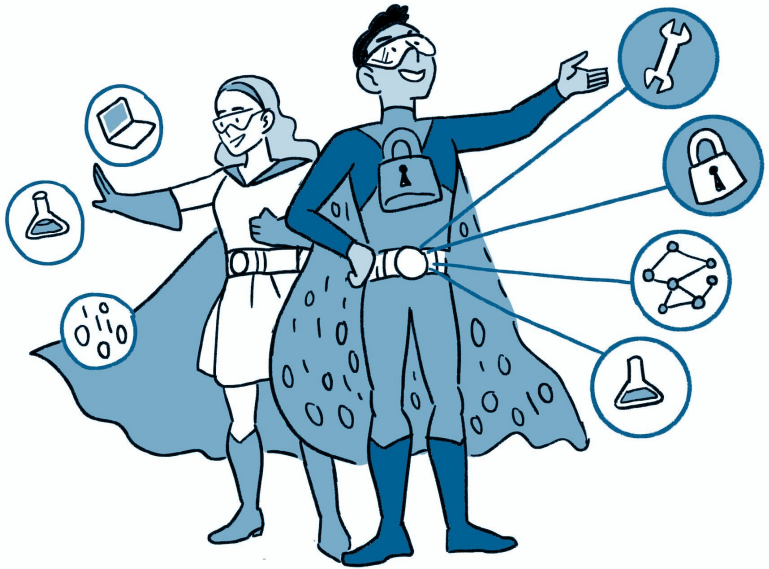
- To offer support and guidance during the different research phases.



The Turing Way Community, & Scriberia. (2020, March 3). Illustrations from the Turing Way book dashes. Zenodo. http://doi.org/10.5281/zenodo.3695300

Our Role

- Interface between policy makers, researchers and IT specialists
- To provide practical approaches to support you in making your data more FAIR and Open.



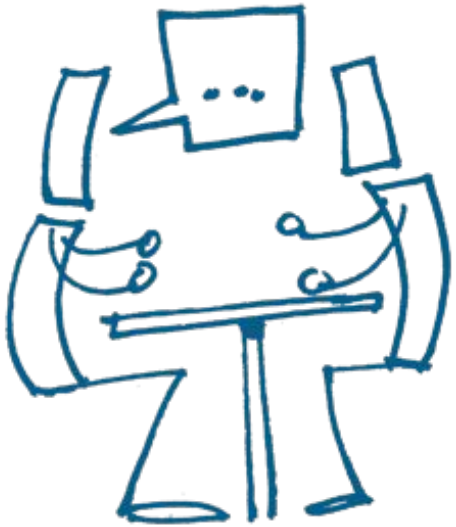
Scriberia

The Turing Way Community, & Scriberia. (2020, March 3). Illustrations from the Turing Way book dashes. Zenodo. http://doi.org/10.5281/zenodo.3695300

Our Role

Planning

- Data management planning (DMPs)
- Data description and metadata extraction
- Data documentation
- Choice of repositories
- Choices of file formats
- File naming
- Data re-use
- Funders requirements
- Ethics and Research conduct
- Funding for RDM activities



https://doi.org/10.5281/zenodo.1212496

Our role

Managing

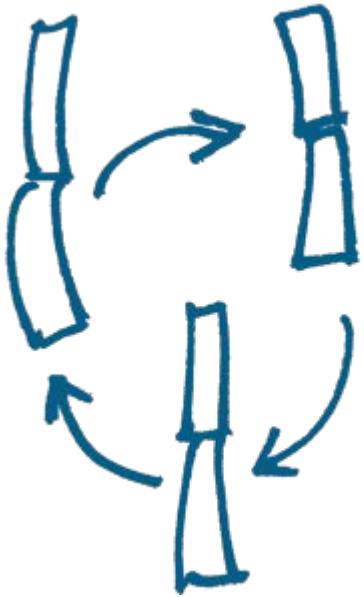
- Storage and backup & security
- Tools and software solutions
- Active Metadata collection
- Curation
- Versioning
- Provenance

https://doi.org/10.5281/zenodo.1212496

Our role

Sharing

- Data access and Sharing rights
- Data privacy and GDPR compliance
- Data ownership, licensing
- Data Transfer

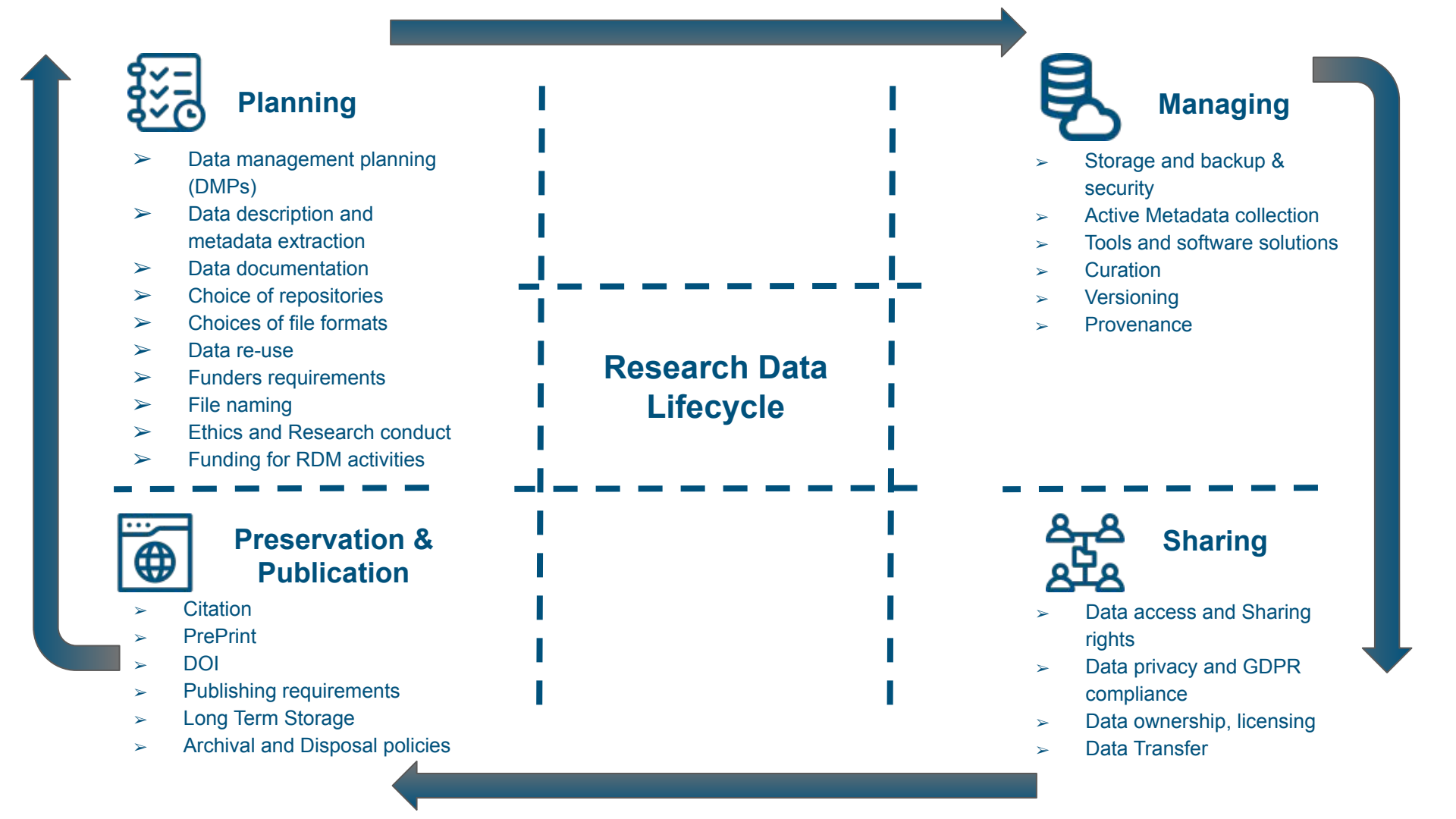


Our role

Publication

- Publishing requirements
- Citation
- PrePrint
- DOI
- Long Term Storage
- Archival and Disposal policies

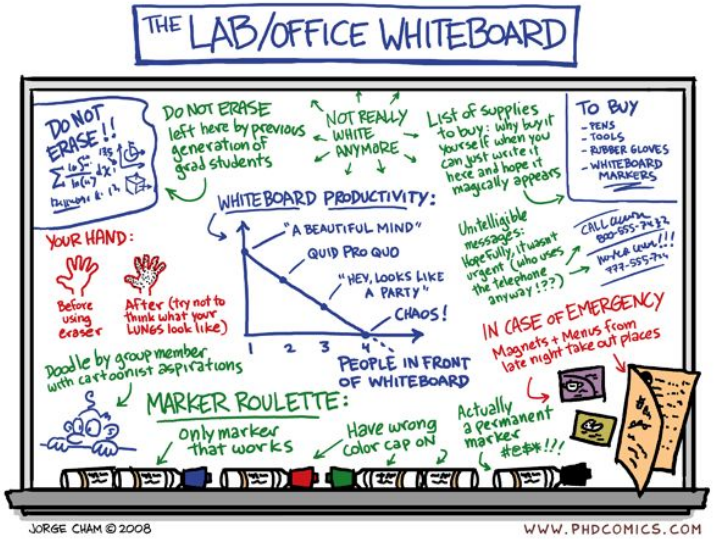
https://doi.org/10.5281/zenodo.1212496


Planning
Data management planning (DMPs)
Data description and metadata extraction
Data documentation
Choice of repositories
Choices of file formats
Data re-use
Funders requirements
File naming
Ethics and Research conduct
Funding for RDM activities
Managing
Storage and backup & security
Active Metadata collection
Tools and software solutions
Curation
Versioning
Provenance
Research Data Lifecycle
Preservation & Publication
Citation
PrePrint
DOI
Publishing requirements
Long Term Storage
Archival and Disposal policies
Sharing
Data access and Sharing rights
Data privacy and GDPR compliance
Data ownership, licensing
Data Transfer

Part 2: What is Data?

Any type of information that is collected, observed, or created, in the context of research, as such, data can be;

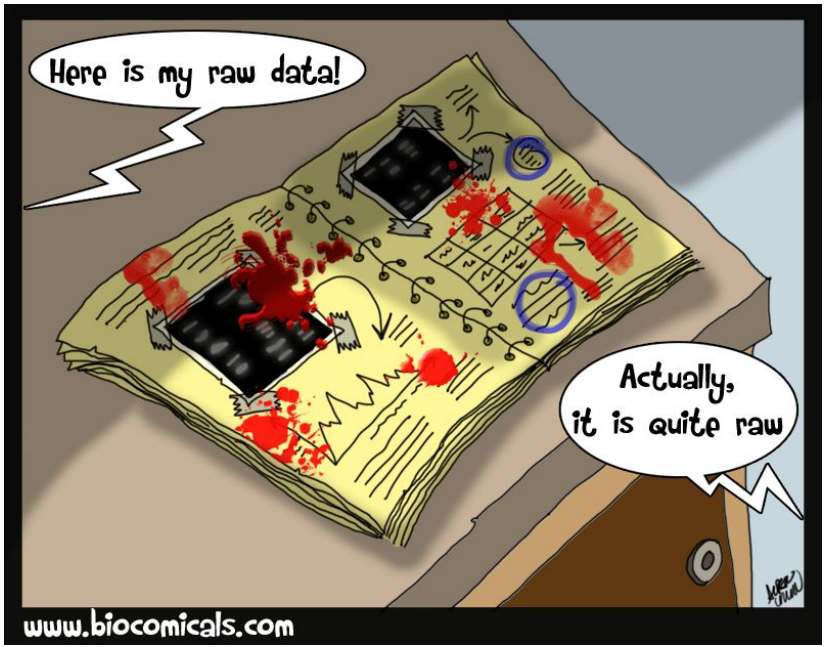
- **Primary-** Raw from measurements or instruments
- **Secondary-** Processed from secondary analysis and interpretations.
- **Published-** final format available for use and reuse
- **Metadata-** data about your data



- It is everything that you need to validate or reproduce your research findings as well as what is required for the understanding and handling of the data.

Primary Data

- Primary data is data collected or observed directly at hand such as those derived from the source e.g.instruments reading.
- Includes raw data for example recordings in lab notebooks



Secondary Data

- Processed files and interpretations,
- Data derived from other sources such as repositories
- Data that was collected for a purpose other than the current one e.g. data from a previous experiment or form another researcher.

This distinction is based on difference in the purpose of the data collection and consequently also on its relation to the research process: Whereas primary data are collected with the purpose to contribute to solving a specific research problem, secondary data are collected with different research purposes than the one for which they are initially used.[21] The researcher using secondary data "by an act of *abstraction* uses questions originally employed to indicate one entity to illuminate other aspects that a former analyst did not have in mind at all." (Hyman, 1972, p. 37)

https://www.liberquarterly.eu/article/10.18352/lq.9173/

Published data

- Journals
- Books
- Conference talks/ posters
- Blogs

https://doi.org/10.5281/zenodo.1212496

Metadata

Independent data that contain structured information about other data, i.e. Data about data.

- It ensures that your data is more reliable and accessible by providing details describing to others what to expect and how they can use this data and under what conditions.
- Increases the value of your data by making it more reusable
- Enhances discoverability of your data, your citation rate and reputation
- Reduces duplication efforts
- Metadata allows us to track people,institutions or publications associated with the original research, which can be very helpful when the original data is no longer available.
- Enables researchers to quickly assess the quality and relevance of the dataset to their research.

Give one example of primary, secondary, published data (if available) and metadata from your work, indicating file formats.

Part 3: FAIR data

- FAIR is a set of principles to define the best practices for data and software to facilitate discovery, access and reuse by humans and machines.
- FAIR is not rules and not a standard, it is an evolving process and a vision.

What does FAIR stand for?

Findable, Accessible, Interoperable and Reusable.

https://doi.org/10.5281/zenodo.1212496

Findable

Your data should be findable, by you and others. What does that mean? It means your data should be available in a discoverable resource, have appropriate description (i.e. metadata), have a persistent identifier.

How to:

- Data and metadata should have a persistent identifier (a stable address where to find it), URL is not a PID.
- Whenever possible, deposit your data to a domain-specific repository related to your field, https://www.re3data.org/
- If that is unavailable, deposit your data in general-purpose repositories such as Zenodo, Dryad, Dataverse.
- Same goes for your metadata.

https://doi.org/10.5281/zenodo.1212496

Accessible

Your data should be accessible for both humans and machines, i.e. retrievable and understandable

How to:

- Deposit the data under well defined conditions for others to be able to access it, i.e. data is accessible at HTTP or public REST API.
- Add clear licenses describing who is allowed to access this data and what they are allowed to do with it.
- Specify what the users need to do to access this data, ideally, a machine can automatically translate those requirements and act on it, i.e. two factor authentication, request access from author, etc…
- For private and sensitive data, the metadata (information about the data) can be made available and accessible.

https://doi.org/10.5281/zenodo.1212496

Remember

No license = No access!

‘As open as possible, as closed as necessary’

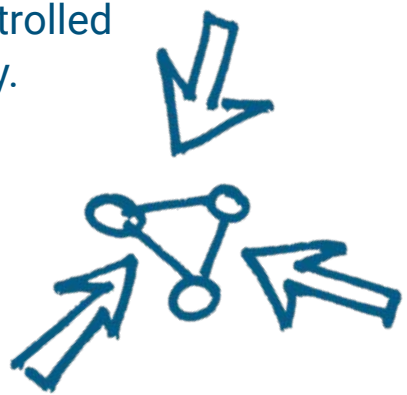
Even heavily protected and private data can be FAIR.

Interoperable

Machines and humans can interpret and use the data in different settings.

How to:

- Describe your data in detail - be generous!
- Describe your data properly, use controlled vocabulary, ontologies (controlled vocabulary with hierarchical relationships) and standardise terminology.
- Use preferred file formats, and open whenever possible.



https://doi.org/10.5281/zenodo.1212496

Reusable

The ultimate goal of FAIR is to advance the reuse of data. Everything you've done so far ultimately leads to this point, ensuring the data can be reused by others.

How to:

→ We will delve deeper into that, over the course of the day!

https://doi.org/10.5281/zenodo.1212496

Summary

1. Deposit your data where others can find it, keep in mind where your peers can find it, i.e. field specific repository and give it a stable unique identifier (PID).
2. Make your data & metadata accessible via standard means such as http/API.
3. Create metadata and explain in detail what this data is about, never assume people know!
4. Deposit metadata with PID and make it available with/out data i.e. in case data itself is heavily protected.
5. Include information on ownership and provenance.
6. Outline what the reusers of your data are/not allowed to do, use clear license. Commonly used licenses like MIT or Creative Commons (keep in mind funders requirements).
7. Specify access conditions, if authentication or authorization is required.
8. Describe your data in a standardized fashion using agreed terminology and vocabulary.
9. Share the data in preferred & open file formats.
10. Start the process early on!

CARE: CARE Principles for Indigenous Data Governance

“ The CARE Principles for Indigenous Data Governance are people and purpose-oriented, reflecting the crucial role of data in advancing Indigenous innovation and self-determination. These principles complement the existing FAIR principles encouraging open and other data movements to consider both people and purpose in their advocacy and pursuits.”

https://www.gida-global.org/care

Questions?

Part 4: Open Data

Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike.

The Open Data Handbook- Open Knowledge Foundation

Part 4: Open Data

FAIR vs Open

“FAIR means thinking about the people who could benefit from your data,”

“When we’re talking about open data, we’re generally referring to data that can be downloaded freely from the internet.”

"A love letter to your future self": What scientists need to know about FAIR data

https://www.natureindex.com/news-blog/what-scientists-need-to-know-about-fair-data

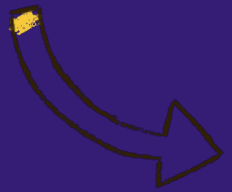
FAIR ≠ Open

‘FAIR is not the equivalent of open, but open data needs to be FAIR to be useful’

Making your data freely and openly available does not translate to it being reusable!

To do so, we need clear, detailed contextual information and data description.

Data can be FAIR but not Open! FAIR data motto “as open as possible, as closed as necessary”



Ideally you want FAIR data shared openly!

What challenges do you face sharing your data openly?

Challenges for open data

I will be scooped!

With very little evidence to support this notion, sharing your data openly, with appropriate licensing ensures your claim of authorship.

This has been one of the major driving factors for the publications of preprints.

Benefits of Preprints

We see preprints as an important step toward a more open and transparent peer review process — one with tremendous benefits for both individual authors and the broader scientific community.

Rapid Dissemination of Your Results

Preprints allow you to share your results when you're ready — whether you're researching an emergent disaster, applying for a grant, or just excited to broadcast your work to a wider audience.

Establishing Priority

It's common for researchers to achieve a similar advance at around the same time but the publication process can artificially delay one paper or favor another. Posting preprints allows researchers to publicly date stamp their discoveries.

Increased Attention (and Citations!)

The sooner research becomes available, the sooner it can begin to receive views and citations. In this case, common sense is backed up by evidence. Research shows that public posting increases the number of times papers are viewed and cited.

https://plos.org/open-science/preprints/

Challenges for open data

I'm not sure I'm allowed to share my data.

A growing number of institutes promote and adopt Open Science practices.

While many funders often require the data made being publicly available, it is always important to check their requirements at the very beginning of the project.

In addition to institutional and funder policies, research is oftentimes funded by public/tax money. Needless to say, it is of utmost importance to gain maximum benefit of the work done by sharing it openly and freely in a reusable format.

Challenges for open data

I'm not sure my data is useful for someone else.

The data itself has an inherent value, that may or may not be realised at the time it was produced.

There are many examples of data reuse in a different context other than the original one it was created for.

Making your data available reduces duplication efforts.

Challenges for open data

What if people misinterpret my data?

Clear and detailed documentation is key! Include rich metadata and outline the restrictions to using this data.

Include your ORCID details to allow a chance for the data reuser to get in touch and inquire before making any judgements or misinterpretations.

Papers Without Code - submission

Papers Without Code: where unreproducible papers come to live.

The goal of this is to save the time and effort of researchers who try to reproduce the results of a paper that is unreproducible. It could either be due to the paper not having enough details or the method straight up not working. In either case, authors will be given the opportunity to respond. The hope is this saves people time and disincentivizes unreproducible papers.

1. First authors of the paper will be informed and given a chance to respond.

2. Submissions with multiple votes and/or a link to a reproduction will be given priority.

3. Every submission will be reviewed to prevent spam. If it is a genuine submission, expect it to be approved within 24 hours.

Note: In order to protect the authors reputation, we will take spamming very seriously.

Contact: papersburned@gmail.com
Results: papers.paperswithoutcode.com

https://www.paperswithoutcode.com/

Challenges for open data

I don’t have time for that, it is too difficult and I don’t know where to start!

Proper data management at early stages reduces time and effort. Consult with data managers and data stewards at your institute. Make use of a variety of online resources such as the Turingway.

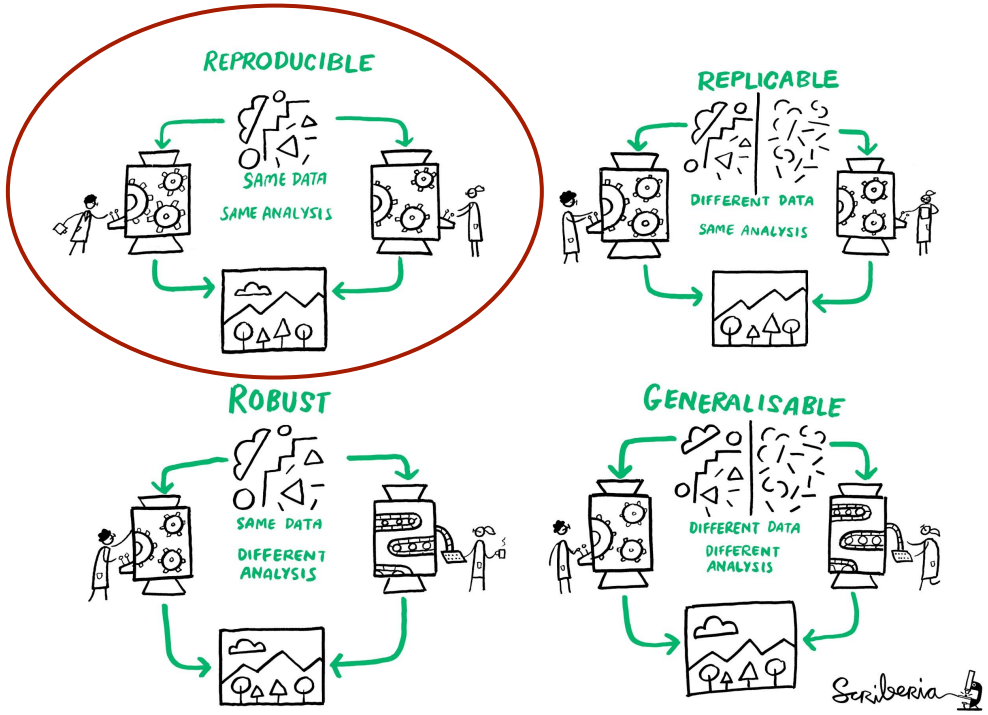
https://the-turing-way.netlify.app/welcome

Questions?

Part 5: Data Reuse and Reproducibility

Reuse: using the data to answer a different question than originally intended

Reproduce: being able to follow the same footsteps to trace back and recreate the same conditions to *hopefully* arrive at the same conclusion!



https://the-turing-way.netlify.app/welcome.html

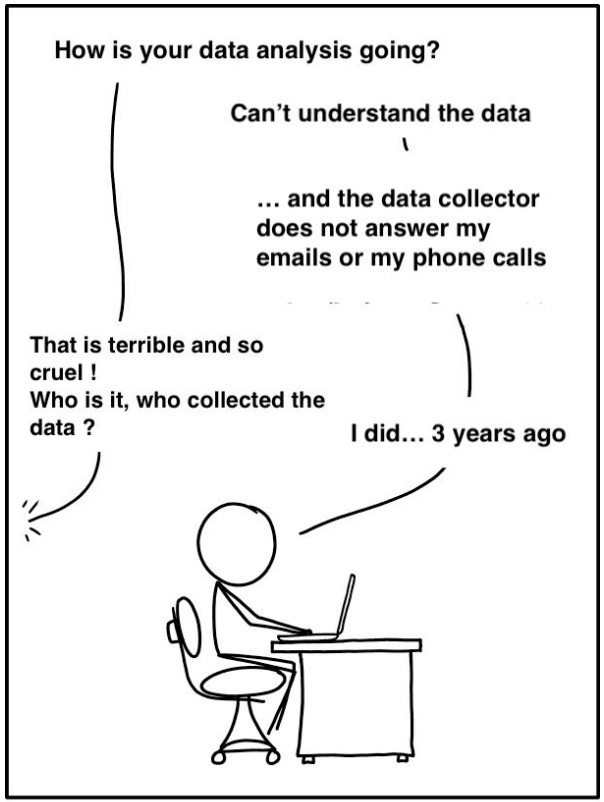
Thanks, but no thanks! - Breakout room

In groups of 2-3 discuss and note down;

- Have you tried replicating an experiment, yours or someone else? What challenges did you face?
- Have you ever received data you couldn't use? & why not?
- What type of information do you wish you received along with the data that you received from external sources?
- What type of information should you include when you share your data with someone else?

Why should I do it?

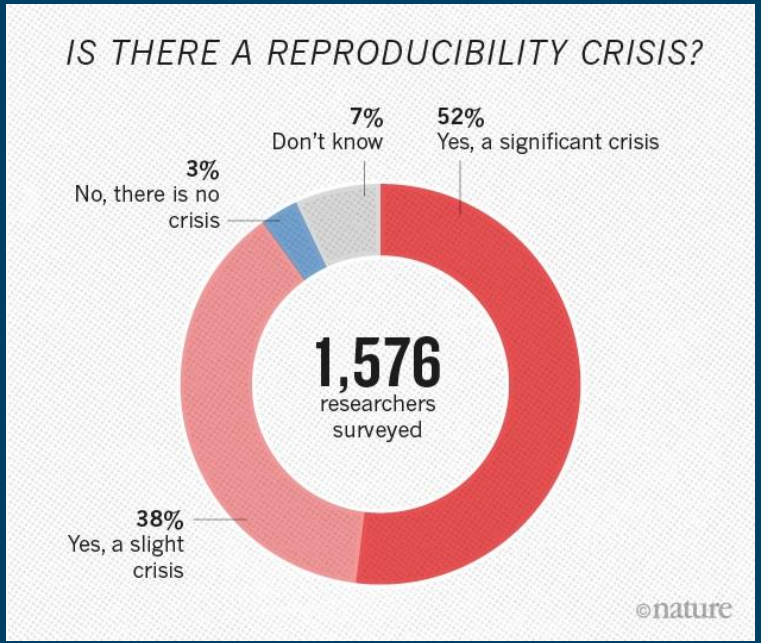
- You are the first one to reuse your data. Do you understand what you did a year ago?
- You are not alone! Research relies on collaboration, can your collaborators understand what you did?



Your first collaborators are your future selves, be nice to them !

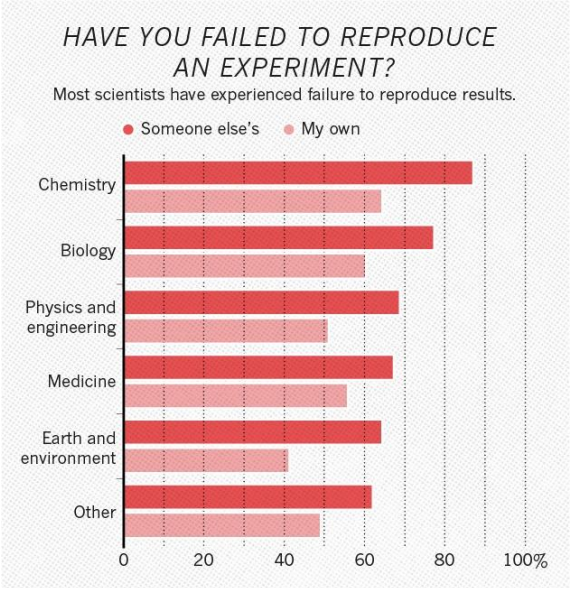
Why should I do it?

Reproducibility crisis!



Why should I do it?

- “More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.”

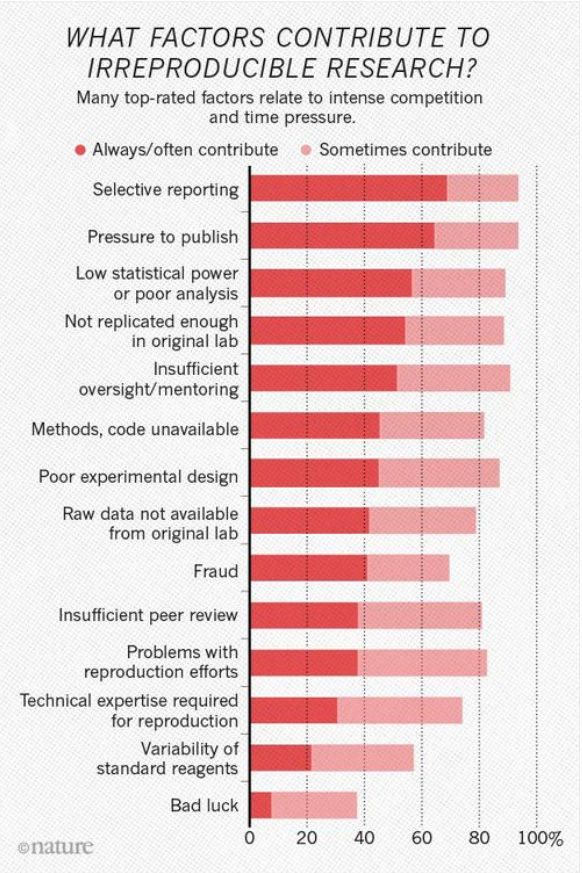


https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Why should I do it?

Factors for irreproducible research include:

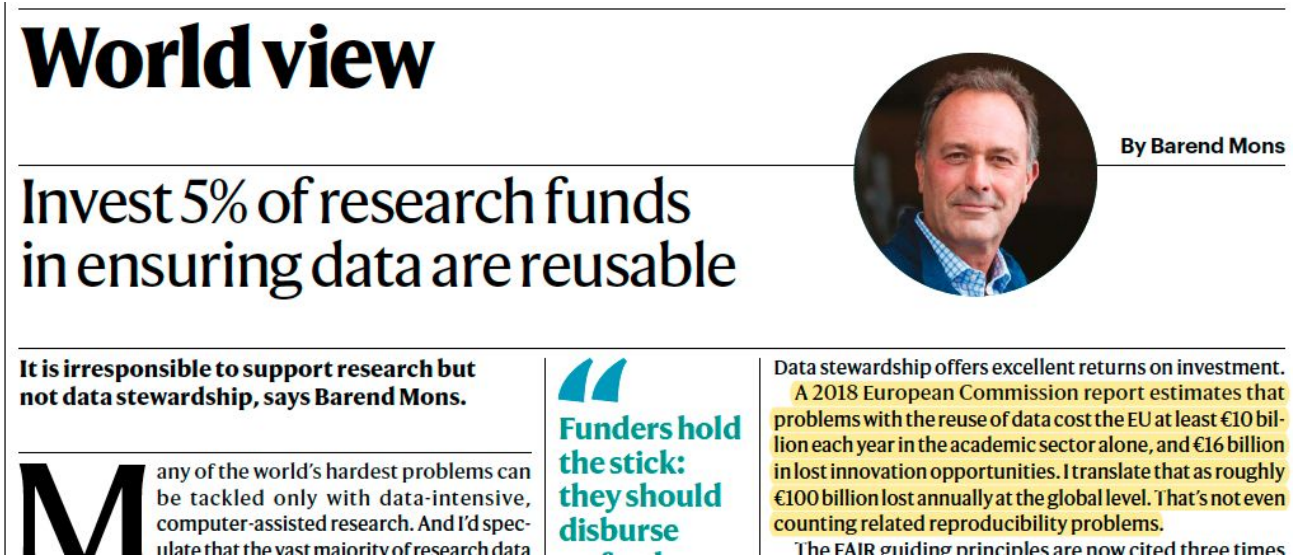
- Selective reporting
- Raw data not available
- Method, code unavailable!



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

Why should I do it?

“A 2018 European Commission report estimates that problems with the reuse of data cost the EU at least €10 billion each year in the academic sector alone”
https://www.nature.com/articles/d41586-020-00505-7



World view

By Barend Mons

Invest 5% of research funds in ensuring data are reusable

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

Many of the world's hardest problems can be tackled only with data-intensive, computer-assisted research. And I'd speculate that the vast majority of research data

Funders hold the stick: they should disburse

Data stewardship offers excellent returns on investment.

A 2018 European Commission report estimates that problems with the reuse of data cost the EU at least €10 billion each year in the academic sector alone, and €16 billion in lost innovation opportunities. I translate that as roughly €100 billion lost annually at the global level. That's not even counting related reproducibility problems.

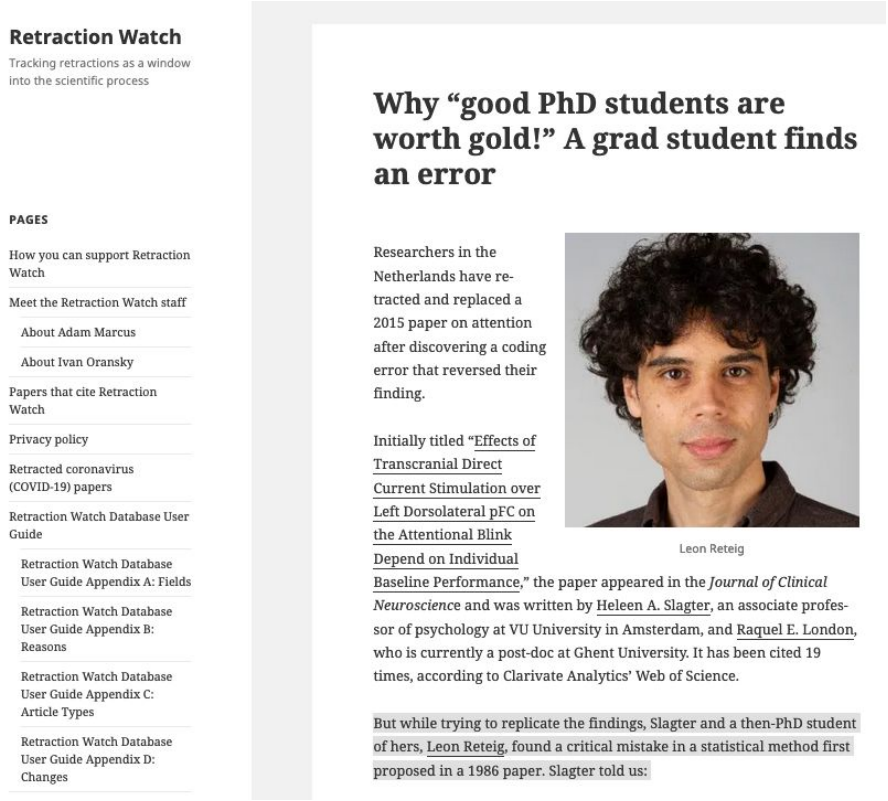
The FAIR guiding principles are now cited three times

Why should I do it?

- Making data available for your reuse, enables new research questions to be answered.
- Good quality data that others want to use, increases citations of the datasets themselves and your research;
- How much time have you spent trying to make sense of your own or someone else's data?
- Compliance with funders and publishers requirements.
- We are losing money because data is not reusable!

Why should I do it?

- Reproducibility ensures the integrity of the data and could affect its use and reuse and is required in order to identify potential problems.
- Problems with reproducibility have real life consequences!!



Retraction Watch

Tracking retractions as a window into the scientific process

PAGES

- How you can support Retraction Watch
- Meet the Retraction Watch staff
  - About Adam Marcus
  - About Ivan Oransky
- Papers that cite Retraction Watch
- Privacy policy
- Retracted coronavirus (COVID-19) papers
- Retraction Watch Database User Guide
  - Retraction Watch Database User Guide Appendix A: Fields
  - Retraction Watch Database User Guide Appendix B: Reasons
  - Retraction Watch Database User Guide Appendix C: Article Types
  - Retraction Watch Database User Guide Appendix D: Changes

Why "good PhD students are worth gold!" A grad student finds an error

Leon Reteig

Researchers in the Netherlands have retracted and replaced a 2015 paper on attention after discovering a coding error that reversed their finding.

Initially titled "Effects of Transcranial Direct Current Stimulation over Left Dorsolateral pFC on the Attentional Blink Depend on Individual Baseline Performance," the paper appeared in the *Journal of Clinical Neuroscience* and was written by Heleen A. Slagter, an associate professor of psychology at VU University in Amsterdam, and Raquel E. London, who is currently a post-doc at Ghent University. It has been cited 19 times, according to Clarivate Analytics' Web of Science.

But while trying to replicate the findings, Slagter and a then-PhD student of hers, Leon Reteig, found a critical mistake in a statistical method first proposed in a 1986 paper. Slagter told us:

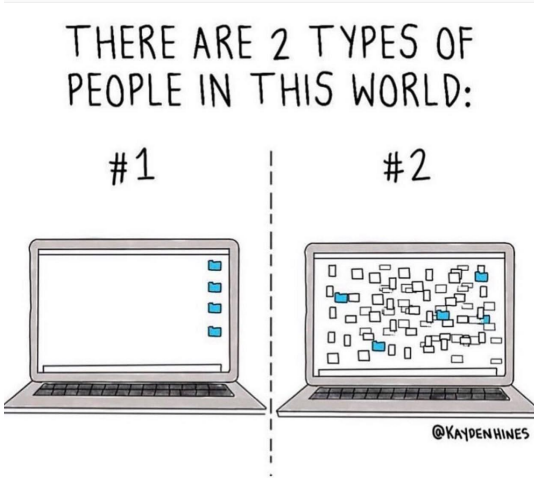
https://retractionwatch.com/page/2/

Part 6: Where do we start?

Proper data management begins at home!

Organize your data

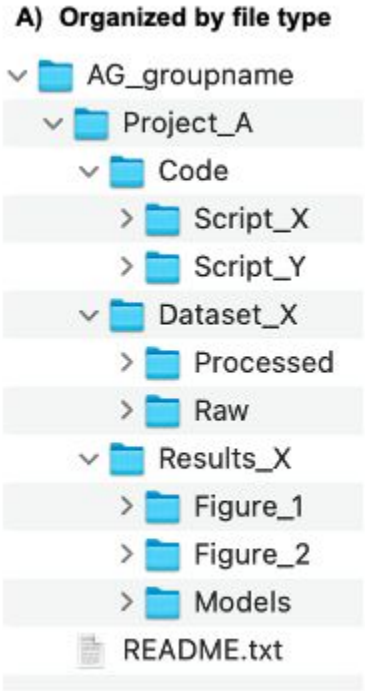
- Organize your data in a logical manner
- Separate the data according to type: i.e. raw data, analysis, code,
- Use directories and folders hierarchy



- A clear directory structure will make it easier to locate files and versions and this is particularly important when collaborating with others.

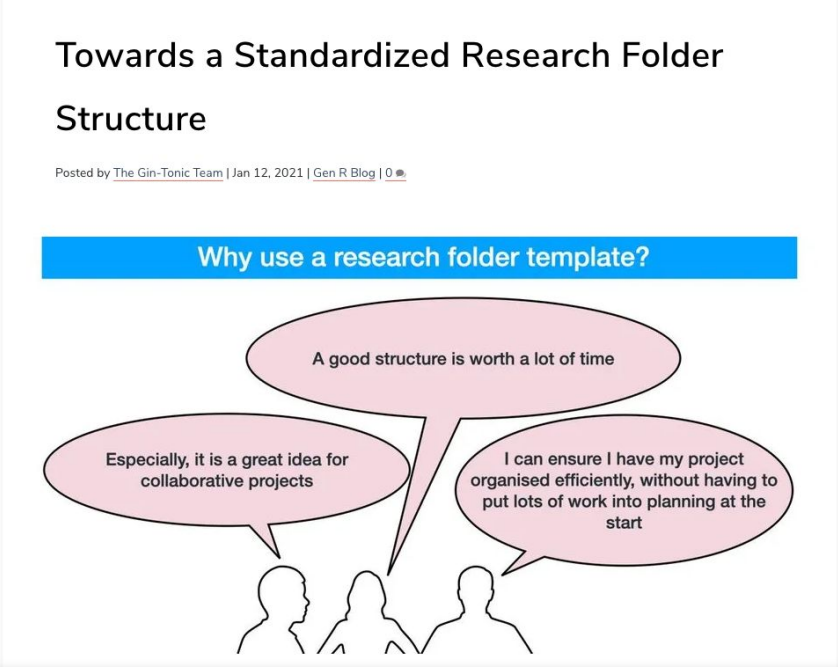
Directory structure guidelines

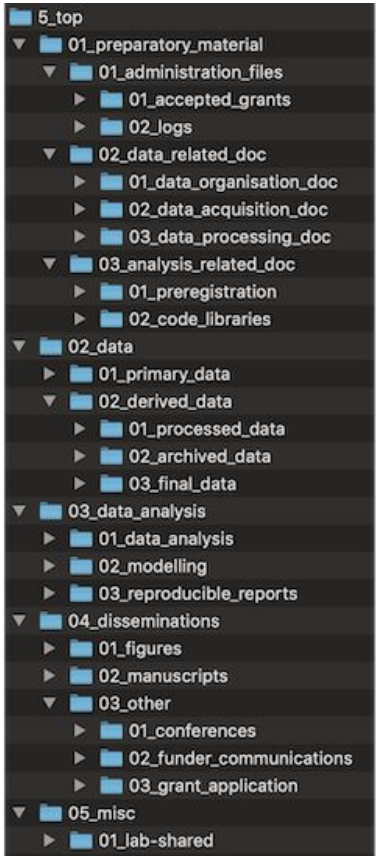
- Consider a hierarchical file structure starting from broad topics to more specific ones nested inside, restricting the level of folders to 3 or 4 with a limited number of items (max. 50 items if possible) inside each folder.

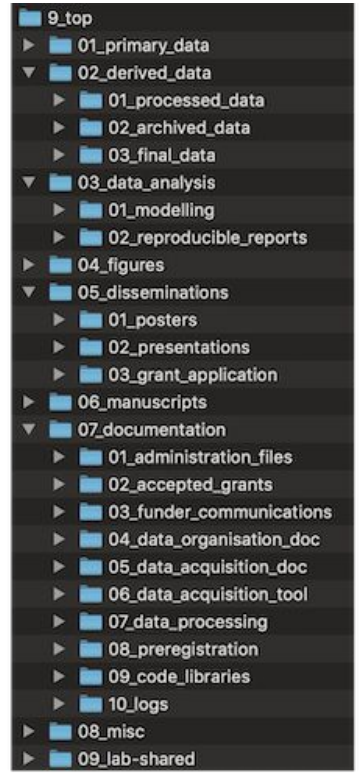


Directory structure guidelines

https://umfrage.hu-berlin.de/index.php/617633?lang=en

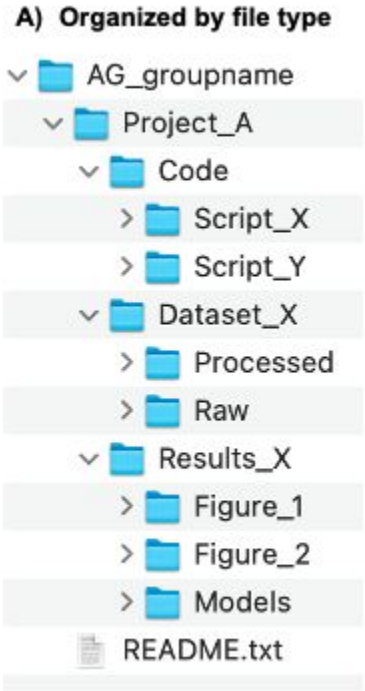






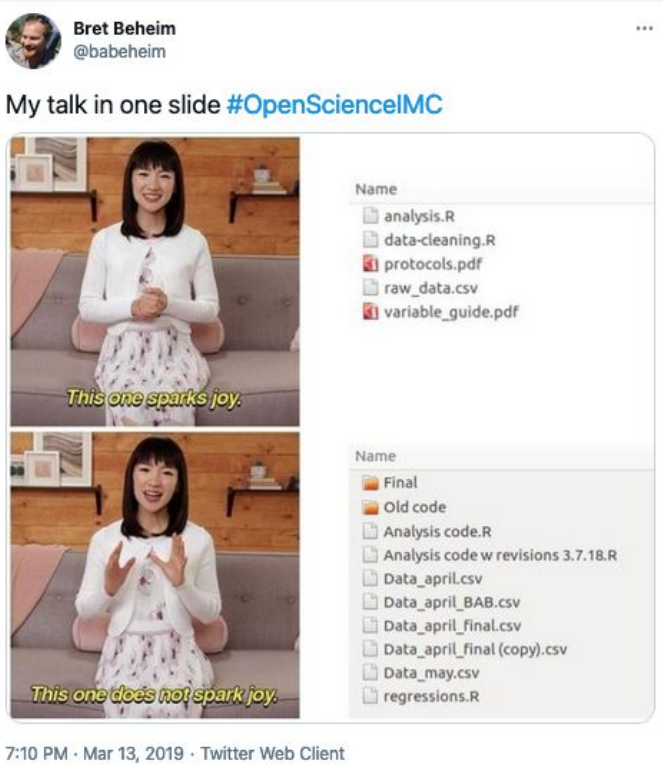
Directory structure guidelines

- Consider a hierarchical file structure starting from broad topics to more specific ones nested inside, restricting the level of folders to 3 or 4 with a limited number of items (max. 50 items if possible) inside each folder.



Name your files

- Common guideline, you should know what your file is before you double click it!
- A file name is the primary identifier to the file and its contents.
- A file name should be unique, consistent and descriptive. This allows for increased visibility and discoverability and can be used to easily classify and sort files.



Do’s and Don’ts of file naming

Do’s

- Create descriptive, meaningful, easily understood names that are not too short or too long, i.e., no less than 12-14 characters except for generic, well-defined names such as README.
- Use identifiers to make it easier to classify types of files i.e., Int1 (interview 1).
- When combining elements in the file name, preferably use underscores (_) or hyphens (-) as an element separator, see examples of commonly used special letter case patterns.
- If applicable, include versioning within file names.
- Make sure the file format extension is present at the end of the name (e.g. .doc, .xls, .mov, .tif, .fasta, .html).
- For dates use the ISO 8601 standard: YYYY-MM-DD and place at the end of the file number **UNLESS** you need to organize your files chronologically.
- For experimental data files, consider using the project/experiment name and conditions in abbreviations.
- Add a README file in your top directory which details your naming convention, directory structure, and abbreviations.

Do’s and Don’ts of file naming

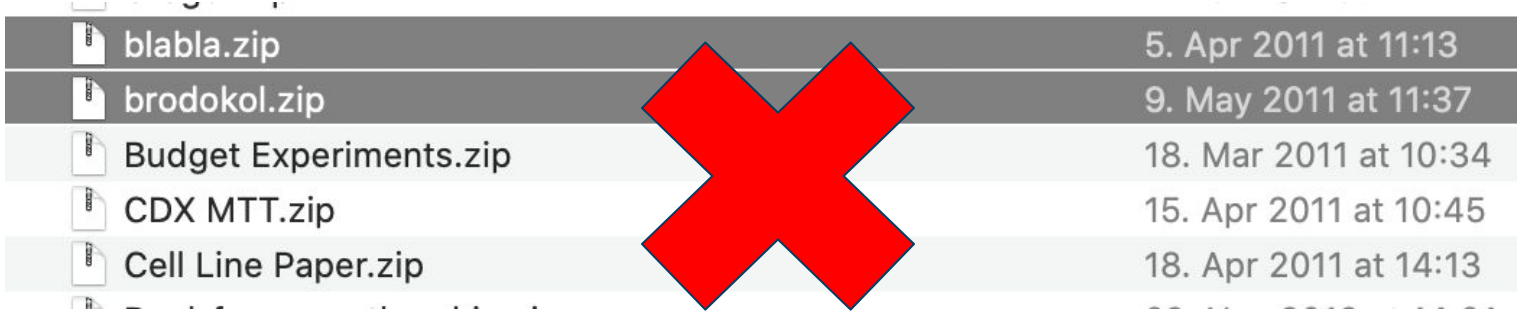
Don’ts

- Avoid using capital letter to separate words such as CamelCase and use underscores or hyphens instead.
- Avoid naming files/folders with individual names as it impedes handover and data sharing.
- Avoid long names. e.g., no longer than 35-40 characters.
- Avoid using spaces, dots, commas and special characters (e.g. " / \ ~ : ! @ # $ % ^ & * ( ) ` ; < > ? , [ ] { } ‘ “ |), or any foreign (Unicode) characters e.g. äöüß r カイダー字 .
- Avoid repetition for ex. Directory name Electron_Microscopy_Images, then you don’t need to name the files ELN_MI_Img_20200101.img.

Do’s and Don’ts of file naming

Examples

- 1900-2000_sasquatch_migration_coordinates.csv
- Smith-fMRI-neural-response-to-cupcakes-vs-vegetables.nii.gz



Tools for bulk renaming

Windows:

- Ant Renamer (www.antp.be/software/renamer)
- Bulk Rename Utility (www.bulkrenameutility.co.uk/)
- Total Commander (https://www.ghisler.com/deutsch.htm)

Mac:

- Renamer 5 (for Mac) (https://renamer.com/)
- Name Changer (https://mrrsoftware.com/namechanger/)
- ExifRenamer (https://www.qdev.de/?location=mac/exifrenamer)

Linux:

- GNOME Commander (www.nongnu.org/gcmd/)
- GPRename (http://gprename.sourceforge.net/)

Unix:

- Rename command

https://doi.org/10.5281/zenodo.1212496

Choose file formats wisely

- During your research, your choice of file formats might be dictated by convenience or instrument provider or team practices.
- To make your data available for others and easy to use, choose file formats that are most commonly used in your field (e.g. fasta/fastq) or open file formats that allow interoperability (e.g. mark down)

	Preferred formats	Accepted formats
Text documents	• ASCII (.txt) • MS Word (.docx) • OpenDocument Text (.odt) • PDF/A (.pdf) • Unicode (.txt)	• MS Word (.doc) • PDF (.pdf) • Rich Text Format (.rtf)
Markup language	• HTML (.html) • JSON (.json) • XML (.xml)	• SGML (.sgml) • Markdown (.md)
Spreadsheets	• CSV (.csv) • MS Excel (.xlsx) • OpenDocument Spreadsheet (.ods)	• MS Excel (.xls) • OOXML (.docx, .docm) • PDF/A (.pdf)
Databases	• CSV (.csv) • SIARD (.siard) • SQL (.sql)	• dBase III or IV (.dbf) • Filemaker Pro (.fmp7, .fmp12) • MS Access (.mdb, .accdb) • OpenDocument Base (.odb)
Statistical data	• OpenDocument (.ods) • SPSS portable (.por) • SPSS SAV (.sav) • STATA (.dta)	• CSV (.csv) • MS Excel (.xls, .xlsx) • R (.rdata, .rda) • SAS (.sas) • SAS transport (.xpt)
Image (raster/bitmap)	• Adobe Digital Negative format (.dng) • DICOM (.dcm) • PNG (.png) • TIFF (.tif, .tiff)	• Adobe Photoshop document file (.psd) • JPEG (.jpg, .jpeg) • JPEG 2000 (.jp2, .jpx) • Raw image data (various formats)

https://snd.gu.se/en/manage-data/guides/suggested-file-formats

Choose file formats wisely

- Choose standard file formats most commonly used in your field.
- Convert data to a standard format.
- Choose a format which is required for data deposition i.e. repository requirements, archival compression.
- Consider exporting or converting from original format to a more open/preferred format but keep in mind that some data/metadata might be lost or altered during the process e.g., text formatting in documents, decimal point formatting, date and time values.
- Keep in mind there are no standard preferred file formats, and none are perfect, but consider choosing open formats that are most applicable for your use and field, specially when sharing!
- When archiving data, combine the whole project (i.e., raw data, analysis, documentation, code and software) in one package.
- For software consider the use of containers to enable interoperability and long-term re-use.

Choose file formats wisely

Tools

- Singularity: https://github.com/hpcng/singularity
- Docker: https://www.docker.com/resources/what-container
- Jupyter : https://jupyter.org/index.html
- Fido: https://github.com/openpreserve/fido
- Vagrant: https://www.vagrantup.com/intro/vs/docker.html

Quality control

- Quality control is a fundamental step in research, which ensures the integrity of the data and could affect its use and reuse and is required in order to identify potential problems.
- It is therefore essential to outline how data collection will be controlled at various stages (data collection, digitisation or data entry, checking and analysis).

How quality control could save your science

It may not be sexy, but quality assurance is becoming a crucial part of lab life.

Monya Baker

27 January 2016

PDF | Rights & Permissions

Chris Ryan/Nature

There are at least six things in this picture that a quality-assurance manager would try to improve. Can you spot them? (Answers, below)

https://www.nature.com/news/how-quality-control-could-save-your-science-1.19223

Quality control

Data collection

- Outline the number of measurements/samples/procedures repeated.
- Outline instrument calibration tests & data set or samples used for calibration.
- Outline standardized controls (e.g., sample controls).
- Use of standardized protocols and methods with clear instructions and documentation.

Quality control

Data entry

- Decide a method for documentation i.e., Electronic lab notebooks vs paper.
- Outline the non-digital data structure and strategy for digitization.
- Collect and create metadata throughout the data collection and handling process
- Use controlled vocabularies.
- Outline how the data/samples/variables are labelled.
- Document terminology used.
- Describe how to flag/tag questionable data.
- Ensure data and time is represented in a machine-readable format and valid.
- Set up validation rules or input masks in data entry software.

Quality control

Data Analysis and checking

- Outline software/code used for analysis.
- Outline strategy for data transfer and controls (e.g., checksum).
- Outline how the data will be cross-checked and validated.
- Assign person/expert for quality assurance and data checks and/or peer review.
- Outline database structure to organize data and data files.
- Document any modifications and outline versioning strategy to avoid duplicate error checking.
- Check and flag questionable data.
- Verify your analysis by using a random data set/samples compare to original data.
- Double-check the code for any errors and ensure appropriate documentation.
- Use statistical analysis to detect erroneous and/or anomalous values.

Quality control

Qualitative data

- For qualitative data such as interviews:
  - Outline guided interview questions.
  - Make use of software tools such as text to speech.
  - Control the quality of audio/video/transcripts files.
  - Refer to the UK data archive guidelines.

Tools

- Open Refine for data quality control https://openrefine.org/
- Numeric data anonymization R-Package: sdcMicro
- UK data archive tools list: https://www.data-archive.ac.uk/managing-data/digital-curation-and-data-publishing/tools-we-use/

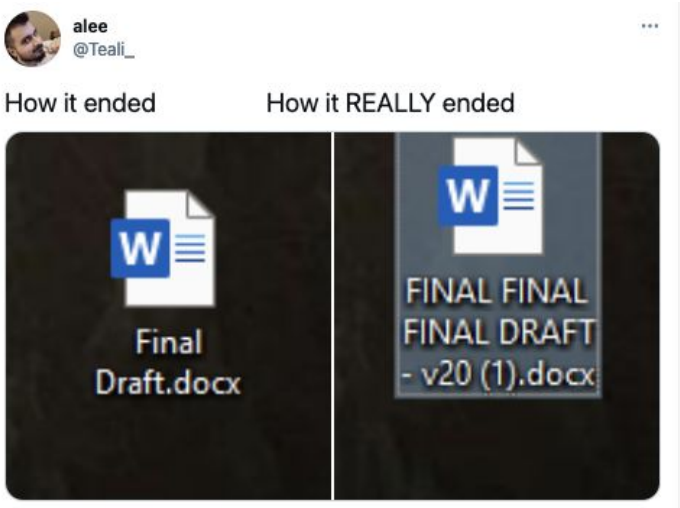
What quality control measures do you use in your experiments?

Versioning

- In order to keep track of changes made to a file/dataset, versioning can be an efficient way to see who did what and when, in collaborative work this can be very useful.

Versioning

- A version control strategy will allow you to easily detect the most current/final version, organize, manage and record any edits made while working on the document/data, drafting, editing and analysis.



Versioning

- Outline the master file and identify major files for instance; original, pre-review, 1st revision, 2nd revision, final revision, submitted.
- Outline strategy for archiving and storing: Where to store the minor and major versions, how long will you retain them accordingly.
- Maintain a record of file locations, a good place is in the README files.
- Record any related files and documents and any updates/changes made to them
- Use a systemic and unique naming system to identify the different versions, e.g., numbers and/or dates.
- Include a version control table that outlines the file history, which version, where the other versions are located, list all associated files and their versions and modifications, add dates, authors, access rights, licensing, and details of changes made since the last version.

Versioning

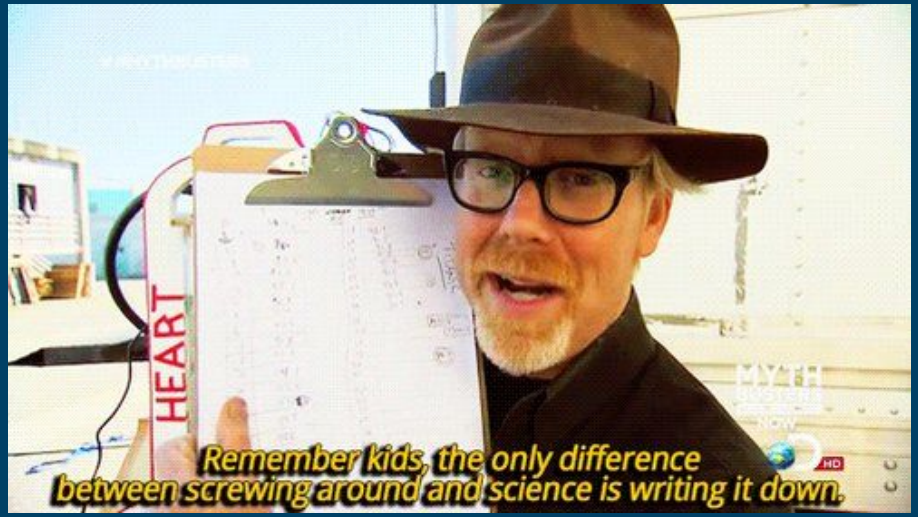
Tools

- Sharepoint (not for personal or sensitive data)
- Github (not for personal or sensitive data)
- DropBox (not for personal or sensitive data
- Google Docs (not for personal or sensitive data)

The Turing Way Community, & Scriberia. (2020, March 3). Illustrations from the Turing Way book dashes. Zenodo. http://doi.org/10.5281/zenodo.3695300

Questions?

Part 7: Documentation



https://www.tested.com/making/557288-origin-only-difference-between-screwing-around-and-science-writing-it-down/

Electronic lab notebooks

Paper Lab-notebooks - in use since the 15th Century!

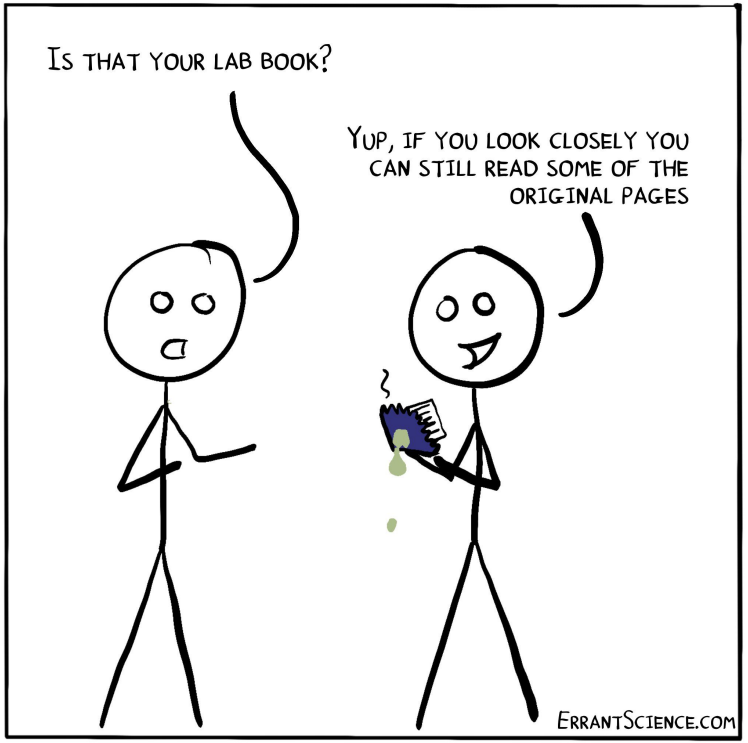
- Not searchable!
- Can be easily damaged, misplaced
- Hard to share and backup
- Legibility issues
- Integration of digital data is not easy

Can you find the nobel prize in here?

https://www.benchfly.com/blog/how-to-keep-a-lab-notebook/

Electronic lab notebooks

- Lab notebooks are the **primary** reporting space, hence the need for ensuring proper documentation.
- Lab notebooks (LN) are used to document research data including; hypothesis, experimental procedures, analysis, interpretation and reporting, which makes them the primary space for data capturing and recording.



Electronic lab notebooks

eLNs offer obvious advantages for researchers:

- Data readily available for reuse,
- Seamless data extraction and data is searchable
- Structured and detailed documentation allowing traceability
- Facilitated data analysis and sharing
- Creation of templates

Electronic lab notebooks

Digital expansion:

- Massive digital expansion in the lab, the one aspect that has not benefited from digitization is record taking,
- Lab inventory management
- Seamless integration with lab equipment and digitally acquired data

https://openworking.wordpress.com/2019/07/05/keep-calm-and-go-paperless-electronic-lab-notebooks-can-improve-your-research/

Electronic lab notebooks

Compliance:

- Proof of provenance and ownership, protection of intellectual property
- Ensured long term availability of the data
- Protection against data manipulation or loss

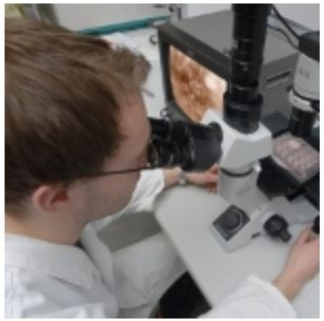
Electronic lab notebooks

EU

- **Good Laboratory Practice (GLP)** introduced by Organization for Economic Co-operation and Development(OECD)
- **Good Manufacturing Practice (GMP)** introduced by EudraLex Volume 4 —GMP Guidelines, Annex 11.

Adherence to compliance requires the software to offer a system which meets the technical requirements and procedural controls within an organization.

GLP Federal Bureau



What is GLP?

"Good Laboratory Practice (GLP) is a quality system concerned with the organisational process and the conditions under which non-clinical health and environmental safety studies are planned, performed, monitored, recorded, archived and reported."

This is the definition of "Good Laboratory Practice" in the "OECD Principles of Good Laboratory Practice" which were then transposed into EC Directives and, after that, into German law as Annex 1 of the German Chemicals Act. The entire sixth section of the Chemicals Act is devoted to Good Laboratory Practice. In paragraphs 19a to 19d the scope and type of monitoring of GLP are laid down by law.

https://www.bfr.bund.de/en/glp_federal_bureau-1488.html

Electronic lab notebooks

USA

Title 21 CFR Part 11 of the Code of Federal Regulations introduced by US Food and Drug Administration (FDA).

https://www.ecfr.gov/cgi-bin/text-idx?SID=140a04c31974ef0891cfb2555bc3a865&mc=true&node=pt21.1.11&rgn=div5

Title 21: Food and Drugs

PART 11—ELECTRONIC RECORDS; ELECTRONIC SIGNATURES

Contents

Electronic lab notebooks

Benefits on institutional level:

- It ensures data integrity, provenance and ownership which is essential in case of resolution of intellectual property issues
- Gaining maximum benefit and impact from research investments
- Reduced financial loss; current estimated loss is around 10,2 billion Euros / year for not having FAIR data implementation.

Fore more information:

- https://op.europa.eu/en/publication-detail/-/publication/d3766478-1a09-11e9-8d04-01aa75ed71a1/language-en
- https://www.nature.com/articles/d41586-020-00505-7

WORLD VIEW · 25 FEBRUARY 2020

Invest 5% of research funds in ensuring data are reusable

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

Barend Mons

What to look for in an ELN

01 Onsite installation

Data security controlled by us, connected to local storage

02 User Friendly

Search, Easy to use, academically oriented, chat support services, training videos

03 Compliant

GLP (Good Laboratory Practice) Compliance & FDA 21 CFR Part 11

04 Workflows

Standardized workflows and templates, saves time, quality assurance

05 Open Export

Export in **open formats**, you can export your data- Data is reusable

06 Inventory Module

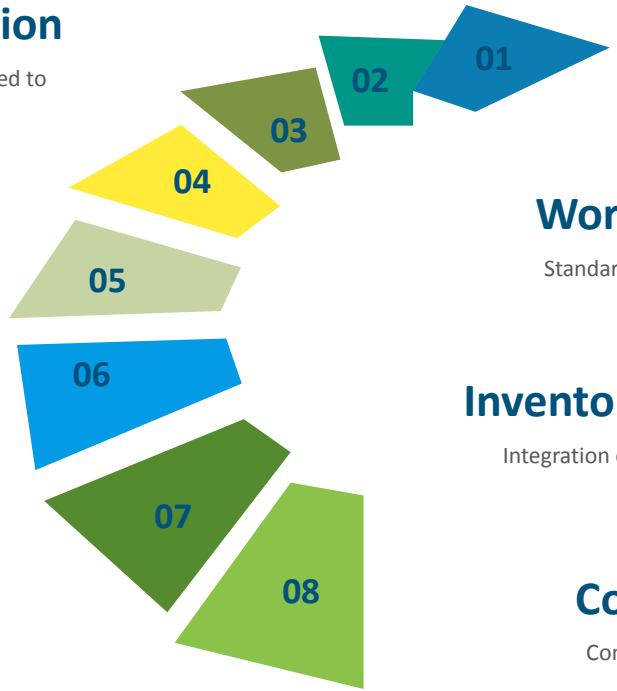
Integration of sample management and tracking system

07 Integration

Integration through API, can connect to instruments

08 Compatible

Compatible with different systems; Windows, Mac, Linux, Android, IOS

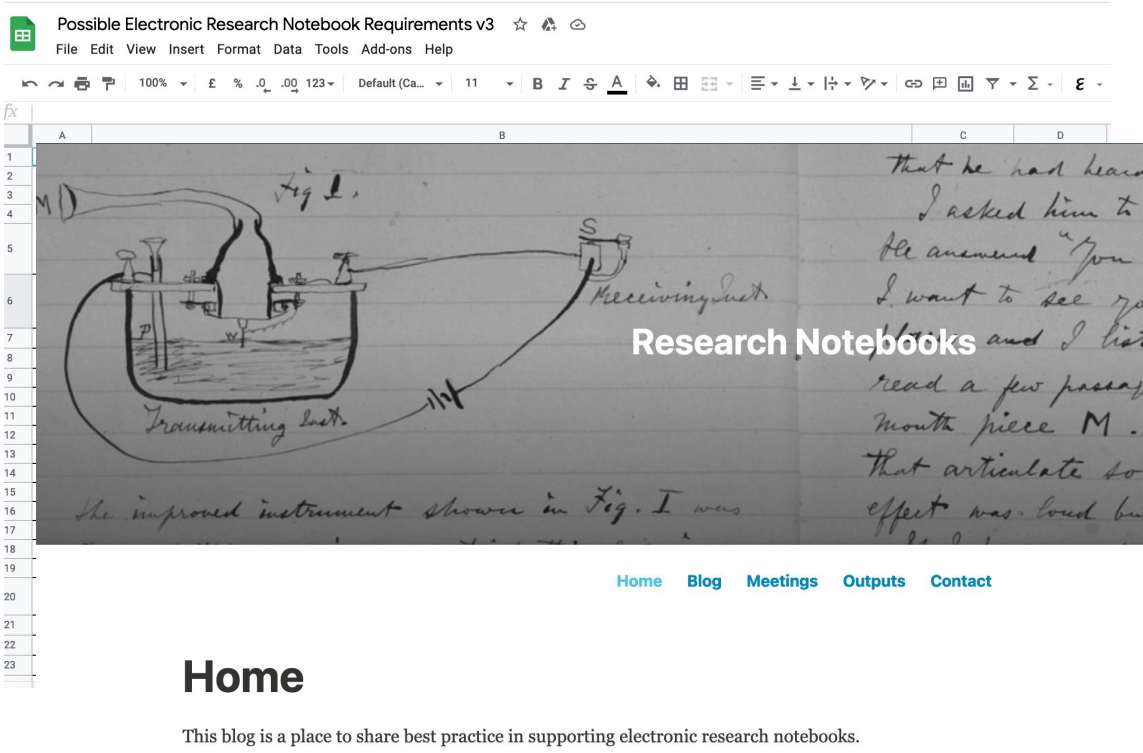


How to choose an ELN

Comparison Matrices:

- Adapted Matrix: https://docs.google.com/spreadsheets/d/1egUW3ZewyIaJ_lhEe8uJrd-69ycVbqQjiqbJxbs_jVY/edit#gid=1172088166

- Research Notebooks Blog https://researchnotebooks.wordpress.com/outputs/



How to choose an ELN

Comparison Matrices:

- Harvard Matrix: https://shrtco.de/G55Dy

HARVARD MEDICAL SCHOOL

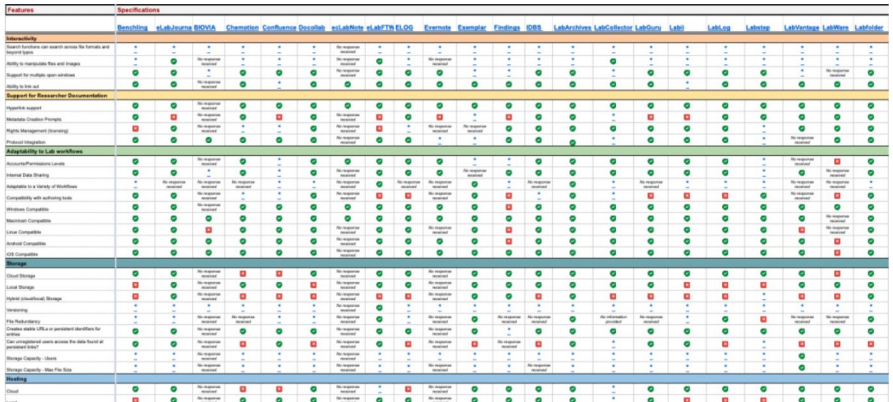
ELN Comparison Grid

As the number of available ELN tools increases and the functions of each quickly change, it may be confusing to evaluate all of the advantages and limitations of each when looking for the right solution for your project.

The Electronic Lab Notebook (ELN) Matrix has been created to aid Harvard researchers in the Longwood Medical Area (LMA) in the process of identifying practical ELN tools to meet their specific research needs. Using the ELN Matrix, researchers can compare and contrast the numerous tools available today, and also explore individual options in-depth.

The tools listed in the ELN Matrix are selected strictly on their relevance to the LMA biomedical research community with respect to their intended function to replace and enhance conventional (hard copy) laboratory notebooks. Inclusion of a tool in the ELN Matrix does not imply an endorsement by Harvard, and exclusion of a tool is based solely on relevance; no evaluations of tool quality were made or acted upon in developing the ELN Matrix.

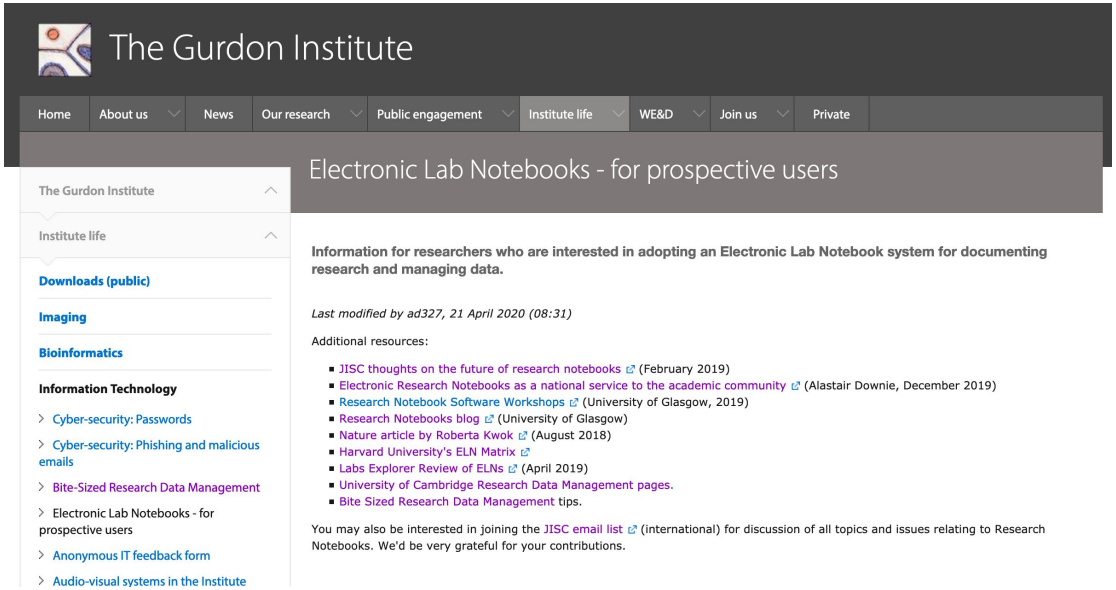
Click on the matrix below to expand the table and see detailed descriptions of each resource.



How to choose an ELN?

Blogs & websites:

- IT and Research Data Management in the Gurdon Institute https://www.gurdon.cam.ac.uk/institute-life/computing/elnguidance & https://gurdoncomputing.blog/
- Open Working from 4TU.ResearchData & TU Delft Library
  Open Science Framework: https://doi.org/10.17605/OSF.IO/JR9U2

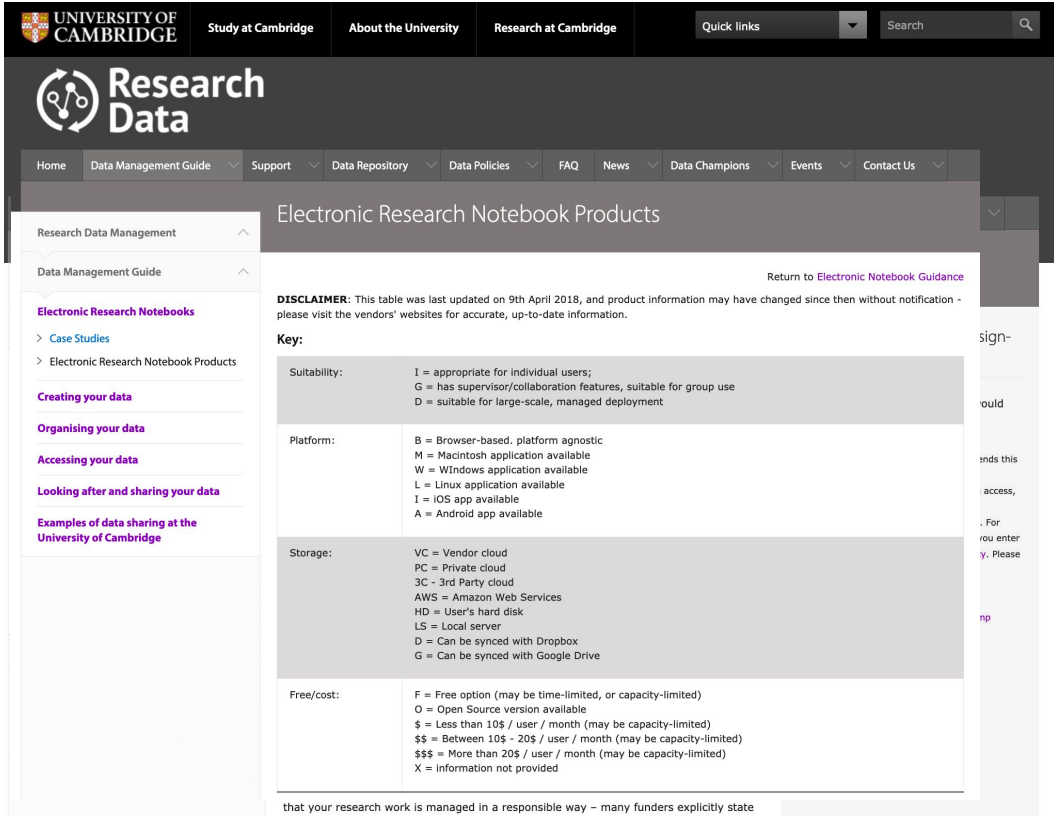


How to choose an ELN?

Blogs & websites:

- Cambridge University guide:

https://www.data.cam.ac.uk/data-management-guide/electronic-research-notebooks/electronic-research-notebook-products



Build your own ELN- Breakout room

In groups of 2-3 discuss and note down;

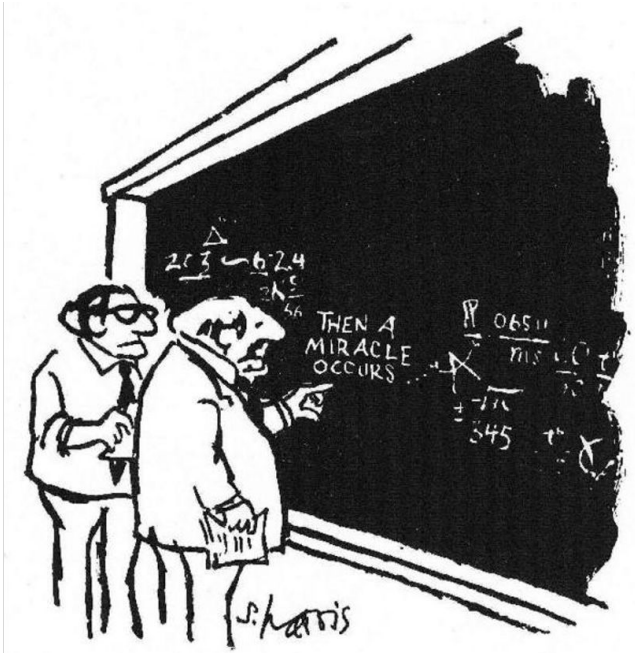
- Choose the features you must have in an ELN
- Choose the features you would like to have in an ELN
- Highlight the top 5 features that you look for when choosing an ELN
- With the help of the top 5 criteria, and the matrices + guidelines mentioned below, indicate your top two choices of ELN solutions
- Please note down your choices in one of the templates according to your group number

Questions?

Metadata

What should I tell others?

- **Who** created and owns this data?
- **What** are the contents?
- **What** output and results?
- **When** was this data created and last updated?
- **Where** is it stored and published?
- **Which** methods were used?
- **Which** instruments were used?
- **How** was the data created, controlled and analysed?
- **How** can I use this data i.e. license?



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

http://www.sciencecartoonsplus.com/pages/gallery.php

Metadata

How to capture metadata?

- Use ELN to record your work 
- Use versioning controls to track history, progress and changes in a descriptive manner 

- Use metadata standards
- Use README files

Metadata

What are Metadata standards (also known as Schemas)?

- Metadata standards enable the structuring of metadata and enhance its interoperability, by using common terms and definitions, to provide consistency and accuracy to data documentation.

- The standards can be discipline specific or general such as Dublin Core, DataCite Metadata Schema, Data Documentation Initiative (DDI) and International Standards Organisation (ISO).

- Example of metadata standards and tools for lab-based research:
  - **ISA framework** and tools: https://isa-tools.org/
  - Minimum Information for Biological and Biomedical Investigations: https://fairsharing.org/collection/MIBBI

Metadata

What are Metadata standards (also known as Schemas)?

- Metadata is frequently required for depositing data in repositories.
- Metadata standards offer controlled vocabularies with predefined terms to ensure consistent use and clear definition of terms and concepts.
- Metadata standards offer technical standards that ensure units of measurement, time, are entered in controlled formats, i.e. date and time formats

How many ways can you say “female”?

18-day pregnant females	female (lactating)	individual female	worker caste (female)
2 yr old female	female (pregnant)	lgb*cc females	sex: female
400 yr. old female	female (outbred)	mare	female, other
adult female	female parent	female (worker)	female child
asexual female	female plant	monosex female	femal
castrate female	female with eggs	ovigerous female	3 female
cf.female	female worker	oviparous sexual females	female (phenotype)
cystocarpic female	female, 6-8 weeks old	worker bee	female mice
dikaryon	female, virgin	female enriched	female, spayed
dioecious female	female, worker	pseudohermaprhoditic female	femlale
diploid female	female(gynoecious)	remale	metafemale
f	femele	semi-engorged female	sterile female
famale	female, pooled	sexual oviparous female	normal female
femail	femalen	sterile female worker	sf
female	females	strictly female	vitellogenic replete female
female - worker	females only	tetraploid female	worker
female (alate sexual)	gynoecious	thelytoky	hexaploid female
female (calf)	healthy female	female (gynoecious)	female (f-o)
hen	probably female (based on morphology)		

female (note: this sample was originally provided as a \"male\" sample to us and therefore labeled this way in the brawand et al. paper and original geo submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a female individual)",

Courtesy of N. Silvester, European Nucleotide Archive, EMBL-EBI

Metadata

What are the different types of metadata?

Descriptive metadata: Information outlining basic facts necessary for discovery and identification, i.e. title, authors, keywords and abstract,

Structural metadata: Information regarding the structure (organisation and relationship) of a data and underlying items. For instance it could be a description of enclosed files and scripts, how they are organized, and structured and how they are related and where they can be found i.e. DOI

Administrative metadata: Information that describes the technical information and information regarding management of the data including, licensing and copyright permissions, technical requirements, file formats, provenance (i.e. history of ownership, who owns the data and where did it come from), access and sharing controls and permissions, quality controls and integrity checks.

Metadata

Which file format should I use for my Metadata?

- A text or html document.
- An XML document linked to data files
- Information embedded in an XML data file

XML (eXtensible Mark-up Language) files includes key data and metadata documentation that is interoperable for web browsers and analysis engines which in turn enables field specific searching,

Questions?

Caption this

https://www.instagram.com/p/CKZQaM2li4l/?igshid=87ua47kvk52l

Metadata

README

The purpose of a README file is to give an overview of the content, aiding individuals in making sense of the data enclosed thereby persevering the long-term value of the data. This can be very helpful if you are sharing your data with others, or to keep track of content and edits or changes made in multiple projects, or revisiting data after some time has passed.

- A README file is better suited for a collection of data such as a directory for a specific project or experiment, software tool, or any data that is related to each other “logically”.
- Place the README file in a parent directory associated with the content described.
- Use plain markdown or a simple text editor to create the README file in either .md or .txt file format.
- For dates use the ISO 8601 standard: YYYY-MM-DD.
- Whenever possible use the standard vocabulary from your field, see metadata standards directory by RDA community.

README file- Breakout room

In groups of 2-3 discuss and note down;

- Identify a data set of interest, assign the researcher most familiar with the dataset to be the “contact person” to describe and answer questions regarding their research.
- The other members of the group take turns to ask questions under the following headings and fill in the information.
- As you are asking questions, consider your own dataset and how you would respond.
- In case you don’t already have an ORCID, create one https://orcid.org/
- Create a text file named README.txt, include information provided below, and deposit it in the parent directory of your project.

→ README file template:
https://github.com/selgebali/RDM_Workshop/blob/master/docs/4.6-READMEfile.md

Wrap up

- Open Discussion
- How likely are you to recommend this session to a friend or colleague?
  - Why or Why not?
  - What worked well?
  - What didn’t work and why?
  - What would you have changed?
  - Anything else you would like us to know?

The End