



Welcome to the Reproducible Research Workshop!

Danny Garside (they/them)



social.coop/@da5nsy



dannygarside@proton.me

Trainer: Danny Garside (they/them)

- Previously: postdoc at University of Sussex
- Currently: Trainer (*and community manager*) with the [Digital Research Academy](#)



Contacts:

- Website: <https://www.dannygarside.co.uk/>
- Email: dannygarside@proton.me
- Socials: <https://social.coop/@da5nsy>



Tour of materials

Link to these slides (to follow along, and/or for you to modify and re-use later!):

- [Google slides](#)
- .pdf/.odp on Zenodo: doi.org/10.5281/zenodo.15769886
- They are based on...

Workshop materials: originally co-developed by Amanda Miotto and Adam Partridge:
<https://carpentries-incubator.github.io/ReproducibleResearch/>

Built using github and The Carpentries lesson template:

- 'Episode' list
- Newly developed
- CC:BY licence, so please re-use!
- **Like all great art, never finished, a continuous work in progress - contributions and feedback welcome!**

Particify activity! (2 questions)

Go to partici.fi, and enter:

7938 9467

Keep the tab open in the background!

(Particify is like menti but open-source)



Particify Qs (for reference - slide is skipped)

How familiar are you with Reproducible Research? (multiple choice - donut visualisation)

- I am a reproducible research expert!
- I use reproducible research techniques frequently
- I use reproducible research techniques infrequently
- I know roughly what is is, but haven't used reproducible research techniques before
- I don't know what reproducible research is

How code-confident are you?

- I code every day
- I dabble
- I've done a little
- I have not coded
- I find code scary

Let's begin...

Learning Objectives

By the end of this session, participants will:

- Learn what is reproducibility and why does it matter?
- Gain a number of skills and resources to help us build reproducibility in our everyday workflows
- Understand the relationship between Reproducible Research and other areas such as Culture and Business Continuity

Training Design Notes

Guiding principles / restrictions:

- Discipline-agnostic
- ~~No~~ *minimal* code
- Actionable! Not 0-100, but a set of practices that can be implemented

Each section has 'next steps' that we will review

What is Reproducible Research?

In this lesson, we'll learn:

- What is reproducible research?
- The different terms around reproducibility

For today:

“Research that is sufficiently transparent that someone with the relevant expertise can clearly follow, as relevant for different types of research:

- how it was done;
- why it was done in that way;
- the evidence that it established;
- the reasoning and/or judgements that were used; and
- how all of that led justifiably to the research findings and conclusions.”

Turing Way: Reproducible vs other terms

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

The Turing Way: <https://book.the-turing-way.org/reproducible-research/overview/overview-definitions.html#reproducible-matrix>

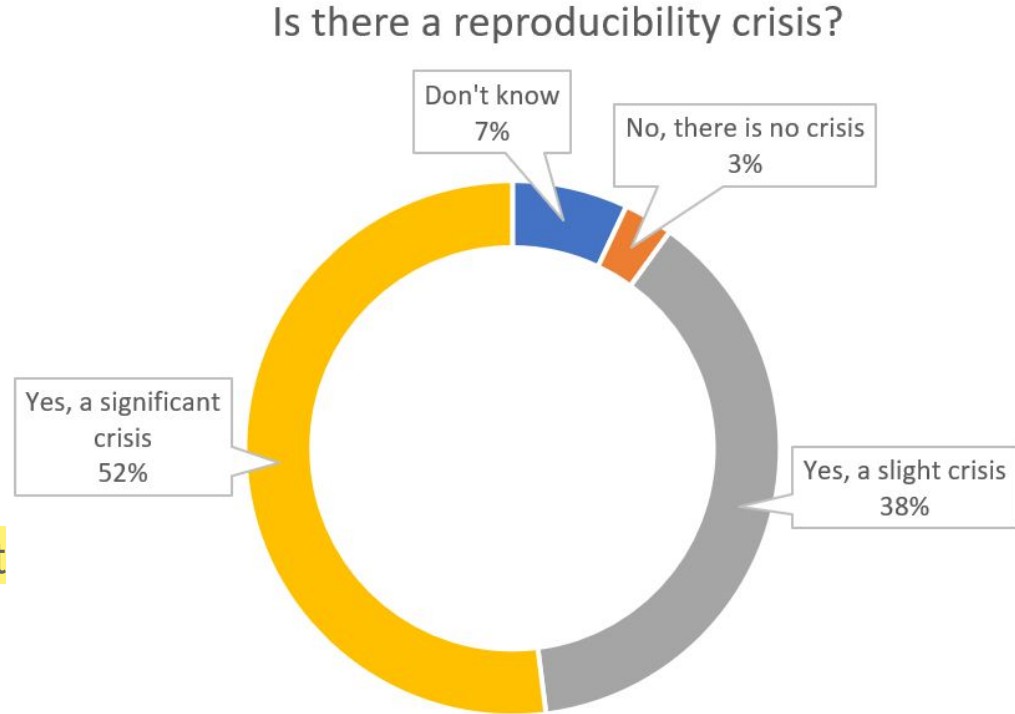
Why does Reproducibility Matter?

In this lesson, we'll learn:

- About the reproducibility crisis
- How reproducibility can be improved broadly
- How making our work more reproducible can also benefit ourselves

Baker (2016): ‘1,500 Scientists lift the lid on Reproducibility’

- Over 70% of researchers who had tried to replicate another researcher's experiments failed.
- Over half had failed to reproduce their own experiments.
- Half of researchers surveyed agreed that there was a significant crisis of reproducibility



Other evidence

- Preclinical Cancer Research (Begley & Ellis, 2012):
 - Only able to replicate 6 of 53 landmark findings
- Reproducibility Project: Cancer Biology (Errington et al., 2021):
 - attempted replication of 193 experiments
 - 2% had open data, 0% of protocols completely described
 - for 32% of experiments, original authors were not helpful/unresponsive
 - only 50 of 193 replications could be conducted
 - using a binary judgement, 46% of effects replicated successfully

Particify activity! (1 question)

Go to partici.fi, and enter:

7938 9467

Keep the tab open in the background!

(Particify is like menti but open-source)



Particify Q (for reference - slide is skipped)

What factors could contribute to low reproducibility?

Particify Q type is 'open ended' free text responses

Explaining low reproducibility

- Research misconduct/fraud (~2%, [Fanelli, 2009](#))?
- Researcher degrees of freedom ([Simmons et al., 2011](#))
- 'Questionable Research Practices' (c.f. [Ulrich & Miller, 2020](#))
 - Publication bias ([Ioannidis, 2005](#))
 - Underpowered studies ([Button et al., 2013](#))
 - P-hacking ([Wicherts et al., 2016](#))
 - HARKing ([Kerr, 1998](#))

However: these issues are often unintentional, not fraudulent!

Many contributing factors!

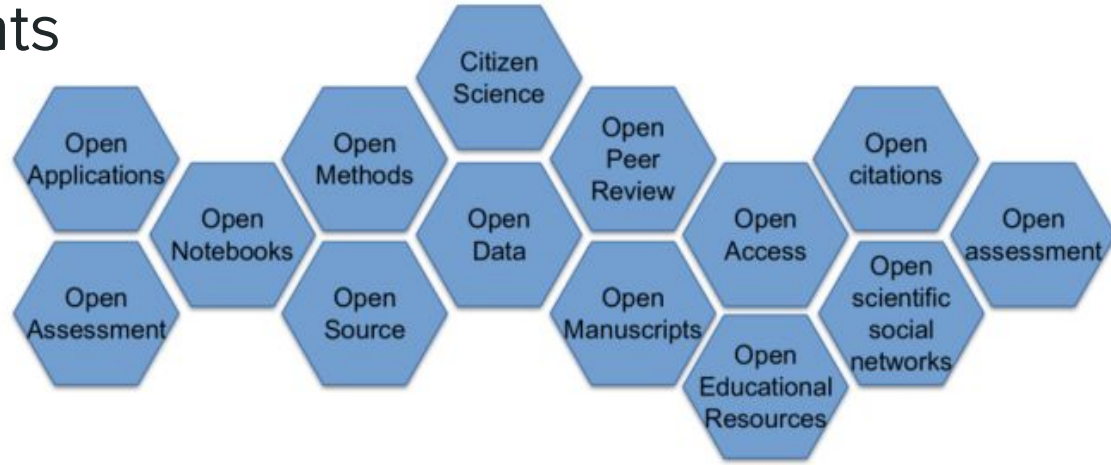
- Difficulty in managing complex datasets or poor statistical practices
- Poor research design, including a failure to control for bias
- A lack of access or detail of the methodology used
- A lack of access to raw data and research materials
- A 'Publish or Perish' culture, that only rewards novel findings

We won't be able to fix all of these today!

We can focus on making small improvements to our own practice (and advocate for wider improvements)

Suggested Improvements

Open research practices, e.g:



Alexander Refsum Jensenius, 2020:
[Why is open research better research?](#)

- Preregistration of experiments
- More training on study design and statistics
- Publishing preprints
- More support for publishing negative results
- More replication and validation studies conducted and shared

“Five selfish reasons to work reproducibly”

Working reproducibly has many benefits for the research ecosystem. However, there are also direct benefits for yourself.

Markowetz (2015) listed these “five selfish reasons to work reproducibly”:

- Reproducibility helps to avoid disaster
- Reproducibility makes it easier to write papers
- Reproducibility helps reviewers see it your way
- Reproducibility enables continuity of your work
- Reproducibility helps to build your reputation

Our 7 steps towards Reproducible Research

- 1) Planning to be Organised
- 2) Keeping your files Tidy and Organised
- 3) Methodology and Protocols
- 4) Documentation and writing it down
- 5) Testing and Controls
- 6) Automation
- 7) Publishing, Persistent Identifiers and Preparing for Reuse

Step 1 - Planning to be Organised

In this lesson we'll learn:

- How to plan our research by considering Data Management Plans and implementing a folder structure
- To keep a copy of raw data safe and secure separate to our working data
- How to consistently name our files
- To consider and map the metadata of our files

Step 1 - Planning to be Organised

Topics

- DMP
- Folder structures
- File naming convention

Challenge 1 - How could these filenames be improved?

What is your next step?

Data Management Plans

A data management plan (DMP) is a living document for a research project, which outlines data creation, data policies, access and ownership rules, management practices, management facilities and equipment, and who will be responsible for what.

They ensure that things like data sharing are considered throughout the process.

These may be recommended or mandated by your institution - check your library for more information!

e.g. The University of Sheffield:

<https://www.sheffield.ac.uk/library/research-data-management/planning>

More resources are linked on the [training materials webpage](#)

Folder Structures

Simple example (Turing Way):

```
compendium/  
├── data  
│   └── my_data.csv  
├── analysis  
│   └── my_script.R  
├── DESCRIPTION  
└── README.md
```

TIER protocol:

```
Project/  
├── The Read Me File  
├── The Report  
├── Data/  
│   ├── InputData/  
│   │   ├── Input Data Files  
│   │   └── Metadata/  
│   │       ├── Data Sources Guide  
│   │       └── Codebooks  
│   ├── AnalysisData/  
│   │   ├── Analysis Data Files  
│   │   └── The Data Appendix  
│   └── IntermediateData/  
├── Scripts/  
│   ├── ProcessingScripts/  
│   ├── DataAppendixScripts/  
│   ├── AnalysisScripts/  
│   └── The Master Script  
└── Output/  
    ├── DataAppendixOutput  
    └── Results
```









File Naming Conventions

Data Carpentry defined three key principles to guide file naming:

- Machine readable
 - No spaces
 - Limited use of special characters, e.g. - or _ used as separators
 - CaSe sEnSiTiViTy
- Human readable
 - Can vary by context, but use descriptive words that make sense
- Plays well with default ordering
 - Operating systems order things by default using numbers in the filename
 - Hence using the date in YYYY-MM-DD format will organise files in date order

File Naming Conventions

You have a DEADLINE TODAY - which would you prefer:

 draft2		 2024-03-01_reportv0_AM
 reportFINALFINAL		 2024-03-03_reportv1_AM
 draftFINAL	vs	 2024-03-05_reportv2_SP
 FINALusethisone!!!!		 2024-04-01_reportv3_AM
 report2FINALusethisone!!		 2024-04-07_reportv4_SP
 draftreport		 2024-04-12_reportv5_SP

What is your next step?

Beginner

Advanced

Your next move can be:

Ensure all your projects are well organised. This could extend to projects across your research group.

Educate others in your space on good naming conventions and build a reference guide to be used across your group.

Get a metadata file set up for your media.

If you are working in Python, you can use a [python package](#) developed by researcher Nikola Vukovic to generate a logical, standardised, and flexible directory hierarchy for academic research.

Step 2 - Keeping Files Tidy and Organised

In this lesson we'll learn:

- How to tidy the data inside our files
- How to organise columns and rows
- How to handle dates
- How to handle missing data

Tidy Data

The fundamental principles of tidy data are:

- Each variable is a column
- Each row is an observation
- Each cell contains one value

What could we do to improve the following example?

Address
170 Kessels Road Nathan Qld 4111 Australia
1 Parklands Dr, Southport QLD 4215

Tidy Data

This format includes a lot of benefits:

- It is clear what is expected to be included in each column
- Easier to search
- Less likely to get extra whitespace or punctuation that could cause issues with analysis software

Here is a better example:

Street Number	Street Name	Street Type	Suburb	State	Postcode	Country
170	Kessels	Road	Nathan	Qld	4111	Australia
1	Parklands	Drive	Southport	Qld	4215	Australia

Other useful concepts

Similar to filenames, use a standardised format for dates e.g. YYYY-MM-DD

- Careful about autocorrect in Excel!

Avoid using punctuation or special characters in cells

Avoid highlighting/colouring cells to add extra information - this will not be picked up by analysis software

Missing Data

A blank cell could mean many different things!

You could use a placeholder like NA.

Generally avoid using 0 as this can affect calculations.

Consistency is key - decide (with your team) how you are going to handle missing data and stick to it!

OK - so now you know the data is missing.. how does this affect your analysis?

- Remove the whole observation? Exclude this participant from which analyses?
- You can pre-register this decision to show transparency

Messy Data - activity

Let's clean this messy data. What changes would you make?

Date	Age	Where was the article	Blood type	Preferred Gender
23/1/2023	32	United States	AB	Female
Feb-17	54		A	N/A
2/2/2000	12	USA	0	Male
23-Mar-93	53	Brisbane, Australia	#NAME?	Mail

Note: #NAME? should be -O

Messy Data - activity solutions

Dates in the same format

Standard way to handle missing data

DoB instead of Age, as Age can change depending on the date of the year

Standard way to handle location

Spelling mistake: 'Mail'

Because it's been migrated from Excel, the format of -O was changed.

Tools to help with this

Tools that don't involve programming:

OpenRefine is an open source (free) tool that is incredibly useful to clean column data. Your data doesn't leave your computer, which makes it safer when working with sensitive data.

Programming Tools:

R has the **Tidyverse library**. Python has the **Pandas library**, which follows similar principles. You may also find **NumPy** useful.

What is your next step?

Beginner

Advanced

A great place to start is:

Decide how to handle missing data

Use YYYY-MM-DD for dates

Beginner

Advanced

Your next move can be:

You can start to work in a data science language such as R or Python. [The Carpentries](#) data science lessons are a great place to start, made for researchers who have never coded before. Workshops are held across the world.

Step 3 - Methodology and Protocols

In this lesson, we'll learn:

- What details to keep about our literature reviews
- About Pre-registrations and where to submit
- What details to include about your methodologies and protocols
- Why and where to publish your protocols

Step 3 - Methodology and Protocols

Topics:

- Literature Review
- Preregistration
- Registered Reports
- Transparent Methodologies and Protocols
- Publishing your Protocol

Literature Review

While you may not be publishing all the details of your literature review, it can be helpful for your future self to retain these details. You may want to use part of this data for your next study, and need to remember why you discounted a paper, or if you felt there was bias unaccounted for in a study.

- What databases did you search?
- What search terms did you use?
- What filters did you use?
- Why did you include or dismiss certain papers?

Pre-registration - what is it?

A plan of key study and analysis details and decisions, made before data collection.

Uploaded to a site like the Open Science Framework (osf.io) or aspredicted.org

- Distinguish between confirmatory and exploratory hypotheses.
- [A plan, not a prison - transparently reporting deviations](#)
- Can be private or embargoed.
- Using secondary data? [Template & tutorial](#)

Registered Reports

The Center for Open Science provides the following definition:

“Registered Reports is a publishing format that emphasizes the importance of the research question and the quality of methodology by conducting peer review prior to data collection. High quality protocols are then provisionally accepted for publication if the authors follow through with the registered methodology...”

Registered Reports

...

This format is designed to reward best practices in adhering to the hypothetico-deductive model of the scientific method. It eliminates a variety of questionable research practices, including low statistical power, selective reporting of results, and publication bias, while allowing complete flexibility to report serendipitous findings. ”

Transparent Methodologies and Protocols

Can you remember what percentage of protocols were completely described in the Reproducibility Project: Cancer Biology (Errington et al., 2021)?

Transparent Methodologies and Protocols

Can you remember what percentage of protocols were completely described in the Reproducibility Project: Cancer Biology (Errington et al., 2021)?

0% of protocols completely described

This is a major barrier to reproducibility! What to report will differ by discipline - health researchers have a wide range of reporting guidelines to inform them:

<https://www.equator-network.org/>

Publishing your Protocol

Builds trust in your work

Enables discovery

Expands your publication records

Offers an early opportunity for peer review, with an opportunity for feedback and improvements

It creates an early record of your novel methodologies, software, and/or innovations

- [Open Science Framework \(OSF\)](#) - A free open source project management tool. You can register all types of review protocols including scoping reviews.
- [Protocols.io](#) is a platform dedicated to protocol publication.
- Many journals will publish protocols - check the journals in your field.

What is your next step?

Beginner

Advanced

A great place to start is:

Learning about how preregistration and registering protocols work by checking out some of these [ReproducibiliTeach videos](#) or check out this [The Turing Way page on Methods and Protocols](#)

Step 4 - Documentation

In this episode we'll learn:

- What documentation is and how it can help us
- How to consider the audience you are writing for
- What to document for staff onboarding and offboarding
- The curse of knowledge and how that skews our perception
- What to include for documentation

What is Documentation?

“Documentation is a love letter to your future self”

- Damian Conway

While it can refer to structured software documentation, when getting started, it just means writing things down! You can always update them as processes change.

- lab handbook ([Fay Lab](#), [Peele Lab](#))
- a group instruction guide
- a team wiki
- a knowledge repository

Why use Documentation? - activity

A key member of your team becomes unavailable:

- What do you need to know?
- Do you know where their work is stored?
- What about their data?
- Do you know where their procedures and protocols are stored?
- How about their research contract or data custodianship details?

‘Bus factor’: the minimum number of team members that have to suddenly disappear from a project before the project stalls due to lack of knowledgeable or competent personnel.



Why use Documentation?

Documentation is for your team but also for you, it:

- Helps you track what you are doing
- Gives a point of reference for future you
- alleviates some of the mental load of remembering everything
- Can act as a history of changes if you realise you've taken a wrong turn.

How to start Documenting?

Your institute may have the eLabNotebooks cloud platform or similar available. But you don't need specialist software - you can have a text file or word document that you store in the same place as your data.

Note-taking can be combined with TODO list management using tools such as logseq or obsidian.

On- and Off-boarding colleagues is an opportunity for documenting

New team members need to learn processes anyway, so are great 'fresh eyes'

- What might they need to know?

You could provide leaving team members with a checklist

- What might they need to document?

Onboarding - examples

- Expected contact and work hours?
- Best methods of communication across the team?
- What is the expected research culture and values?
- How to contact IT/Library/Researcher support/campus security/safety officer?
- Where do you expect research data is stored?
- How do they find out about restrictions and rules around custodianship and sharing of data
- How is authorship and author order decided when submitting papers?
- Does the research group post papers as preprints?
- Do you expect data to be published as open or FAIR? Is there commercial interest around the data? What licences are expected for research data/code?
- How does someone learn about data sensitivity and what they need to be aware of?
- Examples of previous ethics, governance, grant applications
- What technologies are usually used in the group?
- What analysis tools or methodologies are usually used in the group?
- How does hardware hire/booking work?

Offboarding - examples

- Is a copy of their data being stored at your institute? Where?
- Is there any colleagues in the research group or supervisors that currently also have access to the data?
- Is this raw data, processed data and final data? Or just one of these? Which is which?
- What is the details on ownership, custodianship and reuse of this data?
- Who were/are the collaborators?
- What is the retention date of this data?
- Is a copy of the data also going to another institute (whether it be a collaborator or the staff member's future institute)?
- Have you got grant ids, publication links and any other public information associated with this data? Is the dataset published or stored in an external repository
- Where is any ethics or governance approvals for this data?
- What information is there available on analysis, methodology and protocols?
- What tools were used for analysis? What software and software versions? What hardware and hardware models? (if applicable)

The Pizza Protocol - activity

We are going to hone our documentation skills by preparing a pizza protocol!

Write out a step-by-step guide on
how to make a pizza.

You have 1 minute to write your protocol!

Go!



The Pizza Protocol - Expert Bias / The Curse of Knowledge

Did you include a base?

The Pizza Protocol - Expert Bias / The Curse of Knowledge

Did you include a base?

Did you make the base from scratch?

The Pizza Protocol - Expert Bias / The Curse of Knowledge

Did you include a base?

Did you make the base from scratch?

Does that mean using pre-bought flour or did you grow the wheat?!

The Pizza Protocol - Expert Bias / The Curse of Knowledge

Did you include a base?

Did you make the base from scratch?

Does that mean using pre-bought flour or did you grow the wheat?!

Did you use flour?!?!

The Pizza Protocol - Expert Bias / The Curse of Knowledge

Did you include a base?

Did you make the base from scratch?

Does that mean using pre-bought flour or did you grow the wheat?!

Did you use flour?!?!?

Demonstrates:

- 1) Can be tricky to know how much detail to include in a protocol/documentation
- 2) We all have biases/make assumptions based on our experiences

What should you document?

Many things we've already discussed are forms of documentation, but extra details can always help when you return to a project after some time

- Data Management Plan
- Did you pre-register and if so, where is it? You can document justifications for the decisions made in the pre-registration.
- Did you write a protocol? Did you publish it?
- How will you handle missing data, outliers? When will you stop collecting data?

A longer list of things you could include is given on the [training webpages](#).

Other related tools

Standard Operating Procedures (SOP)

- Step-by-step instructions for routine procedures. If you end up teaching people the same thing repeatedly, an SOP may be helpful.

Workflow mapping / analysis pipeline tools

- How did you clean and analyse your data? While tracking this is easier if you're using scripts, you can download the function list from SPSS/NVIVO as the code used to create it. If you're using an analysis program like SPSS, Stata, or Excel.

Password managers

- Can help track different account credentials and help prevent losing access

Documentation - Reflection activity

Imagine you are contacted 2 years from now about a project you're working on now, as a similar paper was published with contradicting results.

- How well (and easily) do you think you'd be able to answer questions about it?
- What can you do to make that process easier?

“Sorting Out the FACS: A Devil in the Details”

Two labs in the US had contradicting results.

They worked together for over a year, swapping machines, samples and even working side by side to find the difference. In the end, the methods for stirring a liquid were different, which caused different results.

They wrote up and published an article describing this:

Hines, 2014. <https://doi.org/10.1016/j.celrep.2014.02.021>.

What is your next step?

Beginner

Intermediate

Advanced

A great place to start is:

Open a document and start a diary. You can write at the end of the day what you did, any results, anything you got stuck on, and any notes to yourself.

You can also write down where your data is saved (On a cloud storage anywhere, external hard drive etc).

That's it. That's a good start.

Step 5 - Testing and Controls

In this lesson we'll learn:

- Why we should be checking our data for validity and integrity during processing
- What we should be looking for when inspecting our data
- Tools for inspecting data
- That physical testing and hardware QA plays an important role too
- Our data may have a lineage of origin, and we need to be aware and document the provenance of our data
- That it is important to track our analysis history (and how to record it)
- That version control is a way to track changes over time

Testing for Validity and Integrity

Unintended changes to datasets happen all the time:

- An added comma in a sentence can throw out a line in a spreadsheet.
- Missing data can alter a calculation.
- A broken link can leave us with unnoticed missing data.

Some quick checks can help avoid errors due to these issues:

- Is there the expected number of lines and columns
- If you find the unique entries, do those entries make sense?
- Do the calculations make sense? e.g. are averages within the range?

Real world example

Genes like SEPT2 and MARCH1 were automatically converted to dates by Excel.

“A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.”

Ziemann, 2016. <https://doi.org/10.1186/s13059-016-1044-7>

“[The previous] article on this topic led the Human Gene Name Consortium to change many of these gene names to be less susceptible to autocorrect.”

Abeysooriya, 2021. <https://doi.org/10.1371/journal.pcbi.1008984>

Testing for Validity and Integrity

Even when everything makes sense, we should still be careful!

One potential bias that can affect reproducibility is trusting a positive result because we are excited and hoping for one! We may check negative results more closely to understand ‘what went wrong’.

*“The first principle is that you must not fool yourself,
and you are the easiest person to fool”*

- Richard Feynman

Physical Testing and Quality Assurance

Consider any hardware you use:

- Is it calibrated?
- Is it producing unexpected effects? (e.g. compressing, resizing, filtering)
- Is it fit for purpose? e.g. headphones produce all frequencies required

Consider any consumables you use:

- Are they still in date?
- Are they stored and labelled correctly?
- Were samples affected by transportation?

Comparison with controls are useful for testing these things

Providing Authenticity and Validity

Data lineage considers the data origin, what happens to it, and where it moves over time.

Consider your research data. What is its original source?

When using secondary data:

- Have you noted where your source obtained the information?
- Your source may not be the data owner - who is?
- Are you able to contact the original creator?
- What copyright and access limitations are on the data?
- If you are using a repository that is regularly updated (satellite images, weather patterns, government policies or legislations etc), have you noted the version of the data?
- Do you have all the necessary information (codebook, raw data etc)?

Tracking your Analysis history

You should have a **backup** of the raw data that you can always go back to.

How can we log the changes we make?

- Open Refine, NVIVO, SPSS all have action logs that you can download and save.
- SPSS analysis pipeline comes as a .sps script file. You may see it called 'Syntax'.
- SAS has a .sas file for pipelines.
- STATA has a .do file for pipelines. You may see it referenced as 'commandlog'.

Coding!

This course aspires to be “code free”, but this is where coding rather than using a GUI (graphical user interface) can be really beneficial:

as you're working directly with the 'commands', writing a script enables you to rerun the same analysis every time, especially if you use dependency management tools (**conda**, **venv**, **apptainer**, **pixi**, etc.).

Version Control

Let's now consider tracking the versions of your analysis pipeline.

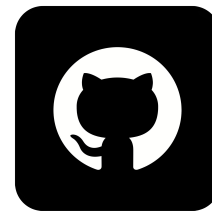
You will likely be making changes to your analysis pipeline as you go

- How are you taking notes of these changes?
- What version of software are you using?
- Have you noted what version of the libraries you are using are?
- Have you noted the name, model and version number of any hardware you may be using (for example, cameras, microscopes, MRI machines, IoT sensors)?

Version control is keeping track of each change you have made, so that if you need to go back to a previous version of your analysis pipeline, you can!

Think of version control as ***having a time machine***.

Version Control - Git



If you are using R or Python, you can use [Git](#) to track changes in your code. You may have heard of [GitHub](#) - this is a cloud platform that you can use to track, share, and collaborate on your code as you write it.

Sharing your analysis pipeline then becomes easy. You can even write reports with both plain text, R or Python code and share the results and graphs in an interactive form using [Jupyter/Quarto](#) or a [Shiny app](#).

More resources for learning Git are given on the [training webpage](#).

What is your next step?

Beginner

Advanced

Your next move can be:

Ensure your data is well described (As per Step 2).

Check that it is clear which of your datasets pair with your Analysis pipelines and in what order.

Publish your protocols and code.

Step 6 - Automation

In this lesson, we'll learn:

- Why automation can be beneficial
- What tools are useful for automation
- Ideas on what to automate

Can you automate any repetitive tasks?

Automation is getting software to do repetitive tasks for you. These tasks often introduce human error and automating them can save you time, energy, and mental load.

‘No code’ solutions for automation include:

- MacOS Automator
- Microsoft Power Automate
- Spreadsheet Macros and formulas (although avoid using these in datasets)
- Workflow systems such as KNIME and Orange Data Mining

Coding is also great for automation. e.g. you could use use shell scripts to run batches of other scripts at once.

Step 6 - Automation

Some ideas:

- Contact me form
- Automated systematic review
- Photos and Videos
- Survey
- Automated Sentiment Analysis
- Take data and automate reports/graphs

What is your next step?

Beginner

Advanced

A great place to start is:

Automate a single thing.

Step 7 - Publishing, Persistent Identifiers, and Preparing for Reuse

In this lesson, we'll learn:

- The difference between an identifier and a persistent identifier
- What a DOI and ORCID is
- How to get a DOI minted for your articles and datasets
- If and how to share your datasets
- What FAIR data is, and how mediated sharing works
- What to consider for licensing
- Where you can deposit your datasets or grey materials
- how negative results can still be important to publish

Identifiers and Persistent Identifiers

An identifier is any label used to name an item (whether digital or physical):

- URLs are an example of digital identifiers
- Personal names are also identifiers, but are not necessarily unique as you may share the same name with other researchers around the world.

<https://www.griffith.edu.au/ereseach-services/hacky-hour> would direct to the correct website, until the team got renamed during a restructure.

Barcode: 32888493 may work in a lab, however may not be unique outside a lab. Or the product with the barcode may be discontinued.

Identifiers and Persistent Identifiers

A persistent identifier is long-lasting unique digital reference to a webpage, digital object, even a person.

- DOI - e.g. <https://doi.org/10.5281/zenodo.10624752>
- ORCID - e.g. <https://orcid.org/0000-0002-0133-4155>

A DOI is a unique alphanumeric string that identifies content and provides a persistent link to its location on the internet.

ORCID provides a persistent digital identifier (an ORCID iD) that you own and control, and that distinguishes you from every other researcher. You can link your professional information — affiliations, grants, publications, peer review, and more.

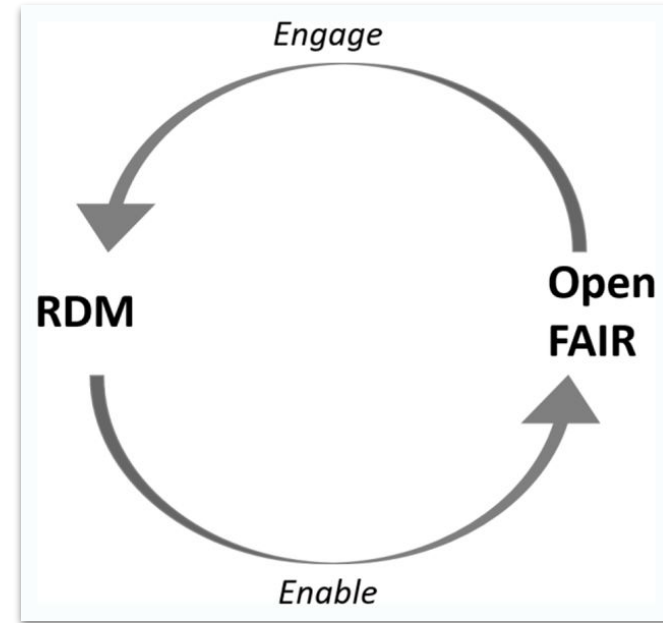
Deposit your final data / analysis - Open/FAIR/can't share?

Findable - e.g. using a persistent identifier like DOI

Accessible - e.g. accessed using standardised protocol

Interoperable - e.g. saved in a non-proprietary format

Reusable - e.g. permissively licensed



Higman, Bangert & Jones, 2019. Three camps, one destination: the intersections of research data management, FAIR and Open. <https://doi.org/10.1629/uksg.468>

Particify activity! (1 question)

Go to partici.fi, and enter:

7938 9467

Keep the tab open in the background!

(Particify is like menti but open-source)



Particify Q (for reference slide is skipped)

Particify Q type is multiple choice

What is your experience with FAIR?

- None before today
- I've heard of it
- I've got some understanding
- I'm a FAIR pro

Licensing

A licence provides guidance and sets legal obligations on:

- Who can reuse the material and for what?
- Can this be used commercially?
- No warranties (or similar) are given
- If someone uses your work as part of their project, are they obligated to also use the same license?
- Do they need to attribute your work?

Common open licenses for articles and datasets are CC-BY, CC-BY-SA, CC-BY-NC

The Center for Open Science has created [this guide to licensing](#).

Your library can also provide further guidance as institutions differ.

Where to deposit?

Deposit final state data to support your publications in an institutional or discipline data repository which **can mint a DOI** for your work.

[FAIRsharing.org](https://fairsharing.org) is a great tool for finding relevant repositories (and standards/policies)

Your institution might recommend a specific repository or there might be a discipline-specific repository.

If not, you can deposit data at:

- [Figshare](https://figshare.com) or [Zenodo](https://zenodo.org)

Or find a repository using [Re3data](https://re3data.org), or [PLOS One guidance](https://plos.org)

Contributor Role Taxonomy (CRediT)

CRediT is a list of 14 types of contribution which are common in academia.

See: <https://credit.niso.org/>

Using the taxonomy (*having a section at the end of the publication saying "Person X did Role Y"*) means that people can be given credit for the specific type of work that they contributed to a project, which can be useful for transparency, recognition, and career progression.

[Conceptualization](#)

[Data curation](#)

[Formal analysis](#)

[Funding acquisition](#)

[Investigation](#)

[Methodology](#)

[Project administration](#)

[Resources](#)

[Software](#)

[Supervision](#)

[Validation](#)

[Visualization](#)

[Writing – original draft](#)

[Writing – review & editing](#)

Publishing negative results

While it can be disheartening to get negative results, these results are still beneficial to the research community at large.

You worked out something didn't work, which is important knowledge all in itself. Sharing this means others don't need to reinvent the wheel, saving resources.

There are a number of journals that specialise in these results:

- [Positively Negative](#)
- [The Missing Pieces](#)
- [Journal of Articles in Support of the Null Hypothesis](#)
- [The All Results Journals](#)
- [Journal of Trial and Error](#)

What is your next step?

Beginner

Advanced

A great place to start is:

Get yourself an ORCID id. This allows people to find and reach you easily. You can even include it in your email signature.

Wrap-up

Where to from here?

For each of our 7 lessons, do we have an action item?

Let's establish concrete action items for the upcoming week or two based on the tasks we've just documented.



CHALLENGE

Take this time to do one of the following:

Email yourself a website you've found today to read through later.

Set a calendar entry to dedicate some time to complete a task.

Pair with someone here to encourage each other.

Talk to your colleagues about a suggested change to make in your project.

Even by only taking small steps, you are now further down the reproducible research path.

Wrap-up

Thank you for attending the Reproducible Research workshop!

Feedback



https://docs.google.com/forms/d/e/1FAIpQLScHfогV C8K_tjIMG9D-O9F07_CdgK02ZHyE3_7WoPrFrM_ExQ/viewform?usp=publish-editor

Feedback frameworks

2x2

	Content	Delivery
Thing you liked		
Thing that could be improved		

One up, one down

- Go around the group and alternate between a positive and negative feedback that hasn't been said yet