



# PSDI

PHYSICAL SCIENCES  
DATA INFRASTRUCTURE

## Promoting Open & Transparent Research Practices in the Physical Sciences through PSDI

Keele University Open Research Network  
10<sup>th</sup> April 2024

Dr Samantha Pearman-Kanza & Dr Nicola Knight  
University of Southampton

<https://www.psdi.ac.uk/>

# Presentation Outline

- ▶ About Us & PSDI
- ▶ A Grand Vision of Process Recording
- ▶ Current Barriers & Challenges to Open & Transparent Digital Research
- ▶ Producing FAIR Data & Research
- ▶ PSDI Initiatives to help with this



# About Us

## Dr Samantha Pearman-Kanza

- ▶ Senior Enterprise Fellow at University of Southampton
- ▶ Pathfinder Lead & Researcher for PSDI
- ▶ Research Interests: Semantic Web Technologies, RDM, Metadata, Process Recording, Interoperability
- ▶ Twitter: @samikanza



## Dr Nicola Knight

- ▶ Senior Enterprise Fellow at University of Southampton
- ▶ Project Coordinator for PSDI
- ▶ Research Interests: Chemical data management, connected lab technologies, RDM
- ▶ Twitter: @njkknight



**PSDI**  
PHYSICAL SCIENCES  
DATA INFRASTRUCTURE

# About PSDI

Through PSDI researchers will be able to:

- ▶ Find and Access to reference quality data from commercial and open sources
- ▶ Combine data from different sources
- ▶ Share data, software and models including experimental and simulation data
- ▶ Use AI to explore data
- ▶ Learn how to make the results of their research open and FAIR

## Physical Sciences Data Infrastructure

An Integrated Data Infrastructure for the Physical Sciences

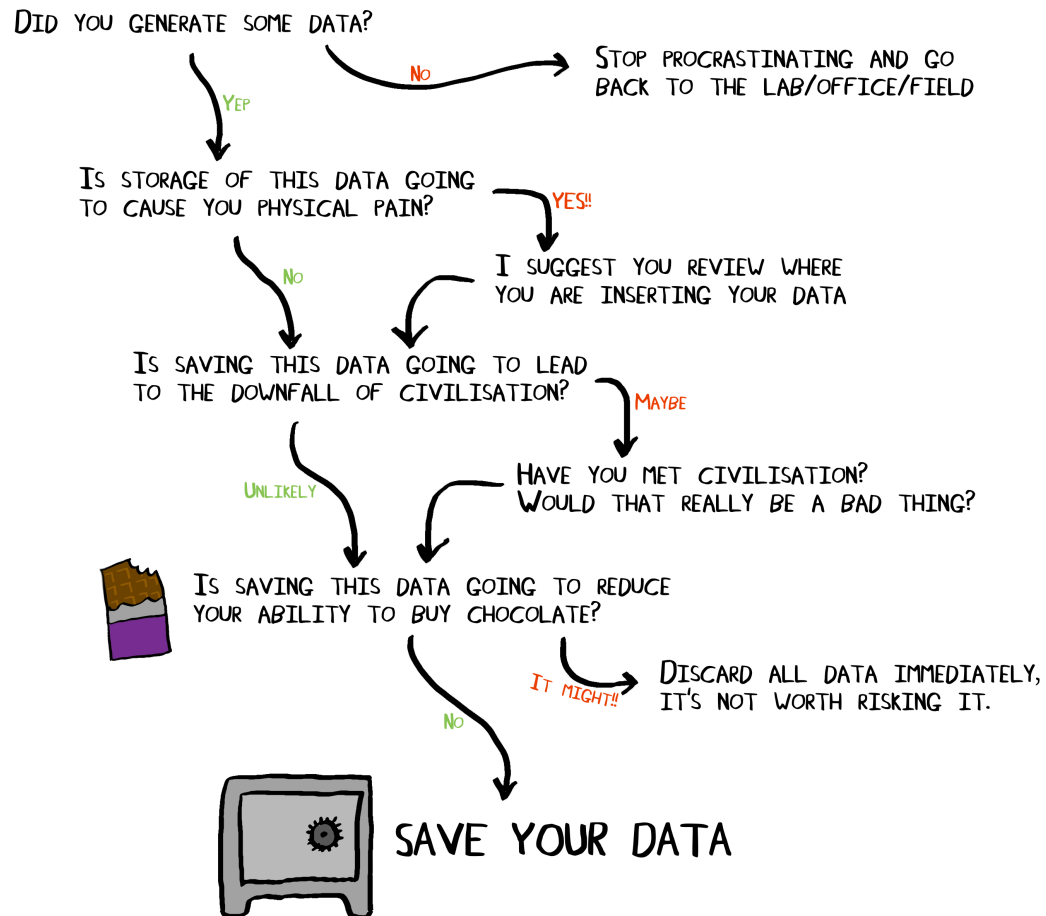
PSDI aims to accelerate research in the physical sciences by providing a data infrastructure that brings together and builds upon the various data systems researchers currently use.

# A grand vision

## SHOULD YOU SAVE YOUR DATA?

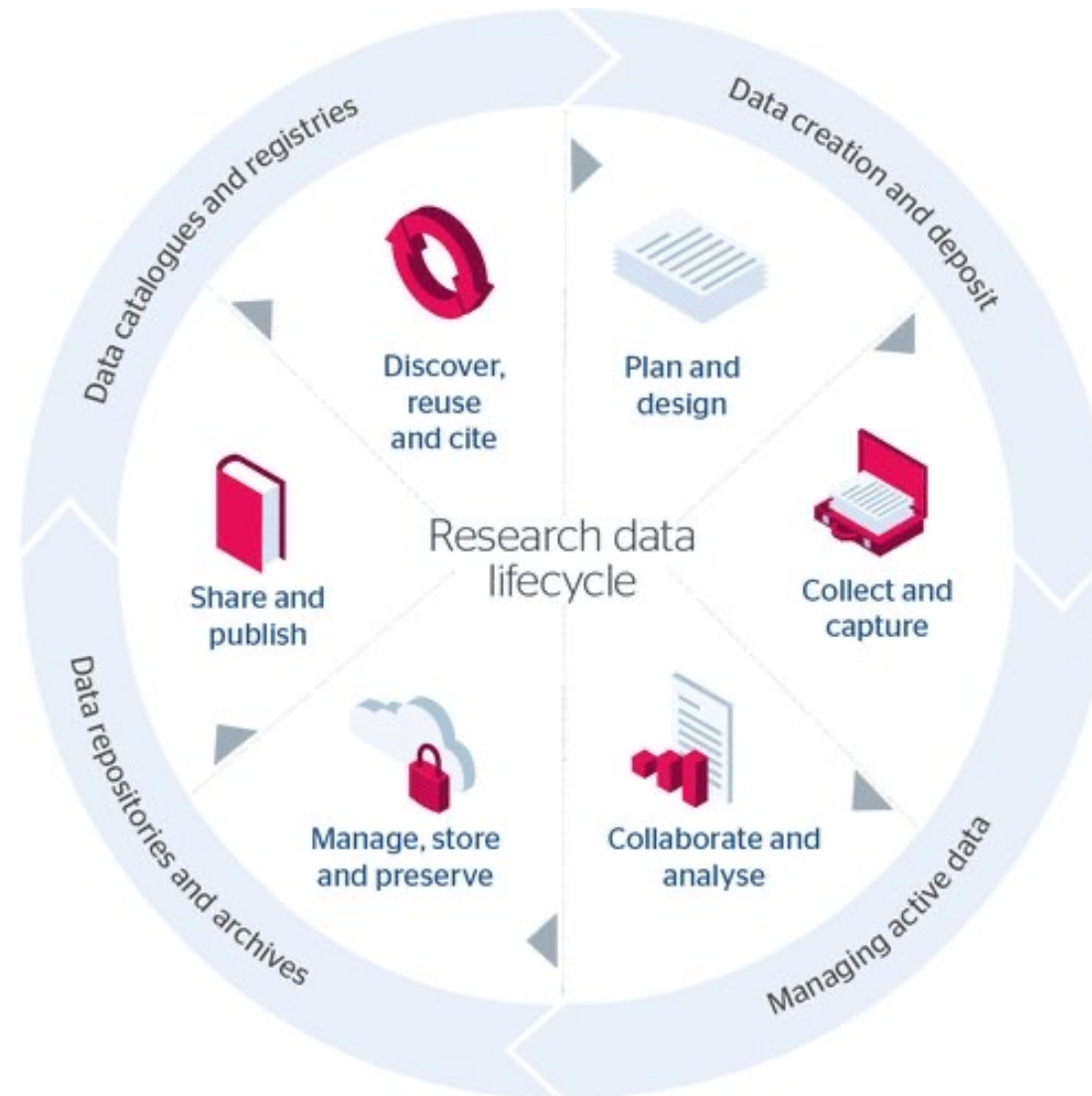
### A FLOW DIAGRAM

ERRANTSCIENCE.COM



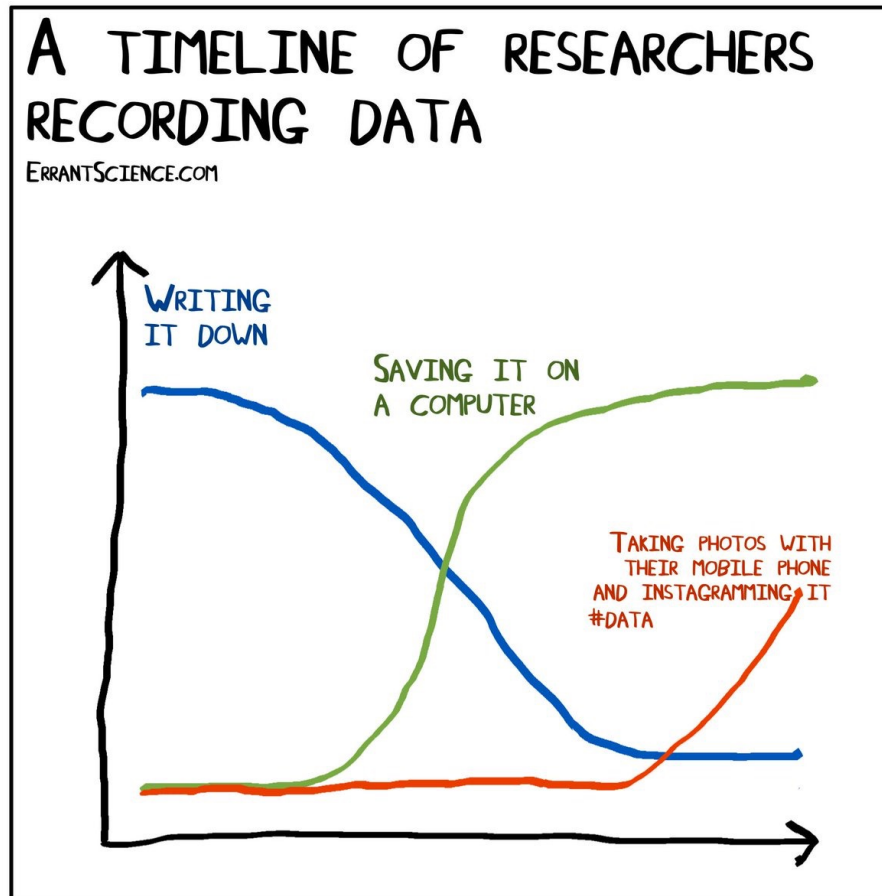
A wonderful world of researchers capturing and sharing all their data, code, and methods in a re-useable way

# Research Data Lifecycle





# Process Recording



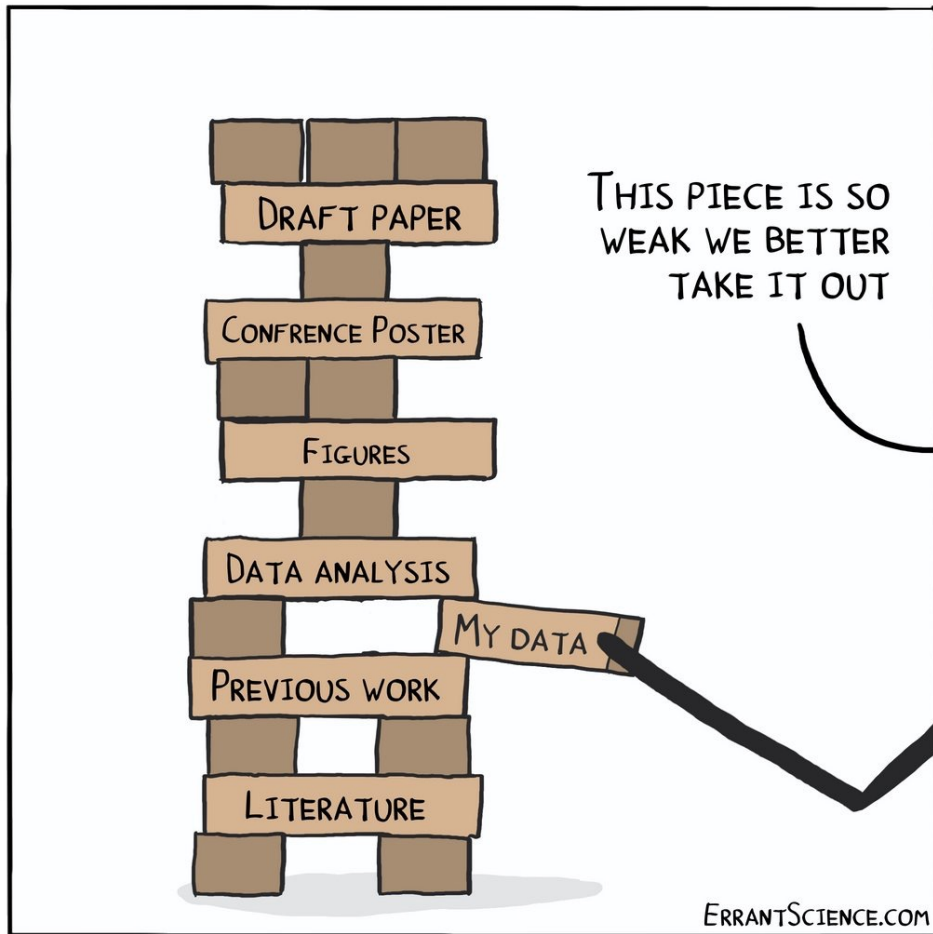
CC BY-ND 4.0 Errant Science - <https://errantscience.com/>

- ▶ It is vital that we capture our data and processes throughout our research so that:
  - ▶ Our future selves can find it and understand it
  - ▶ We can share it alongside our publications and others can **ACTUALLY USE IT**

**So why haven't we achieved this already?**



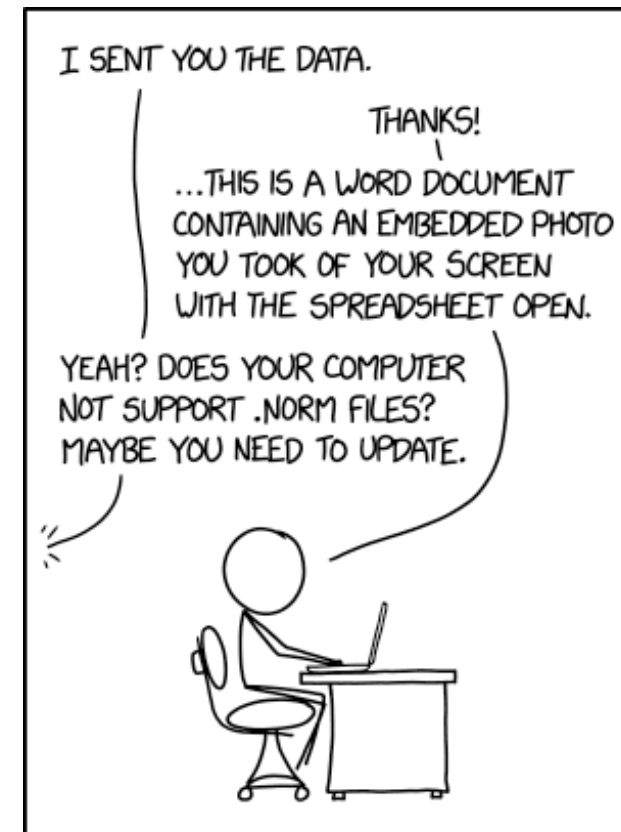
# Current Barriers & Challenges to Open and Transparent Digital Research



- ▶ Data
- ▶ Standards
- ▶ Software
- ▶ Hardware
- ▶ Cost
- ▶ Time
- ▶ Trust
- ▶ People / Adoption

# People/Adoption Barriers

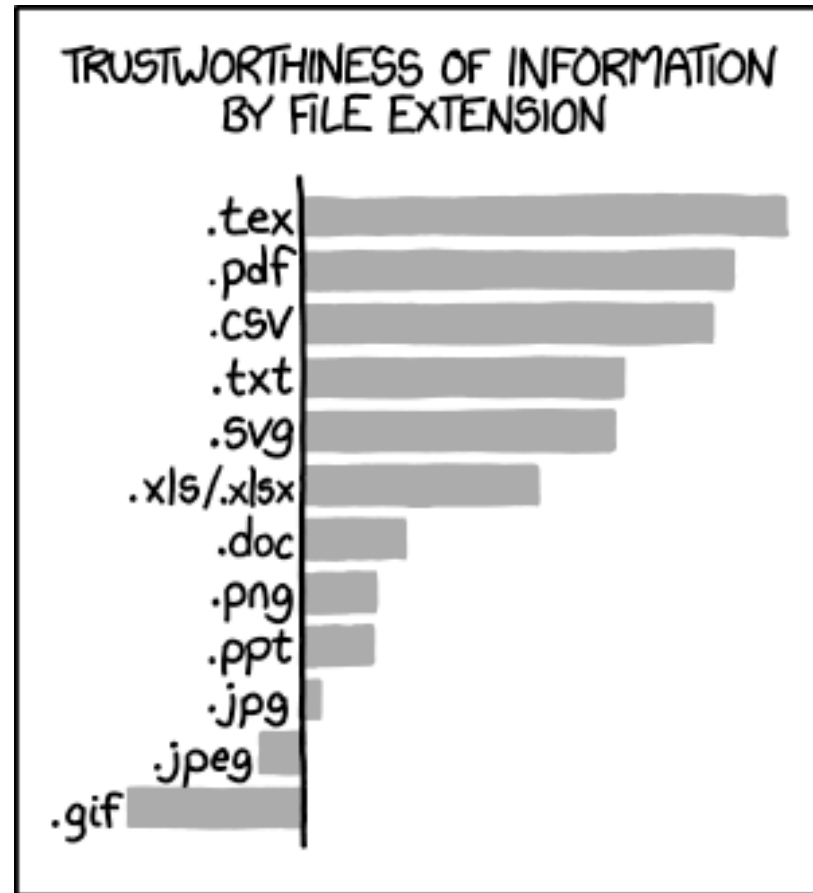
- ▶ People are arguably one of the biggest barriers
- ▶ Top-down influence can make or break this
- ▶ Concerns about changing processes
- ▶ Hard to persuade people to embark on a journey with a lot of front-loaded work, unless they really understand the benefits



SINCE EVERYONE SENDS STUFF THIS  
WAY ANYWAY, WE SHOULD JUST  
FORMALIZE IT AS A STANDARD.

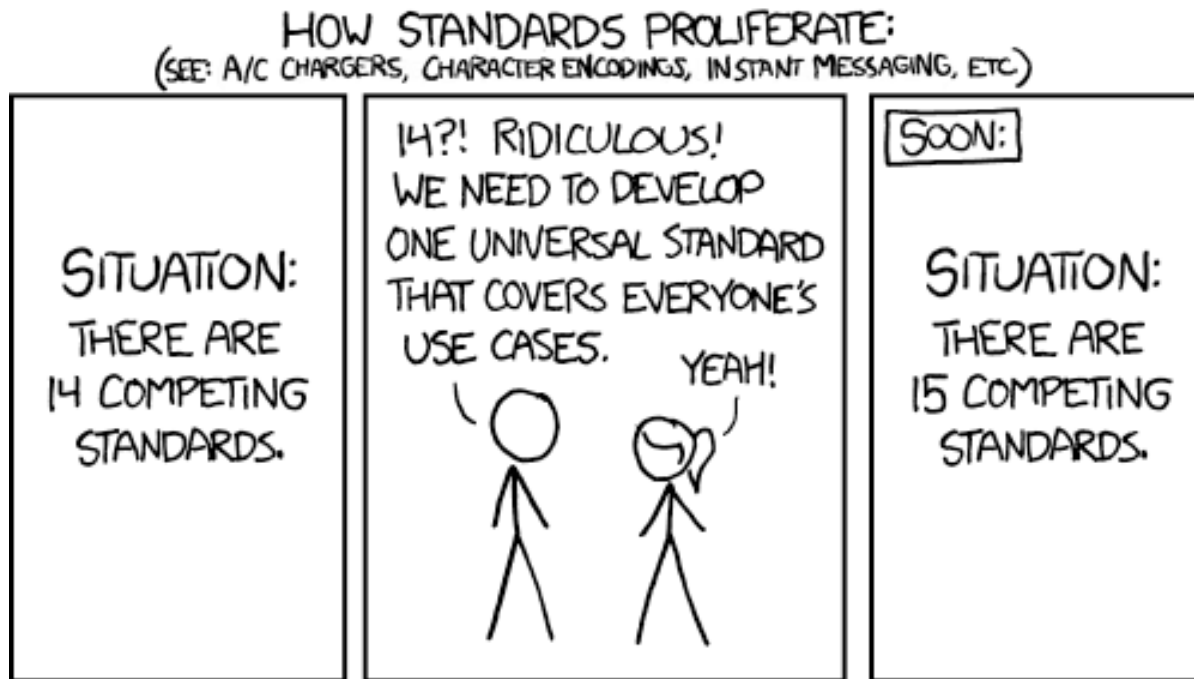
# Data Barriers

- ▶ **Un-FAIR Data**
  - ▶ Much data doesn't adhere to FAIR standards. Data isn't findable, accessible, interoperable or re-useable
- ▶ **Metadata/Provenance**
  - ▶ Data often lacks context
  - ▶ Time consuming to capture metadata
  - ▶ Leads to not being able to trace the provenance of research
- ▶ **Data Size**
  - ▶ Scientists frequently work with large datasets that are harder to store and share



"File Extensions" by XKCD is licensed under [CC BY-NC 2.5](https://creativecommons.org/licenses/by-nc/2.5/)

# Standards Barriers

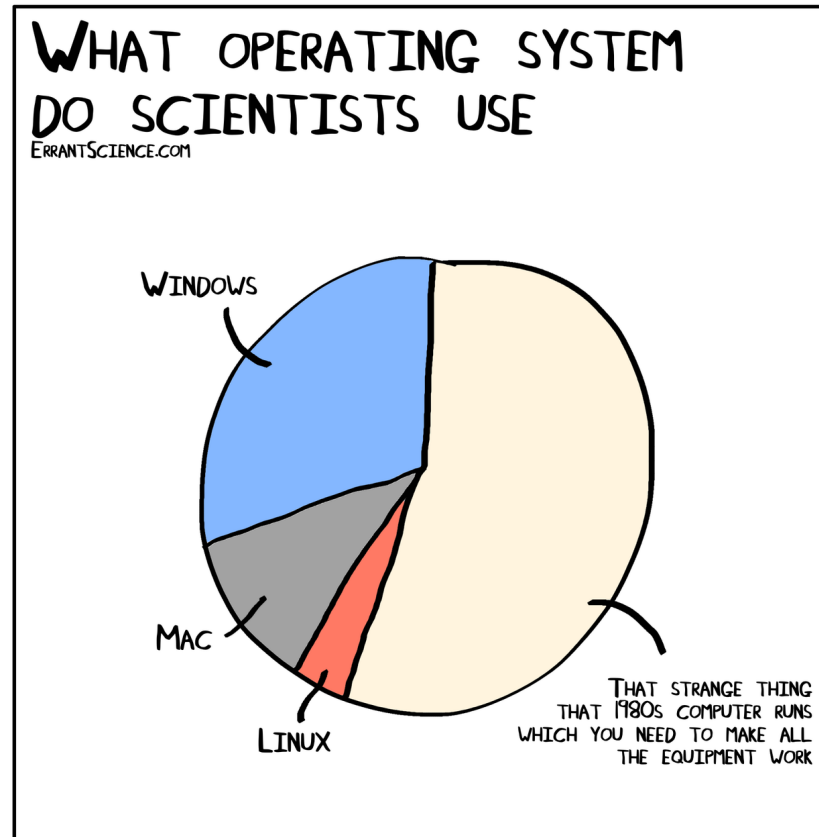


"Standards" by XKCD is licensed under [CC BY-NC 2.5](https://creativecommons.org/licenses/by-nc/2.5/)

- ▶ Too many Standards
  - ▶ We are drowning in standards, and yet still lack them in many areas
- ▶ Proprietary Formats
  - ▶ Lots of software uses proprietary formats that won't work with other software to lock vendors in
- ▶ Lack of Interoperability
  - ▶ There are many formats that won't work across multiple pieces of software

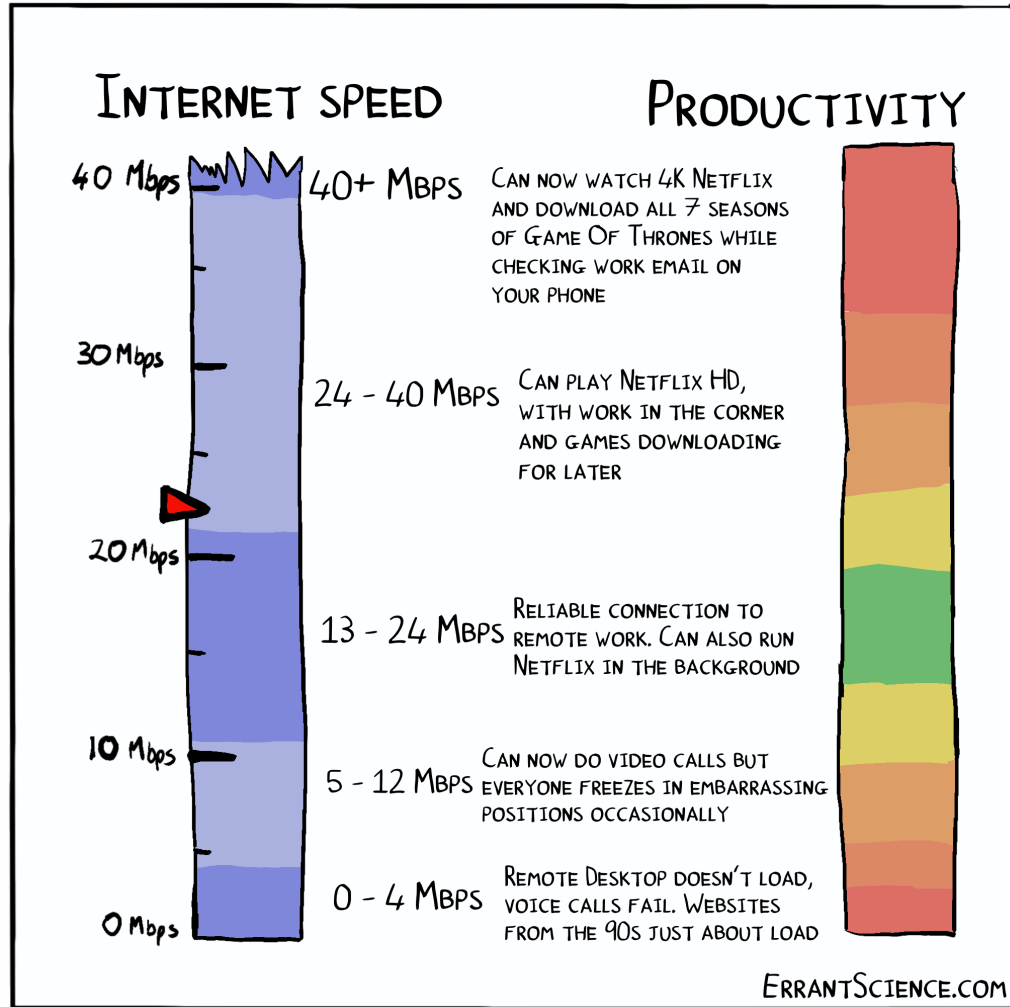
# Software Barriers

- ▶ **Software Overload**
  - ▶ There are so many different pieces of software for capturing data and research digitally, how on earth can we choose?
- ▶ **Software Compatibility**
  - ▶ Software often doesn't play well together, which makes having an overarching digital ecosystem challenging
- ▶ **Software Quality**
  - ▶ Software often isn't at a high enough quality for users to want to engage with it
- ▶ **Online Software**
  - ▶ Online Software is great, until the internet stops working



"Scientists Computer Use" by ErrantScience.com is licensed under [CC-BY-NC](#)

# Hardware Barriers



CC BY-ND 4.0 Errant Science - <https://errantscience.com/>

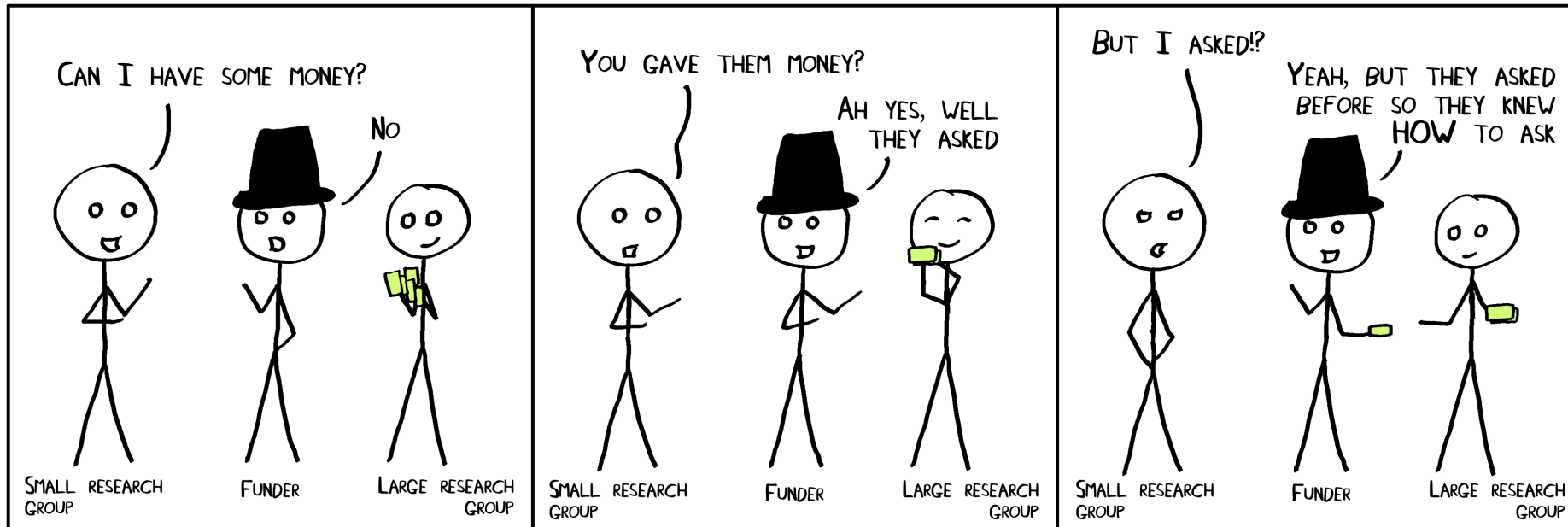
- ▶ Data Storage Capabilities
  - ▶ Many scientists say they do not have sufficient data storage capabilities
- ▶ Clunky Hardware
  - ▶ Often researchers struggle to gain access to hardware that will run their digital tools well - which then leads to them not wanting to use them at all
- ▶ Legacy Equipment
  - ▶ Many laboratories use legacy equipment which requires legacy software and outdated data formats
- ▶ Hardware Cross Contamination
  - ▶ Where digital tools are required in the lab, we need dedicated hardware to run these tools otherwise there are contamination risks moving computers/laptops in and out of the lab

# Cost Barriers

- ▶ Cost
  - ▶ Funding, Research, Software, Hardware, Publishing

## UNIVERSITY FUNDING IN A NUTSHELL

ERRANTSCIENCE.COM

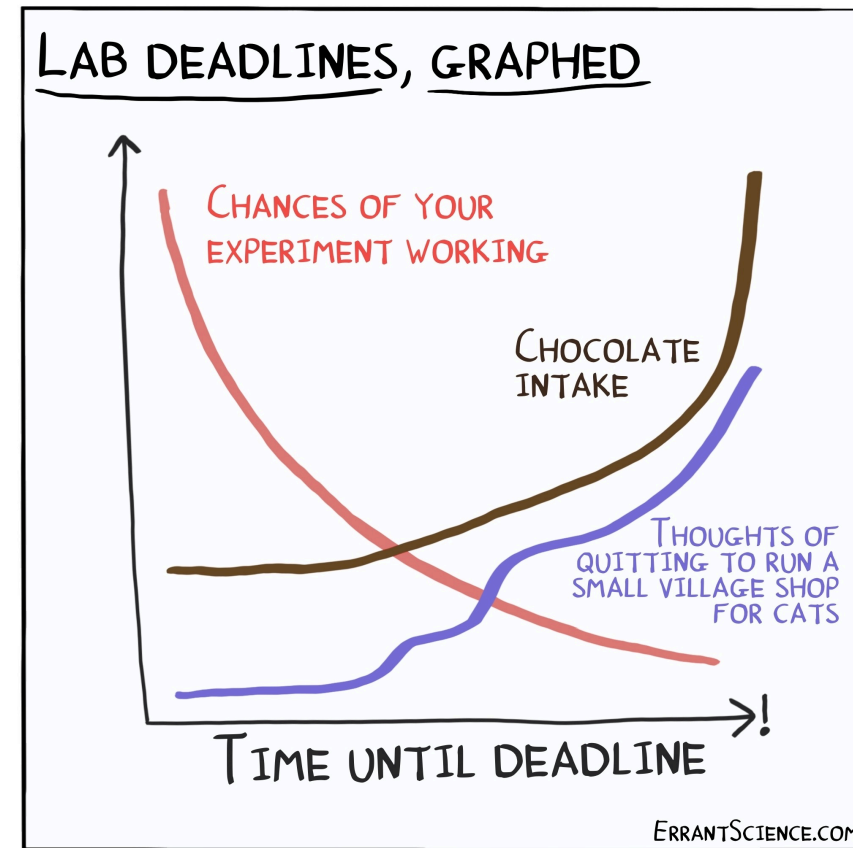


"University Funding in a Nutshell" by ErrantScience.com is licensed under [CC-BY-NC](https://creativecommons.org/licenses/by-nc/4.0/)

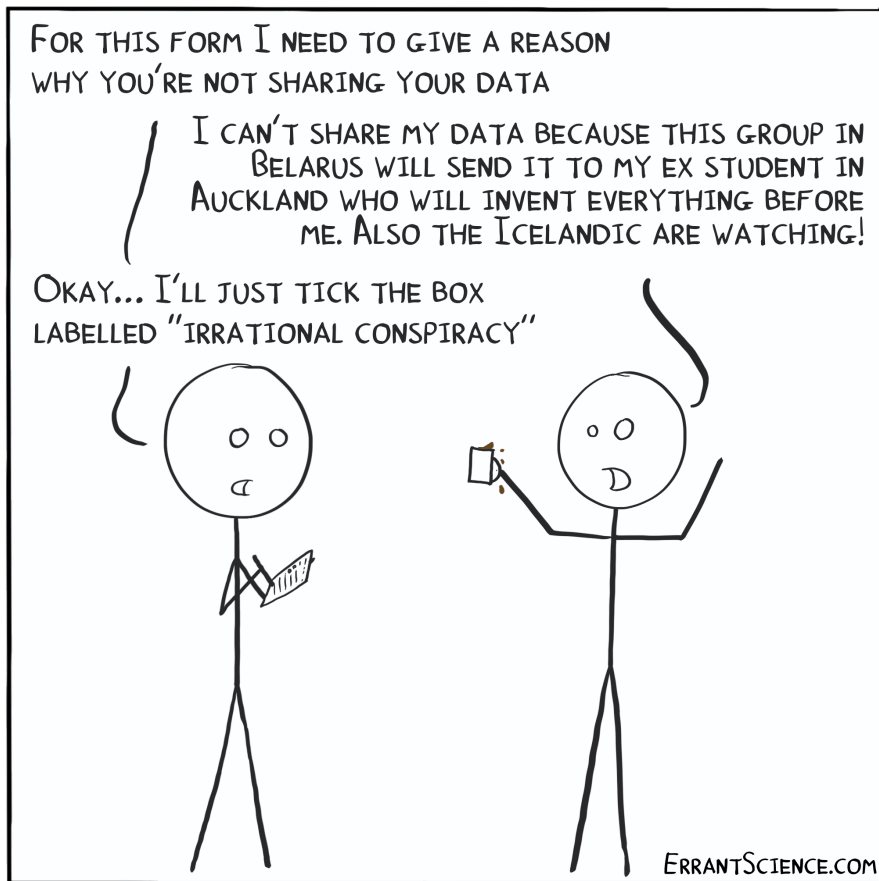


# Time Barriers

- ▶ Lack of time for projects
- ▶ Time taken to learn and use new systems
- ▶ Time to do research data management **PROPERLY**
- ▶ Current systems in place make digitizing data and capturing research in a transparent way very time consuming



# Trust Barriers



CC BY-ND 4.0 Errant Science - <https://errantscience.com/>

- ▶ Many researchers do not trust storing the data/research online
- ▶ There are many concerns to consider:
  - ▶ Data privacy (Sharing/Hacking)
  - ▶ Software using proprietary formats
  - ▶ Lack of cohesive data exit strategy

**So how do we overcome these barriers?**

# Lets talk about FAIR

From 'The FAIR Guiding Principles for scientific data management and stewardship'<sup>1</sup>

- ▶ F – Findable
- ▶ A – Accessible
- ▶ I – Interoperable
- ▶ R – Reusable

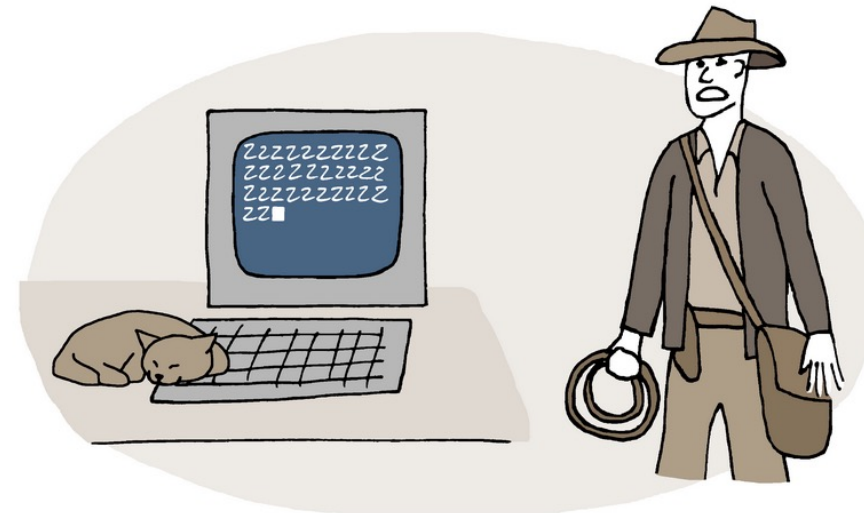


Image created using imgflip.com

<sup>1</sup> Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>

# F is for Findable

- ▶ To be Findable:
  - ▶ It needs to exist
  - ▶ But existing != findable
  - ▶ Provide your users with pointers!
  
- ▶ Are all your code/data/lab book/notes actually there?



FINALLY! AFTER ALL THOSE YEARS  
I FINALLY FOUND  
THE SOURCE OF THE DATA!

 Dataedo /cartoon

Piotr@Dataedo

[CC BY-ND 3.0 Dataedo - https://dataedo.com/cartoon/](https://dataedo.com/cartoon/)

# A is for Accessible

- ▶ What should and shouldn't be accessible?
- ▶ What is the use case?
- ▶ If access is restricted or complex, have you provided relevant information?



 Dataedo /cartoon

*PioloDataedo*

[CC BY-ND 3.0 Dataedo - https://dataedo.com/cartoon/](https://dataedo.com/cartoon/)

**Technically accessible != Easily accessible**

# I is for Interoperable

- ▶ Consider your data standards
- ▶ Use Common and Shared Vocabularies
  - ▶ For Data and Metadata
- ▶ Use Ontologies/Knowledge Graphs to the best of their potential



Michael J. Swart <https://michaeljswart.com/2011/06/meta-aggregate/>

**Even standards need standards**



# R is for Re-useable

- ▶ This isn't JUST about the data! You need to consider:
  - ▶ Data, Tools, Code, Methods, Context
  - ▶ How could/would your work be re-used, replicated, reproduced or repurposed
    - ▶ Re-use – re-use the data (or run the software) in the same manner
    - ▶ Replicate – repeat entire research from scratch including data collection and analysis
    - ▶ Reproduce – reanalyse the existing data in the same manner
    - ▶ Repurpose – use existing data or software for a new purpose

This is only  
the tip of the  
“R” Iceberg



# FAIR Details

## Data

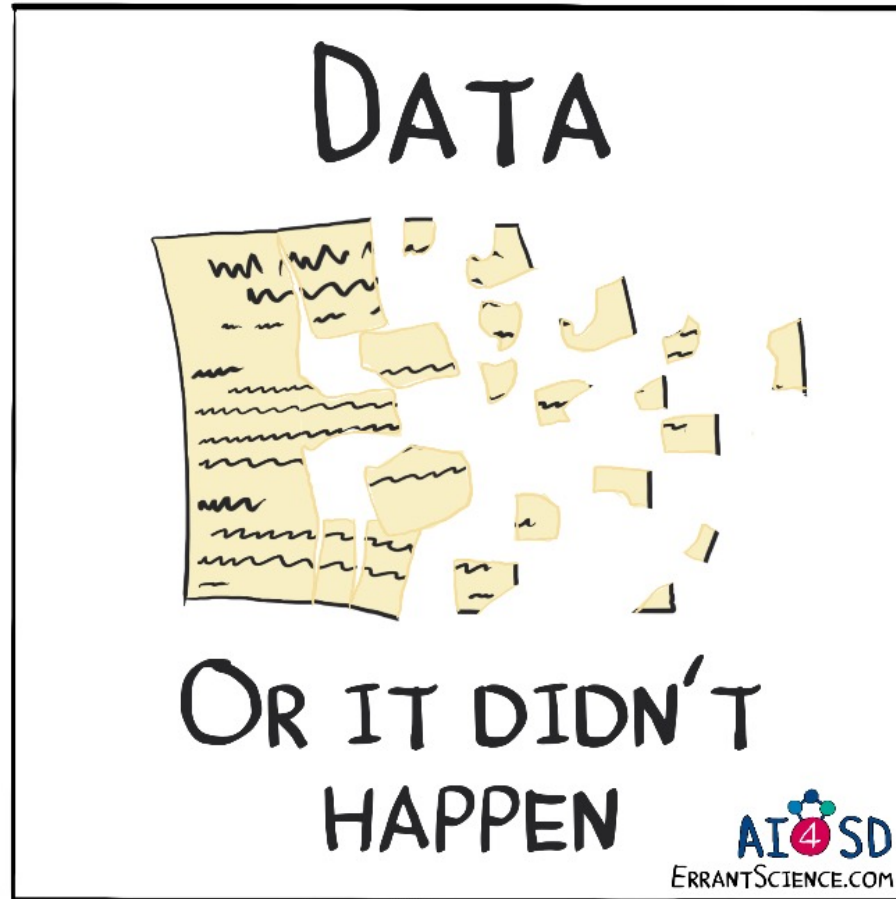
- ▶ Do your data file names make sense
- ▶ Do your data headings make sense?
- ▶ Are your files understandable?

## Code

- ▶ Do your code files make sense
- ▶ Is your code all there?
- ▶ Is it commented?

## Lab Books

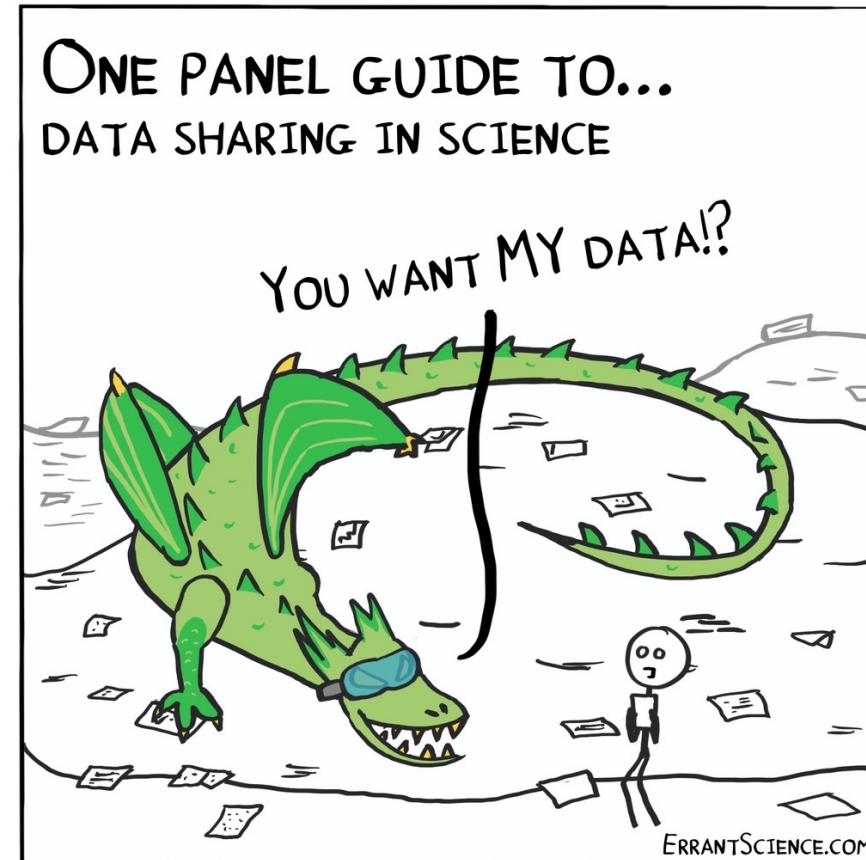
- ▶ Does your lab book fully detail your reagents, samples, experiment parameters?



[CC BY-ND 4.0 Errant Science - https://errantscience.com/](https://errantscience.com/)

# FAIR Pre-requisites

- ▶ Performing any of our 'R' operations on data or software is complex
- ▶ Data
  - ▶ Is this stored on outdated media?
  - ▶ What tools/software/dependencies do we need to use the data
- ▶ Databases
  - ▶ How do we use these? Are there database dumps? Schemas? Instructions?
- ▶ Software
  - ▶ What coding libraries are required?
  - ▶ Are there dependencies?
  - ▶ What installations and drivers are required?
  - ▶ Is all the underlying data included and accessible
- ▶ Lab Books
  - ▶ What were the experimental conditions?
  - ▶ What was the experimental setup?
  - ▶ What context exists for the experiment that you haven't recorded?

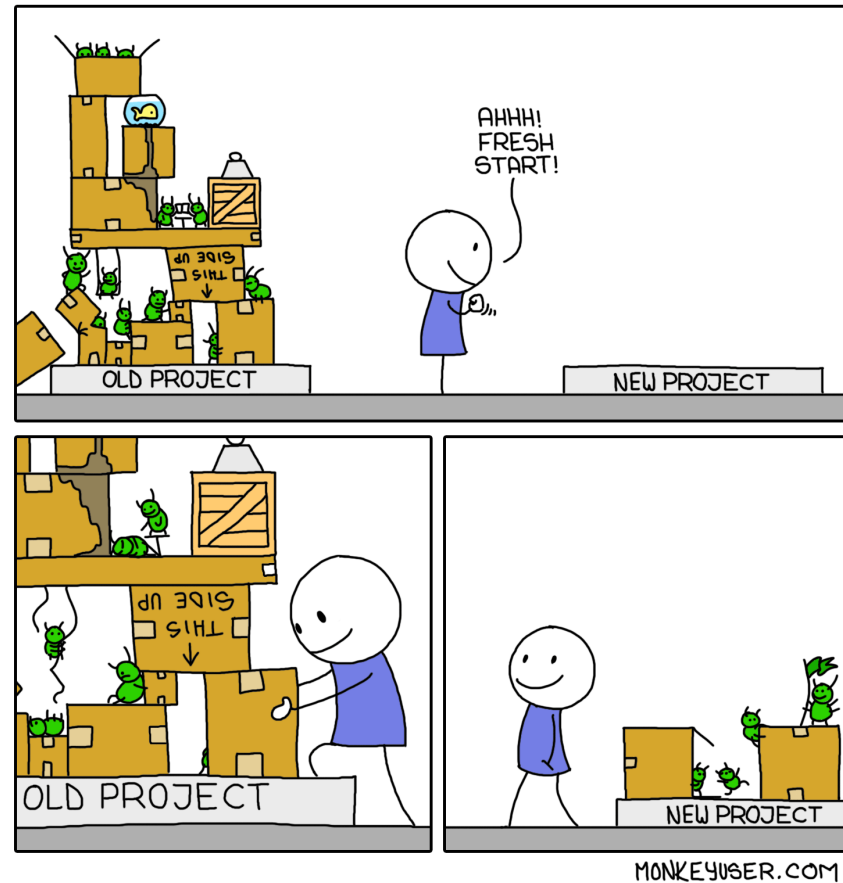


[CC BY-ND 4.0 Errant Science - https://errantscience.com/](https://errantscience.com/)

# FAIR Instructions

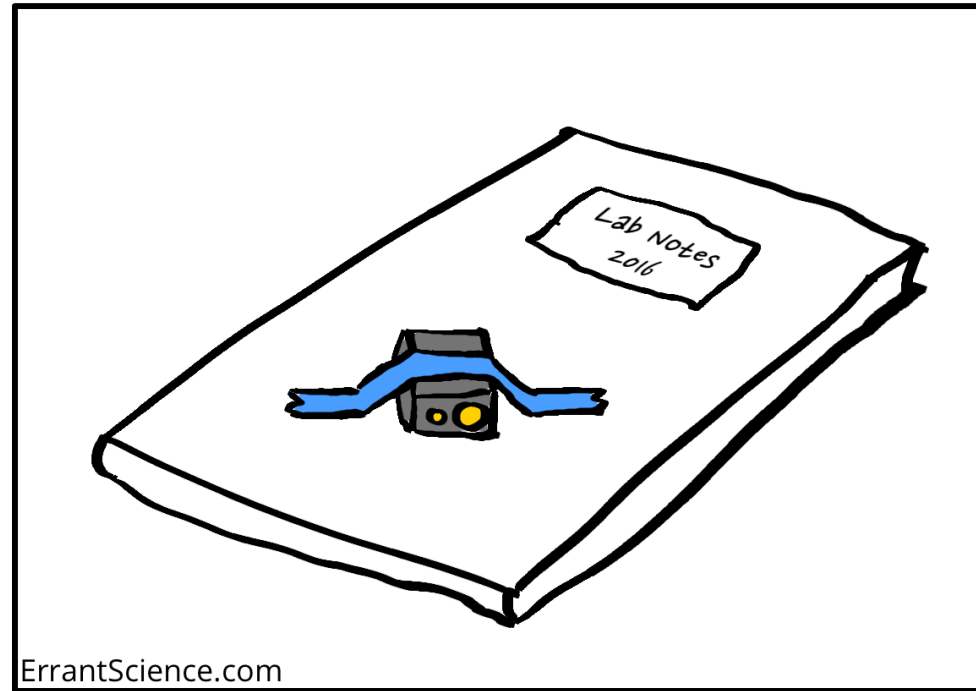
- ▶ Be clear
- ▶ Do not assume prior knowledge
- ▶ Include all steps from start to finish (which means documenting as you go along)
- ▶ How was the data collected?
- ▶ What scripts/parameters were used?
- ▶ How did you get your database to interface with your code?
- ▶ How do you access the data?
- ▶ How do you run the software locally?
- ▶ If someone had your lab book and all your data could they re-run your experiment?
- ▶ Could someone else really re-use, reproduce, replicate or repurpose this?

## CODE REUSE



# FAIR Considerations for Selecting Digital Tools

- ▶ What data are you currently recording?
- ▶ How are you recording it?
- ▶ Where are you recording it?
- ▶ Is there extra data that you should be capturing?
- ▶ Who needs to access the data?
- ▶ What tools are people already using, and why?



If your electronic lab book looks like this,  
you're doing it wrong

[CC BY-ND 4.0 Errant Science - https://errantscience.com/](https://errantscience.com/)



# Further FAIR Considerations

- ▶ Investment through people and finance are imperative
- ▶ We are the problem so we can be the solution
- ▶ Investing in the relevant hardware and software tools will make these processes easier
- ▶ Should be considered early in the project, not just at the end

"ALL RESEARCH SHOULD AIM TO BE F.A.I.R."

#FIGSHAREFEST

	GOOD	BAD
<b>F</b> INDABLE	ONLINE DATABASE	FILING CABINET IN A BATH IN THE BASEMENT UNDER A LEAKING PIPE
<b>A</b> CCESSABLE	OPEN ACCESS FOR EVERYONE (NO LOGIN)	THE FILING CABINET ALSO IS HOME TO A NEST OF WILD BADGERS
<b>I</b> NTEROPERABLE	ALL DATA IS IN OPEN FORMATS	ALL DOCUMENTS ARE PRINTED IN COMIC SANS AND WRITTEN IN ESPERANTO
<b>R</b> EUSEABLE	GOOD META DATA AND SECURELY STORED FOR 10 YEARS	THE PAPER EXPLODES IF IT'S READ

ERRANTSCIENCE.COM

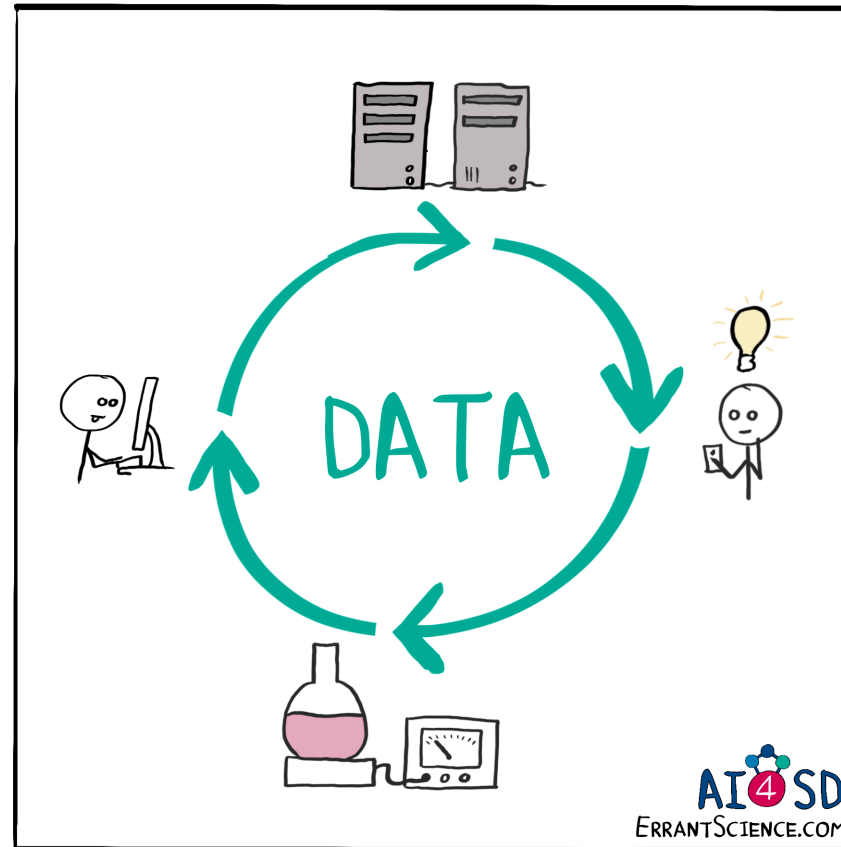
**What is PSDI working on to help this?**



# PSDI Focus Areas

Our current (but growing) focus is on:

- ▶ Process Recording
- ▶ Metadata
- ▶ DMPs
- ▶ Access to trusted data resources



[CC BY-ND 4.0 Errant Science - https://errantscience.com/](https://errantscience.com/)

Domain Exemplars:

- ▶ Biomolecular simulation
- ▶ Catalysis
- ▶ Machine learnt interatomic-potentials
- ▶ Materials
- ▶ NMR Crystallography


# Tools to help researchers (in development)

- ▶ Recording workflows for reproducibility
  - ▶ Developing tools in both AiiDA and Galaxy to enable easy and thorough recording of the steps taken in computational processes
- ▶ Data Conversion
  - ▶ Lookup and conversion between data formats to enable interoperability
- ▶ Data Revival
  - ▶ Scan in paper lab books and get data back in a machine-readable form
- ▶ Metadata Generation
  - ▶ Generate semantically rich metadata records, template READ-ME's, license files
- ▶ Toolkits for creating better structured data / databases

# Training & Guidance

- ▶ DMPs
- ▶ FAIR data publication
- ▶ Skills4Scientists
  - ▶ Technical skills
  - ▶ Soft skills
- ▶ Metadata creation
- ▶ Tool selection (e.g. picking the right process recording tool, implementing tools in teaching / research environments)
- ▶ Software guides (for PSDI developed tools)

DATA PUBLISHING	
GOOD	BAD
DATA REPOSITORY	STICKY NOTE ON YOUR DOOR
INSTITUTIONAL ARCHIVE	SUPPLEMENTARY DATA
	BOTTOM OF A WELL

  
ERRANTSCIENCE.COM

# How to get involved

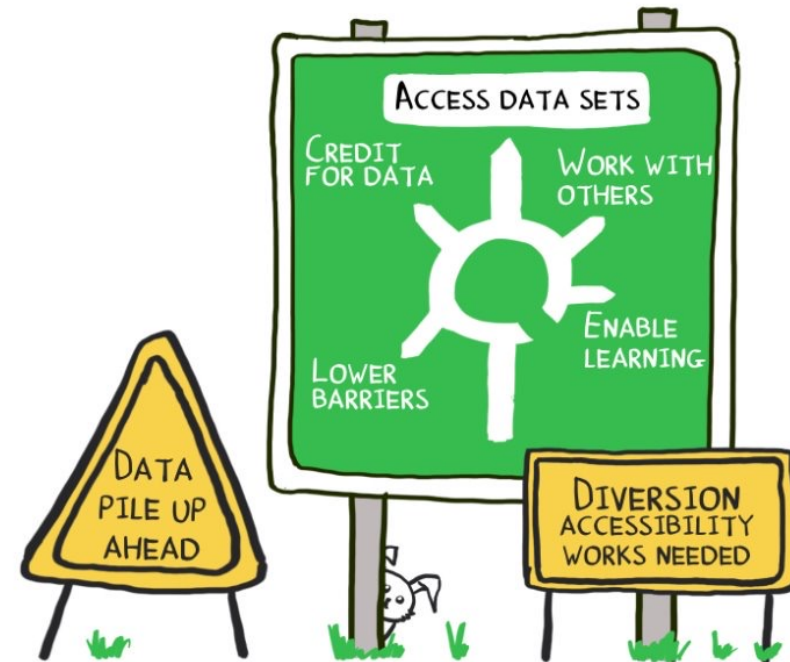
- ▶ User focus groups
- ▶ Come to our webinars
- ▶ Send us an email

## Townhall

- ▶ Add the other things here...

# Conclusions

- ▶ There are still many barriers to overcome
- ▶ But PSDI is working towards solutions
- ▶ We need to remember the following:
  - ▶ Ask the right questions, about your data, your tools, your situation
  - ▶ FAIR is a FOUR letter word, but it has many many nuances
  - ▶ Collaboration is key - This is as much a human endeavor as a software/data one
  - ▶ We must all strive to be better



[CC BY-ND 4.0 Errant Science - https://errantscience.com/](https://errantscience.com/)

**To the well organised FAIR dataset, re-use, replication, reproduction and repurpose are but the next great adventure**

# Relevant Talks

- ▶ Kanza, S. (2022, June 7). The effects of COVID-19 on the digitisation of Scientific Research - Presentation at Future Labs Live 2022. Future Labs Live 2022 (FLL2022), Basel. Zenodo. <https://doi.org/10.5281/zenodo.10118139>
- ▶ Kanza, S. (2022, October 4). To Digitisation And Beyond! The Digitisation Requirements Of A 21st Century Scientist - Presentation at Drug Discovery World 2022. Drug Discovery World 2022 (DDW2022), London. Zenodo. <https://doi.org/10.5281/zenodo.10142544>
- ▶ Kanza, S. (2022, December 6). Technical and Data Requirements of Digitalising Scientific Research - Presentation at Smart Labs & Automation 2022. Smart Labs & Automation, London. Zenodo. <https://doi.org/10.5281/zenodo.10142749>
- ▶ Kanza, S. (2023, January 25). The Digitisation of Scientific Research: Requirements, Barriers and Logistics - Presentation at Lab of the Future 2023. Lab of the Future 2023, Online. Zenodo. <https://doi.org/10.5281/zenodo.10142604>
- ▶ Kanza, S. & Knight, N. (2023, March 29). Process recording and digitisation requirements for the 21st century scientist - Presentation for ACS Spring 2023. ACS SPRING 2023 Crossroads of Chemistry (ACS SPRING 2023), Indianapolis, IN & Hybrid. Zenodo. <https://doi.org/10.5281/zenodo.10144147>
- ▶ Kanza, S. (2023, May 31). ELNs are Dead! Long Live ELNs! - Presentation at Future Labs Live 2023. Future Labs Live 2023 (FLL2023), Basel. Zenodo. <https://doi.org/10.5281/zenodo.10138225>
- ▶ Kanza, S. (2023, August 13). We don't talk about Semantic Web Technologies - Presentation at ACS Fall 2023. ACS FALL 2023 Harnessing the Power of Data (ACS FALL 2023), San Francisco, CA & Hybrid. Zenodo. <https://doi.org/10.5281/zenodo.10149599>
- ▶ Kanza, S. (2023, August 14). Electronic Lab Notebooks and Beyond! The evolution of process recording tools for scientific research - Presentation at ACS Fall 2023. ACS FALL 2023 Harnessing the Power of Data (ACS FALL 2023). Zenodo. <https://doi.org/10.5281/zenodo.10149499>
- ▶ Pearman-Kanza, S. (2023, November 1). To the well organised FAIR dataset, re-use is but the next great adventure - Presentation at Lab Innovations 2023. Lab Innovations 2023, NEC, Birmingham. Zenodo. <https://doi.org/10.5281/zenodo.10119611>
- ▶ Pearman-Kanza, S. (2023, December 7). How can we combat heterogeneous, unfair and disparate data in digital chemistry? Presentation at the ChemSpider Webinar Series. ChemSpider Webinar Series: Challenges and opportunities for digital chemistry data, Online. Zenodo. <https://doi.org/10.5281/zenodo.10417786>
- ▶ Pearman-Kanza, S. (2024, March 13) Electronic Lab Notebooks and Beyond! The evolution of process recording tools for scientific research'. RSC Historical Group Open Meeting, Zenodo <https://doi.org/10.5281/zenodo.10818945>

# Relevant Publications

- ▶ Kanza, S., Willoughby, C., Gibbins, N., Whitby, R., Frey, J.G., Erjavec, J., Zupančič, K., Hren, M. and Kovač, K., 2017. Electronic lab notebooks: can they replace paper?. *Journal of cheminformatics*, 9(1), p.31. <https://doi.org/10.1186/s13321-017-0221-3>
- ▶ Kanza, S., 2018. What influence would a cloud based semantic laboratory notebook have on the digitisation and management of scientific research? (Doctoral dissertation, University of Southampton). <https://eprints.soton.ac.uk/421045/>
- ▶ Kanza, S., Gibbins, N. and Frey, J.G., 2019. Too many tags spoil the metadata: investigating the knowledge management of scientific research with semantic web technologies. *Journal of cheminformatics*, 11(1), p.23. <https://doi.org/10.1186/s13321-019-0345-8>
- ▶ Knight, N.J., Kanza, S., Cruickshank, D., Brocklesby, W.S. and Frey, J.G., 2020. Talk2Lab: The Smart Lab of the Future. *IEEE Internet of Things Journal*, 7(9), pp.8631-8640. <https://doi.org/10.1109/JIOT.2020.2995323>
- ▶ Kanza, S., Willoughby, C., Bird, C.L. and Frey, J.G., 2021. eScience Infrastructures in Physical Chemistry. *Annual review of physical chemistry*, 73. <https://doi.org/10.1146/annurev-physchem-082120-041521>
- ▶ Kanza, S., 2021. Guidelines for Chemistry Labs Looking to Go Digital. *Digital Transformation of the Laboratory: A Practical Guide to the Connected Lab*, pp.191-197. <https://doi.org/10.1002/9783527825042.ch13>
- ▶ Kanza, S., 2021. Understanding and Defining the Academic Chemical Laboratory's Requirements: Approach and Scope of Digitalization Needed. *Digital Transformation of the Laboratory: A Practical Guide to the Connected Lab*, pp.179-189. <https://doi.org/10.1002/9783527825042.ch12>
- ▶ Kanza, S., 2021. Academic's Perspective on the Vision About the Technology Trends in the Next 5–10 Years. *Digital Transformation of the Laboratory: A Practical Guide to the Connected Lab*, pp.297-301. <https://doi.org/10.1002/9783527825042.ch22>
- ▶ Kanza, S. and Knight, N.J., 2022. Behind every great research project is great data management. *BMC Research Notes*, 15(1), pp.1-5. <https://doi.org/10.1186/s13104-022-05908-5>
- ▶ Kanza, S., Willoughby, C., Knight, N.J., Bird, C.L., Frey, J.G. and Coles, S.J., 2023. Digital research environments: a requirements analysis. *Digital Discovery*. <https://doi.org/10.1039/D2DD00121G>



# Acknowledgements



**PSDI Team:** Simon Coles, Jeremy Frey, Nicola Knight, Cerys Willoughby, Colin Bird, Ray Whorley, Mark Anderson, Stephen Gow, Samuel Munday, Thomas Allam, Hannah Gittins (*University of Southampton*), Juan Bicarregui, Barbara Montanari, Brian Matthews, Vasily Bunakov (*Science & Technologies Facilities Council*)



# PSDI

PHYSICAL SCIENCES  
DATA INFRASTRUCTURE



[www.psdi.ac.uk](http://www.psdi.ac.uk)



PSDIUK

X @PSDI\_UK



@psdi-uk.bsky.social



@PSDI\_UK

<https://www.jiscmail.ac.uk/PSDI>