PSDI

PHYSICAL SCIENCES DATA INFRASTRUCTURE

# How can we combat heterogeneous, unFAIR and disparate data in digital chemistry?

ChemSpider Webinar 3: Challenges & Opportunities

7th December 2023

Dr Samantha Pearman-Kanza
University of Southampton

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

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

# About Me & PSDI

- Senior Enterprise Fellow at University of Southampton
- Pathfinder Lead & Researcher for PSDI Project: Process Recording
- Research Interests: Semantic Web Technologies, IoT, Research Data Management, Digitisation, Lab of the Future, Paperless Labs, Re-use of Technology
- @SamiKanza

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

## Physical Sciences Data Infrastructure

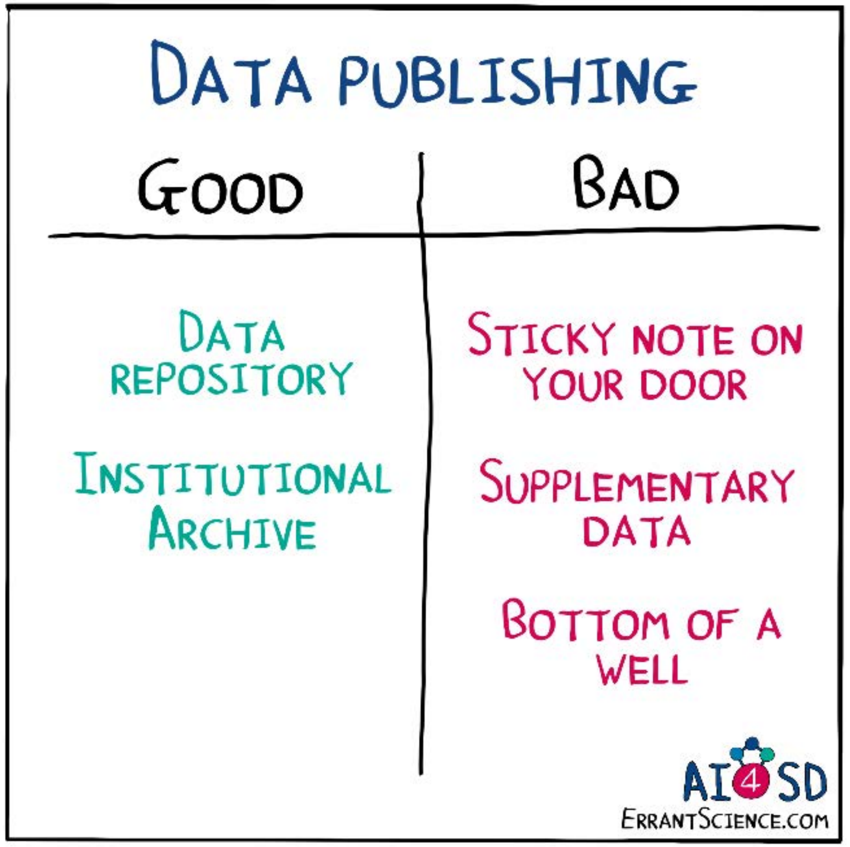
### An Integrated Data Infrastructure for the Physical Sciences

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.

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

# How can we combat heterogeneous, unFAIR and disparate data in Chemistry?

- Understand the environment and the challenges
  - Barriers & Challenges to Digitisation
- Process Recording
  - Digitisation Requirements
  - Choosing your tools for process recording
- Producing FAIR Data AND Research AND Code
  - Considering all aspects of FAIR and going beyond the guidelines
  - Establish common vocabularies and practices (data and metadata)



DATA PUBLISHING

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

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

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

# Barriers & Challenges to Digital Research

- Logistical Barriers
  - Cost
  - Time
- People Barriers
  - Attitude & Adoption Factors
  - Training
- Data Barriers
  - Un-FAIR Data
  - Metadata/Provenance
  - Size of data
- Standards Barriers
  - Too Many Standards
  - Proprietary formats
- Software Barriers
  - Oversaturated Market for ELNs, Notebooks & Domain Based Software
  - Software Integration/Compatibility
  - Trust in Software
- Hardware Barriers
  - Data Storage
  - Legacy Equipment

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

# What do Users want from ELNs?

## Notebooking Features

- Alternative input methods (voice/handwriting/text recognition)
- Searching/Tagging/Indexing
- Colour Coding/ Personalisation
- Links with reference management software
- Collaboration features

## Domain Specific Features

- Integration with Chemical Equipment
- Integration with Chemical Data
- Attach and view characterization data in ELN directly
- Setup for multiple domains

## Data Features

- Data Management features
- Version Control
- Linking between records
- Archiving old data
- Store structured data
- Flexible data export/data portability

## Technical/Logistical Features

- Integration with Hybrid Devices
- API Access
- More Storage
- Open Source / Development Capabilities
- Cost

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

# What do Users want from Notebooks?

## Notebooking Features

- Alternative input methods (voice/handwriting/text recognition)
- Create/Use Templates
- Add schemas/diagrams/images
- Searching/Tagging/Indexing
- Collaboration features
- “Be just like paper”
- Integrate with Project Management Software (ToDo lists/Gantt Charts)

## Domain Specific Features

- Interface with Chemical Structure Editor/have features inbuilt
- Pasting Chemdraw Structures
- Integrate with ELN

## Data Features

- Linking between records
- Flexible data export/data portability
- Excel features to work with data/plot graphs
- Link to external data sources

## Technical/Logistical Features

- Mobile Support
- Interoperability between devices
- Speed
- Cost

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

# What do Users Want from a Digital Research Environment?

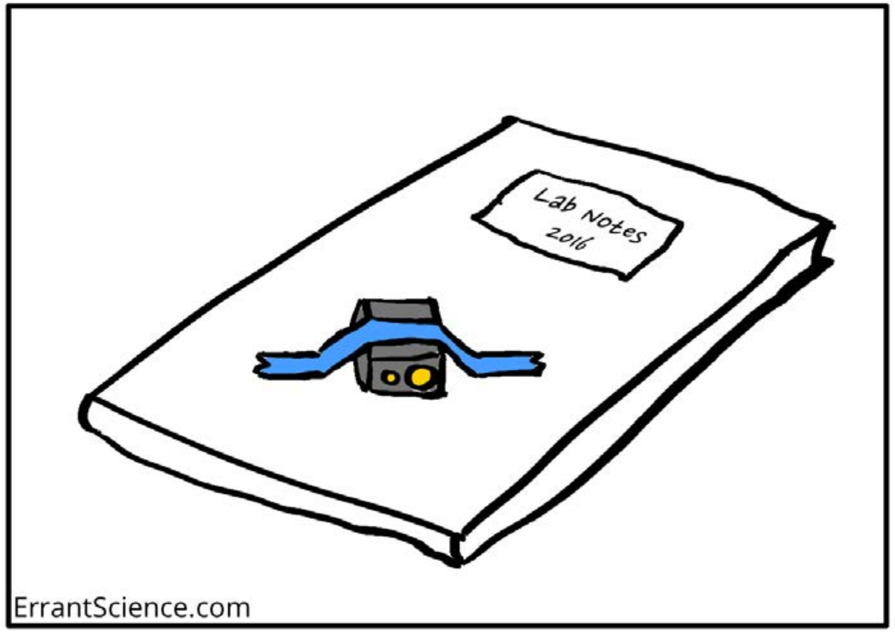
| Feature Category | Description |
|---|---|
| Generic | **API Access**, Automation, GUI, Localisation, Remote Access, Synchronisation |
| Notebooking | **Content Support,** Interaction/Access, File Links, Organisation/Reconfiguration, **Paper Integration, Referencing/ Literature,** Word Processing |
| Data | Access, Conversion, Exchange, **Integration, Management**, Quality, Retention, Security, **Standards**, Support, FAIR, Identifiers, Provenance |
| Publishing & Sharing | Documentation & Instructions, DOIs, Export, Licensing, Open Access, Publishing, Sharing, Social Media, Researcher Attribution, Repositories |
| Collaboration & Management | **Auditing, Comments, Notifications, Subscribe, Team Management** |
| Domain Based Features | **Chemical/Molecules**, Default Lists, **Equipment Interface**, Experiment Planning/Recording, Health & Safety, **LIMS/ELN**, **Link to Domain based databases & software** |
| Coding Support | Coding, **Versioning** |
| Metadata, Semantics & AI | AI Tools/Integration, Metadata, Semantics |
| Searching | **Search By: Domain, Characteristics Search, Keyword/Concept via Content Types, Literature & Notebook, Indexing** |
| Customisation & Extension | Personalisable, **Templates** |
| Training & User Support | Training, User Documentation |

● ELN Features ● Notebook Features ● Both

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

# Choosing tools and methods for Process Recording?

- What data are you recording?
- How are you recording it?
- Where are you recording it?
- What data is not being recorded?
- What are the pain points?
- What is the actual problem you are trying to fix?

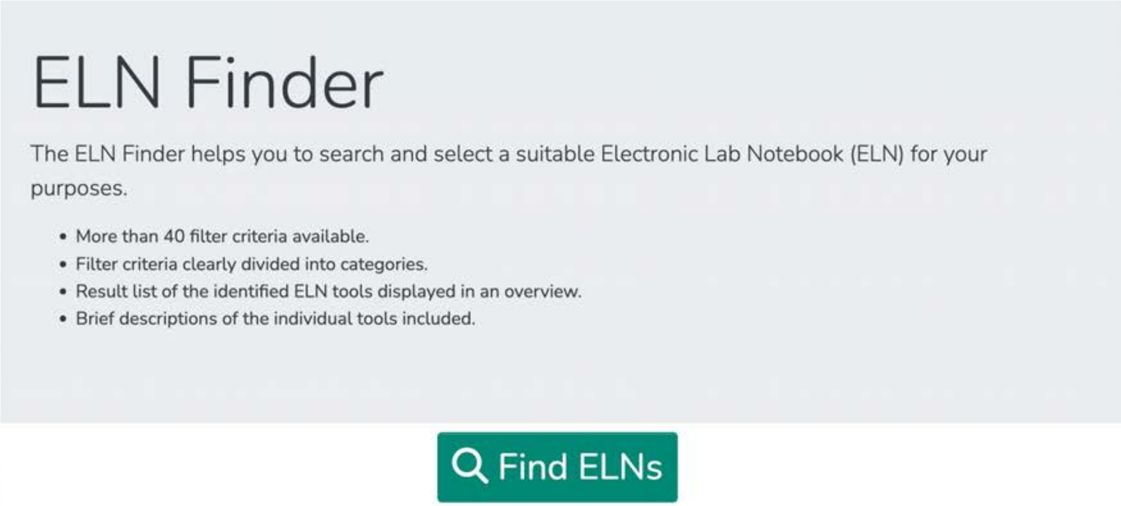


If your electronic lab book looks like this, you're doing it wrong

"Electronic Lab Notebooks" by ErrantScience.com is licensed under CC-BY-NC

# ELN Finder

## https://eln-finder.ulb.tu-darmstadt.de



# ELN Finder

The ELN Finder helps you to search and select a suitable Electronic Lab Notebook (ELN) for your purposes.

- More than 40 filter criteria available.
- Filter criteria clearly divided into categories.
- Result list of the identified ELN tools displayed in an overview.
- Brief descriptions of the individual tools included.

- Detailed hierarchical criteria catalogue created, defines and describes the metadata structure for the ELNs (Excel):
- > 40 criteria and associated values, attributes (e.g. name/URL).
- Summary of criteria in categories
- Fully functional first version developed on the basis of the open source software DSpace 7:
- External ELN information collection created for individual ELNs
- Entering data from the information collection
- 35 ELNs entered

## Filter Criteria

- APIs
- Automation
- Collaboration
- Compliance
- Controlled vocabulary
- Customizable user interface
- Data access
- Data export
- Data import (formats)
- Data import (method)
- Data input
- Data storage location
- Device connection
- Laboratory management functions
- Languages  Support
- License
- Location of provider
- Offline functionalities
- Operating system
- Plug-Ins
- Preservation of evidence
- Pricing
- Project management tools
- Search functions
- Standard interfaces
- Subject
- Templates
- Usage option
- Usage statistics
- Versions
- Workflows

# Lets talk about FAIR

From ‘The FAIR Guiding Principles for scientific data management and stewardship’[1]

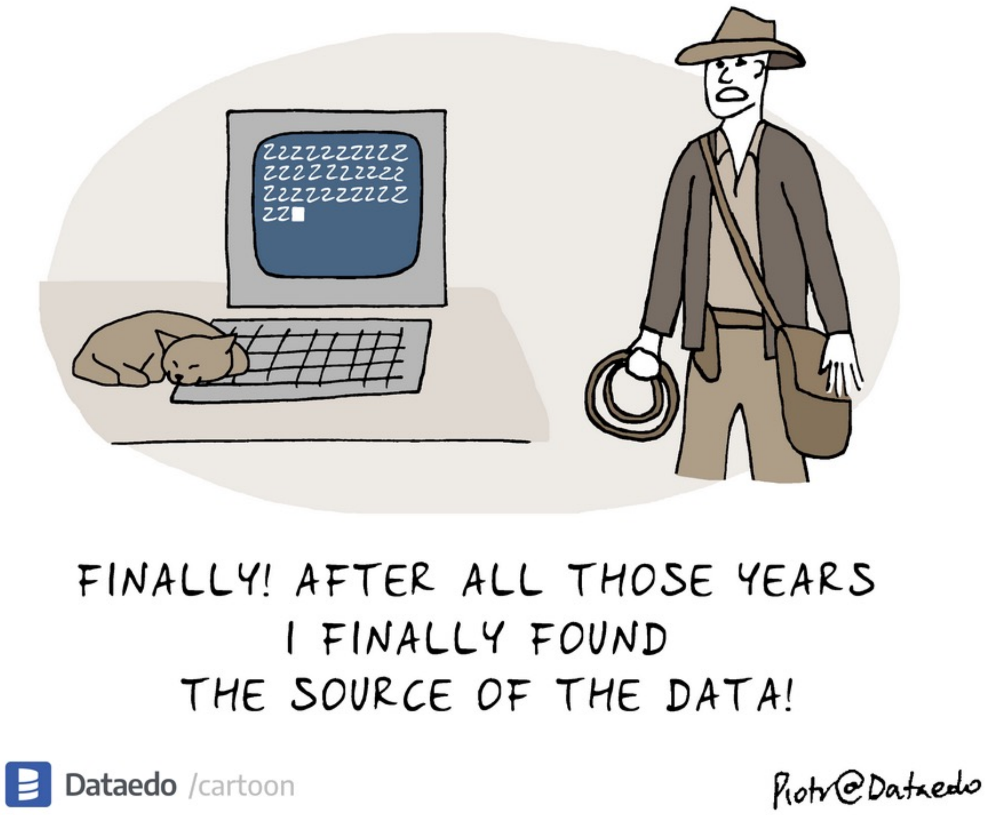
- F – Findable
- A – Accessible
- I – Interoperable
- R – Reusable

Image created using imgflip.com

[1] 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?**



CC BY-ND 3.0 Dataedo - https://dataedo.com/cartoon/

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

# 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?

CC BY-ND 3.0 Dataedo - https://dataedo.com/cartoon/

Technically accessible != Easily accessible

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

# 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

https://www.pinterest.co.uk/jaci_mize/metadata/

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

CARTOONSTOCK
Search ID: shrn5456
J. Harris
"OF COURSE YOU CAN'T REPLICATE MY EXPERIMENTS. THAT'S THE BEAUTY OF THEM."

https://www.cartoonstock.com/directory/s/scientific_method.asp

## 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?

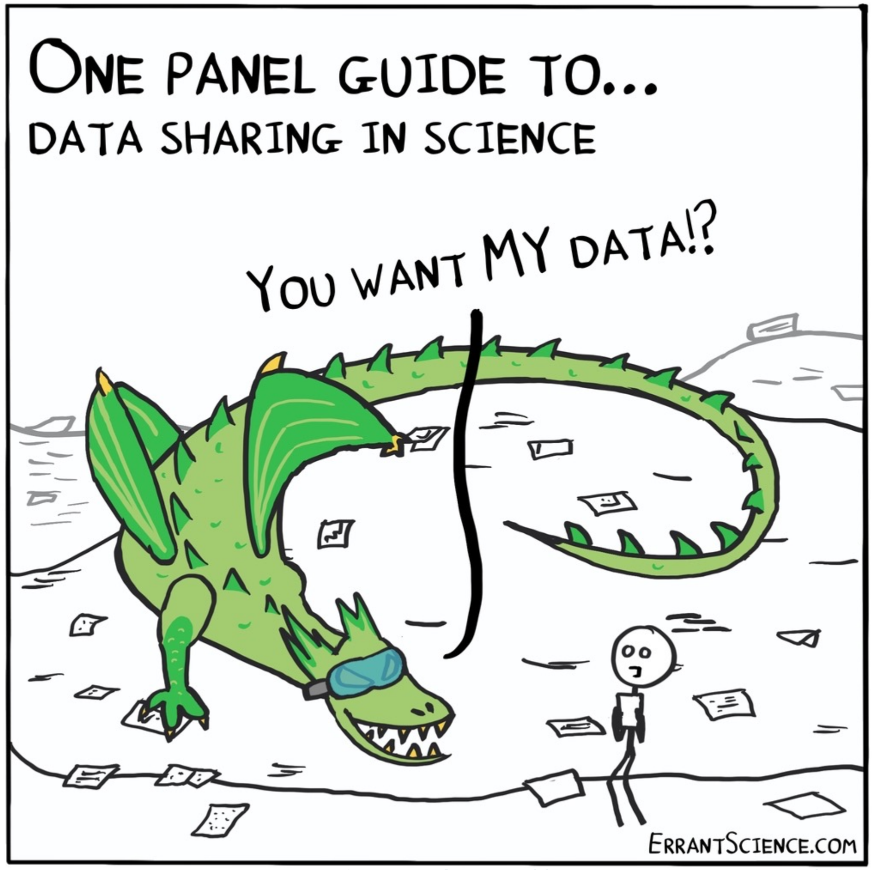
## Lab Books

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

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

# FAIR Pre-requisites

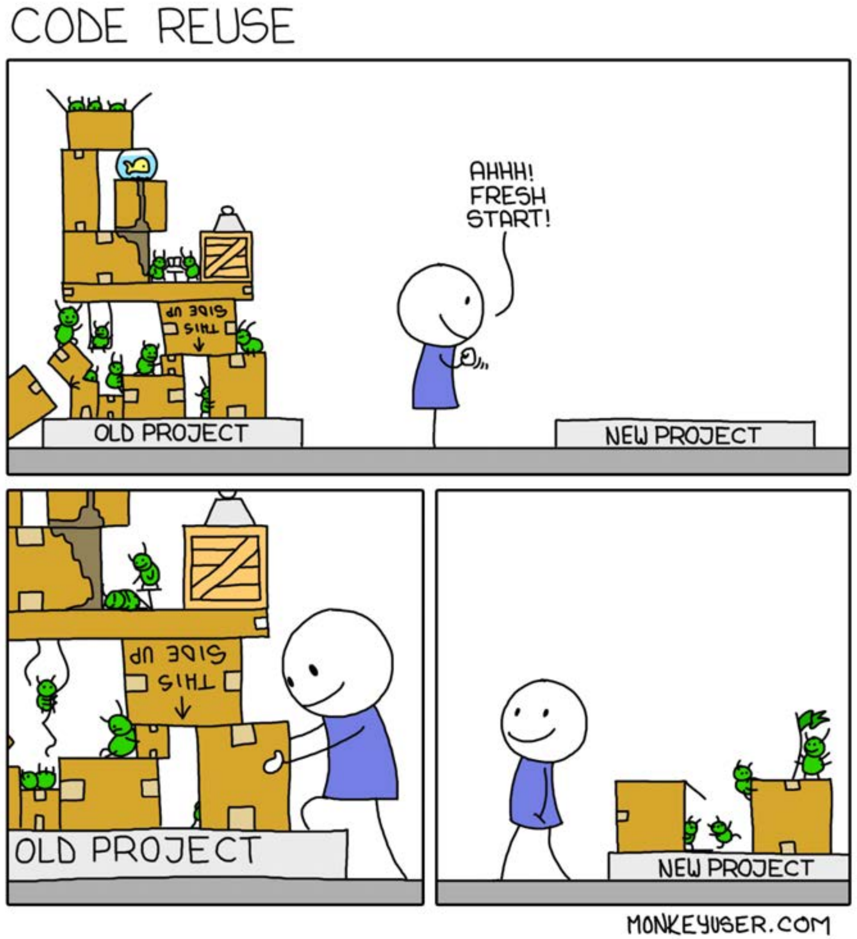
- Performing any of our ’R’ operations on data of software is complex
- Data
  - Is this stored on outdated media?
  - What tools/software/dependencies do we need to use the data
- Databases:
  - How do we use these? Are there database dumps? Schemas? Instructions?
- Software:
  - What coding libraries 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/

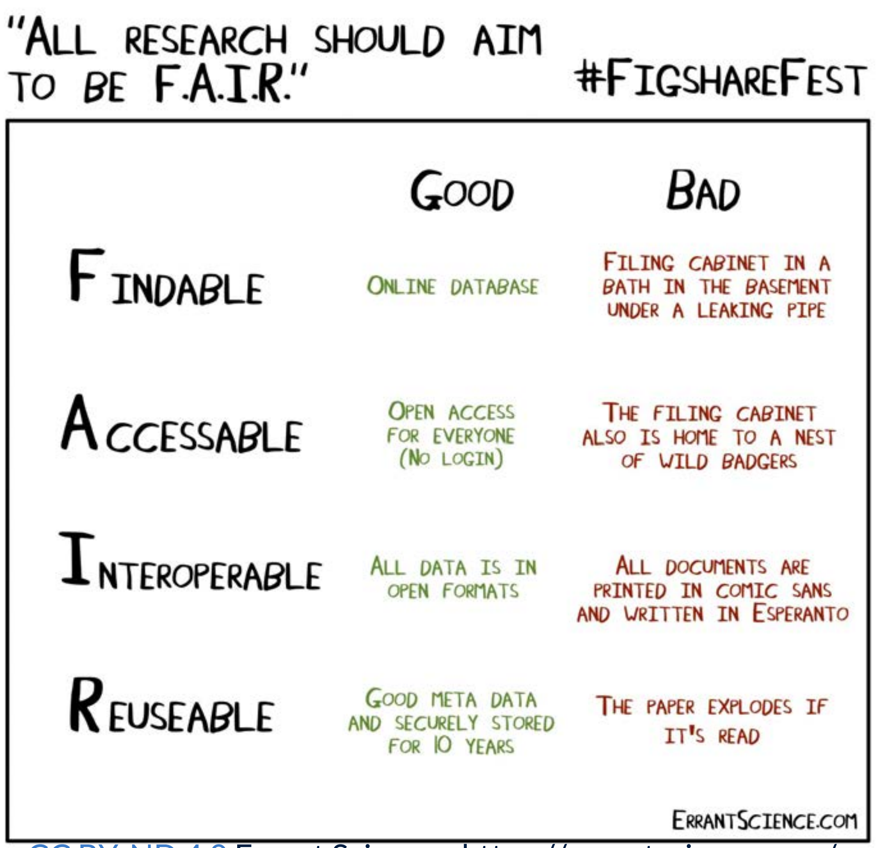
# FAIR Instructions

- 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?



# Conclusions

- There are still many barriers to overcome
- But the community 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



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

| | GOOD | BAD |
|---|---|---|
| FINDABLE | ONLINE DATABASE | FILING CABINET IN A BATH IN THE BASEMENT UNDER A LEAKING PIPE |
| ACCESSABLE | OPEN ACCESS FOR EVERYONE (NO LOGIN) | THE FILING CABINET ALSO IS HOME TO A NEST OF WILD BADGERS |
| INTEROPERABLE | ALL DATA IS IN OPEN FORMATS | ALL DOCUMENTS ARE PRINTED IN COMIC SANS AND WRITTEN IN ESPERANTO |
| REUSEABLE | GOOD META DATA AND SECURELY STORED FOR 10 YEARS | THE PAPER EXPLODES IF IT'S READ |

ERRANTSCIENCE.COM

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

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

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

# Relevant Talks

- Kanza, S. P. (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. P. (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. P. (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. P. (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. P., & 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. P. (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. P. (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. P. (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
- Kanza, S. P. (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

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

# 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

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

# 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

# PSDI & Personal Details - Questions

www.psdi.ac.uk

@PSDI_UK

@PSDI_UK

linkedin.com/company/psdiuk

linktr.ee/samanthakanza

Mailing List: https://www.jiscmail.ac.uk/PSDI