Pacific Northwest
NATIONAL LABORATORY

PNNL-34942

Understanding Technical and Psychosocial Barriers to Realizing FAIR Data Process

September 2023

Nancy M Washton
Caitlyn M Ackerman

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, **makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

**Available to DOE and DOE contractors from
the Office of Scientific and Technical
Information,
P.O. Box 62, Oak Ridge, TN 37831-0062**
www.osti.gov
**ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov**

**Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov**

PNNL-34942

Understanding Technical and Psychosocial Barriers to Realizing FAIR Data Process

September 2023

Nancy M Washton
Caitlyn M Ackerman

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Executive Summary

The present study investigates barriers and facilitators to the implementation of Findable, Accessible, Interoperable, and Reusable (FAIR) data processes within the Physical Sciences Division of the Computational Sciences Directorate (PCSD). Employing a dual-method approach consisting of surveys and focus group discussions, the study aims to illuminate the complex interplay between technical and psychosocial factors that influence FAIR data adoption.

Key Findings:

- Surveys indicated that while staff generally understood the merits of FAIR data, the implementation was hampered chiefly due to concerns of accuracy, trust, and resource constraints.
- Focus group discussions further elucidated the nature and extent of these barriers, revealing issues ranging from career risk to administrative burdens.
- Despite general apprehensions, there was a common acknowledgment of the positive potential of FAIR data, such as streamlining research processes and fostering a community of shared insights and failures.

Recommendations:

- Convene a cross-disciplinary working group to facilitate implementation strategies and serve as FAIR data ambassadors.
- Implement NEMO, an open-source software for streamlined data handling and robust cost-benefit analyses.
- Engage in meta-data identification congruent with community practices.
- Foster dialogues with Principal Investigators and Project Managers regarding DataHub costs.
- Employ a dedicated “Data Librarian” to manage and curate data repositories.

The report underscores the necessity of a nuanced approach that considers both technical and psychosocial variables to accelerate FAIR data integration into the PCSD's research ecosystem. The detailed insights and recommendations aim to provide a roadmap for cultivating a data culture that is both rigorous and collaborative, thereby potentially expediting scientific discovery.

Acronyms and Abbreviations

ESC – Earth Sciences Capability

FAIR – Findable, Accessible, Interoperable, Reusable

IRB – Internal Review Board

PCSD – Physical Computational Science Directorate

PNNL – Pacific Northwest National Laboratory

Contents

Figures

Tables

Introduction

Background and Significance

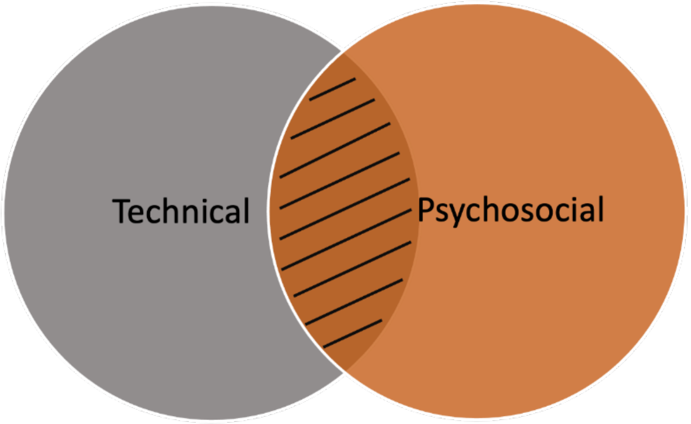

Figure 1. Barriers to generating FAIR data arise in technical and psychosocial arenas.

Over the past decade, there has been a burgeoning emphasis on the necessity for the physical sciences to produce robust, reusable data amenable to broad scientific inquiry. The FAIR data principles—findable, accessible, interoperable, and reusable—have been substantially embraced in data-intensive fields like astronomy and atmospheric chemistry due both to technical imperatives and established cultural norms of data sharing. However, this adoption is far from uniform across the physical sciences, and chemistry, in particular, lags materials science and physics in institutionalizing FAIR data standards. The impediments to FAIR data generation manifest along two principal axes: technical and psychosocial (Figure 1).

Technical roadblocks encompass issues like the absence of specialized ontologies, software inadequacies in capturing metadata, and unsupported data repositories. Psychosocial constraints range from entrenched domain-specific cultures and expectations for early-career scientists[1], seeking interactions with familiar collaborators to limit risk of data provenance[2,3], a deep sense of individual ownership of the intellectual property that the data represents[3], and the discomfort brought on by changing well established workflows. Intersecting these dimensions is the quandary of resource allocation: who funds the necessary infrastructure, and how will compliance be monitored and enforced? These questions, devoid of pre-existing frameworks, generate psychosocial unease and uncertainties about technical feasibility.[4]

While many initiatives targeting FAIR data adoption prioritize technical solutions[5], we posit that such unidimensional approaches are doomed to partial success due to the potent psychosocial factors influencing researcher behavior and receptivity to new workflows. This project aims to dissect both the technical and psychosocial intricacies of:

- Current workflows related to data generation, usage, storage, and sharing
- Potential workflows for FAIR data generation, usage, storage, and sharing

The catalysis sub-domain of chemistry is chosen as representative of chemistry due to the broad variety of workflows, scientist educational variability (*e.g.,* chemist, chemical engineer, physicist, materials scientist) and high need for more effectively engaging researchers to adopt FAIR data principles moving forward.

Research Design and Methodology

Workflow Structures

Catalysis research embraces multifaceted workflows which, for the purposes of this study, have been distilled into three archetypal categories: Instrumentalists, Lab-based, and Synthetic scientists (Figure 2). *Instrumentalist* focus on a specialized class of instruments, such as TEM, SEM, NMR, or Mass Spectrometers. Researchers in this bracket span multiple domains, including chemistry, materials science, physics, geochemistry, and biochemistry. *Lab Based* scientists conduct experiments to assess sample properties under varied conditions, often using multiple instruments for characterization. They exhibit a higher degree of workflow complexity compared to *Instrumentalists* and hail from a diverse range of scientific domains, such as chemistry, materials science, physics, geochemistry, biology, chemical engineering, or biochemistry. *Synthetic* scientists primarily engage in sample creation and subsequent characterization, again with more intricate workflows than *Instrumentalists.* These researchers also span multiple scientific fields, such as chemistry, materials science, geochemistry, biology, and biochemistry. These archetypes are further subdivided into workflow stages I-IV (Figure 3), delineating considerations for metadata inclusion and data storage. Information from a survey (Table 1) will have refined the questions for ensuing focus group discussions.

Survey and Focus Groups

A survey will be administered to the Physical Sciences Division research cohort to gain an understanding of scientist's perspectives of their current workflows. This data will be used in conjunction with the workflow stages (*vide supra*) to formulate and refine the structure for conducting focus group interviews which will be used to collect more detailed information on the technical and psychosocial attitudes and viewpoints related to FAIR data processes. A thematic analysis will be conducted to assess focal group responses as a function of research descriptor (*i.e., Instrumentalist, Lab Based, Synthetic Scientist)*, educational background and career stage. The information gained from these investigations (Figure 4) will provide insight into specific technical and psychosocial pain points, which will allow us to formulate and implement strategies to alleviate friction on the path to realizing FAIR data practices. Both the survey and focus groups content will be reviewed by the IRB for categorization and defined as either human subjects or human subjects' research.

Survey
Focus Group
Thematic Analysis

Figure 2. Workflow determining current practices, openness to alternate processes, and analysis of cohort and individual responses.

Case Study in Catalysis Data

To fully investigate technical and psychosocial barriers associated with a FAIR data workflow process, a case study utilizing nuclear magnetic resonance (NMR) data generated as part of a catalysis study will be undertaken. We will identify two coherent data sets that represent solution- and solid-state NMR experiments. These modalities vary in the types of research problems investigated and common workflow processes. Similarities and differences will be determined, as will the amount and robustness of meta data recorded at each step of the workflow as shown in Figure 3. Deficient areas will be identified and strategies to modify the workflow to include necessary meta data logging will be formulated. Although ontology development is outside the scope of this project, a simplified framework will be drafted and assessed. The feasibility of hosting an open database containing catalysis data generated within the Physical Sciences Division will be investigated, with outcome dependent upon the availability and ease of associating meta data with raw, processed and analyzed data. Assessments in this case study will be conducted within the context of technical and psychosocial barriers.

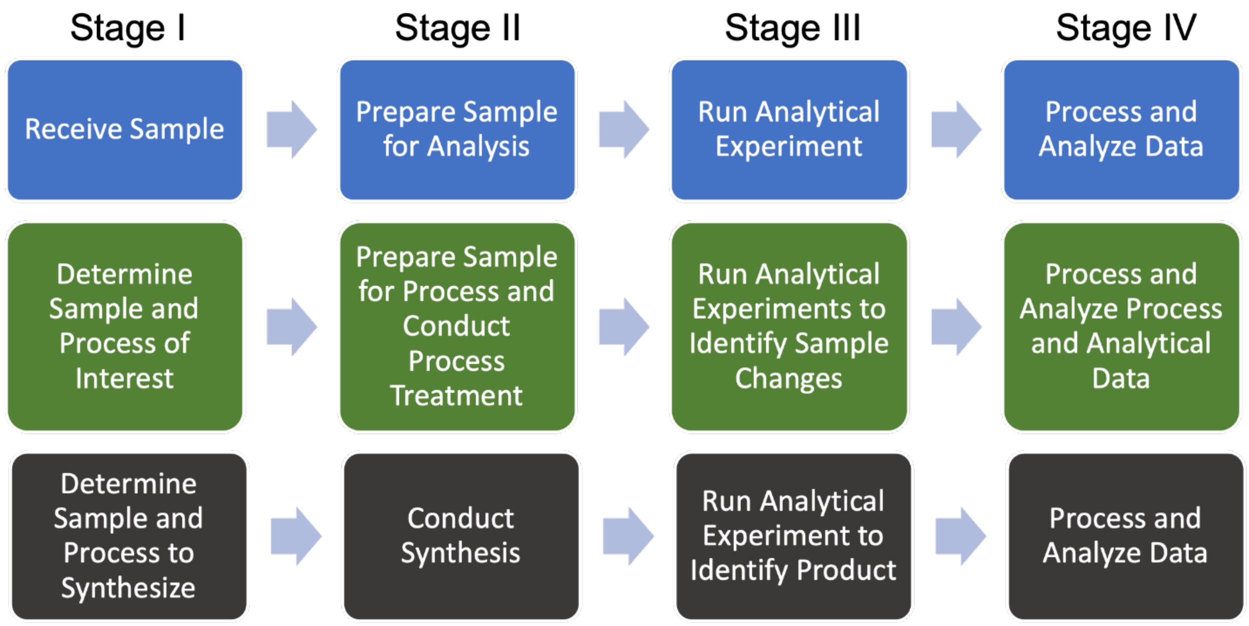
Stage I
Stage II
Stage III
Stage IV
Receive Sample
Prepare Sample for Analysis
Run Analytical Experiment
Process and Analyze Data
Determine Sample and Process of Interest
Prepare Sample for Process and Conduct Process Treatment
Run Analytical Experiments to Identify Sample Changes
Process and Analyze Process and Analytical Data
Determine Sample and Process to Synthesize
Conduct Synthesis
Run Analytical Experiment to Identify Product
Process and Analyze Data

Figure 3. Workflows for: Top, Instrumentalist; middle, Lab Based; bottom, Synthetic Scientists.

Table 1. Workflow stage information and data details as a function of researcher descriptor: Instrumentalist (top), Lab Based (middle), and Synthetic Scientists (bottom).

Stage I	Stage II	Stage III	Stage IV
Extend of information about the samples varies widely. Samples are typically provided by collaborators, and they are not always forthcoming with information, such as how the sample was prepared, whether it underwent post fabrication processes (*e.g.*, heat, pressure, etc.), or other analytical information they may have already obtained.	Researcher typically prepares the sample for the instrumental analysis and has this information and meta data associated with the preparation, but it is probably unusual for this information to be included in the instrument output files, although in many cases it is possible to add it as a text file in the software.	Instrument software logs much of the meta data associated with experiment parameters (some, like temperature, may not be auto logged).	The parameters used for data processing may or may not be captured in the software but at least some are typically noted in journal publications. Process and analysis software can include instrument associated, 3rd party software or both. Parameter capture is ad hoc at best.
Researcher typically has information on the sample and process chosen, and why each was chosen. This information may be input to a lab notebook.	Sample preparation may involve heat or acid treatment, typically *bench top* treatment. Researcher knows the information, but it is most likely not captured in any instrument software (best case is in lab book). Data may include temp, pressure, time, etc.	Instrument software logs most of the meta data associated with experiment parameters (some, like temperature, may not be auto logged). These analyses may use small bench top and large instruments (*e.g.,* EM, NMR, etc.).	The parameters used for data processing may or may not be captured in the software but at least some are typically noted in journal publications. Process and analysis software can include instrument associated, 3rd party software or both. Parameter capture is ad hoc at best.
Researcher knows exactly why the specific sample and process were chosen. This information may input to a lab notebook.	This process may be based on a preexisting or new method. The details may have been obtained from journal articles and may be input to a lab notebook.	Instrument software may log meta data associated with experiment parameters (some, such as temperature, may not be auto logged). The extent of captured meta data depends on the type of instrument used and whether the researcher manually includes meta data (*e.g.,* text file).	The parameters used for data processing may or may not be captured in the software but at least some are typically noted in journal publications. Process and analysis software can include instrument associated, 3rd party software or both. Parameter capture is ad hoc at best.

Methods

Primary methods for this study included a survey and focus group discussions with staff within PCSD to gain an understanding of the technical and psychosocial barriers to realizing findable, accessible, interoperable, and reusable (FAIR) data processes in the physical sciences with a focus on the intersection of technical and psychosocial arenas.

Prior to beginning the project, a project plan was submitted to the PNNL Institutional Review Board, and the project was deemed exempt from human subjects' research requirements.

Survey

To gather information from the staff within PCSD a survey was distributed via email and through Microsoft Teams chat. Detailed information about that survey and distribution of it is stated below.

Participants and Procedures

Participants were contacted via email or were given a survey link during a PCSD division meeting via Microsoft Teams chat. The initial contact to the participants described the topic and goals of the project and highlighted that participation was voluntary and anonymous.

Fifty-eight staff participated in the survey and eighteen staff left their contact information expressing their interest to participate in the focus group discussions. Table 2 shows the survey questions asked from the participants.

Table 2: Survey Questions

Questions
Which of the following best describes your role?
What are your most common daily tasks while conducting direct scientific research? (Check all that apply)
What are the most common methods/tools you use to capture the scientific context of your projects? (Check all that apply)
What are the most common methods/tools you use to store/save the raw data generated by instruments for your projects? (Check all that apply)
Do you manually include meta data with your raw data file?
Which type of software do you use as the primary tool to process the raw data with?
What are the most common methods/tools you use to store/save the processed data for your projects? (Check all that apply)
Do you manually include meta data with your processed data file?
What are the most common methods/tools you use to analyze data for your projects? (Check all that apply)
What are the most common methods/tools you use to store/save the data analyses for your projects? (Check all that apply)

Questions
Do you upload any data files to journal sites in conjunction with your article manuscripts?
Please note whether you upload raw, processed and/or analyzed data to the journal site.
What percentage of your journal submissions over the last two years include some form of uploaded data?
Rank your willingness to use new technologies and/or new software programs to capture and store data
What kind of learner are you when it comes to approaching a new technology or program that you are unfamiliar with?
Is there anything else you would like to share about your experience with utilizing technology or software programs for capturing raw data, meta data, processed data and/or analyses of processed data?
Would you be interested in participating in a focus group on this subject?

Focus Group

To gather more detailed information from a handful of staff within PCSD six focus group discussions were conducted. Detailed information about those discussions are below.

Participants

Participants were recruited via survey and then reached out to via email. The email again highlighted the topic and goals of the project and mentioned that participation was completely voluntary, and any comments provided during the focus group discussions would be kept anonymous.

Seventeen staff from a variety of physical science domains participated in the focus group discussions. All information connected to staff participants identity and responses is confidential.

Procedures

Six focus group sessions were conducted in a hybrid setting both onsite in the ESC building and via Microsoft Teams. Sessions were approximately 90 minutes in length and included between one and three participants each. At each session, there were a maximum of five project team members present. Two team members served as the moderator, asking technical questions, and leading the discussion, and three others attended the sessions to observe participant behavior and take notes for later analysis.

At the start of the discussion, the moderator reiterated the topic of conversation (FAIR data in the physical sciences) and emphasized that all comments made during the session would be kept confidential and not attributed to any participant. The moderator emphasized that participation was voluntary, and results would be provided for review at the end.

After this introduction, the moderator posed questions to the participants, the content of which are shown in Table 2. Because the conversation was interactive, some follow on questions were asked for clarification.

Table 3: Focus Group Questions

Questions
Do you agree or disagree with this statement: Open data results in efficiencies in research, more reproducible science, maximizing the use of a valuable resource, and the democratization of knowledge.
What do you perceive the impact of open data (FAIR) to be on your career, both from a generator of open data to a user of open data (with the understanding that you may not have generated nor used open data to-date)?
How do you feel about the scientific community having full access to data that you've already published? Please define data as you understand it here.
How do you feel about the scientific community having full access to data that you will you never publish (i.e., poor quality data sets, projects that didn't yield results, etc)?
If you were given a data set that contained the same information as data sets you generate, would you be able to reconstruct enough relevant information to contextualize the data in a way that would make it useful?
Can you envision a process that would allow you to connect all your data associated with a specific research thrust into a single package (e.g., a data directory with all files associated with the raw data)? What would a useful process look like to you? What type of process would you oppose?
Would you consider using an electronic lab notebook? Please expound on your answer
Have you used DataHub or other open data platforms? Why or why not?

During the session, the three team members assigned to take notes recorded responses to the questions using participant initials for later analysis and collation.

Thematic Analysis

After completion of the six focus group sessions, the three team members' notes were consolidated into a single set for thematic analysis. This helped ensure that the most information possible was captured from the notes.

Results

Survey

The survey elicited 58 responses from staff members, yet its utility proved to be suboptimal for delivering actionable insights for the project's primary objectives. Instead, the survey served as a constructive blueprint for designing the focus group discussions. A graphical representation of survey questions is available in Appendix A.

Focus Group

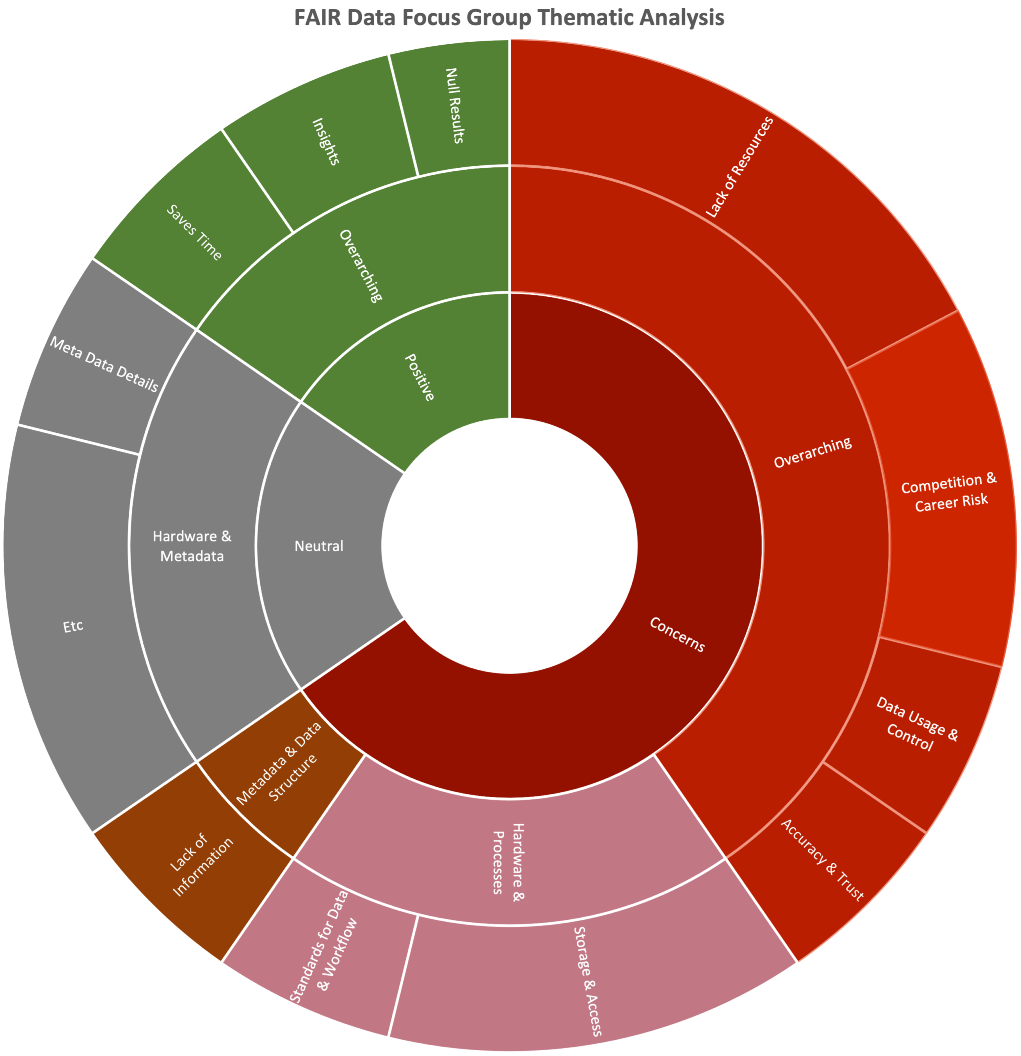

Figure 4. Themes present in focus group responses.

The thematic analysis of the survey outcomes isolated several key dimensions of perspectives towards FAIR data, categorized into positive, neutral, and concern-centric domains. Concern-centric themes dominated the narrative, as illustrated in Figure 4.

Positives

- Allows for new work to be done or different interpretations on past work
- Creates a starting point for your work
- Thoughts and insights from others
- People can disagree and provide their insights and comments
- Saves time and resources
- Create a community of failure to learn from one another
- Usefulness of null results or detailed failed mistakes

Concerns

Concerns fell into several broad categories:

- Accuracy and trust
    - Is the information correct
    - Paucity of accurate results
    - General lack of trust in the data and associated information

- Desire to control data usage and risk
    - Desire to control usage as some users do not have the context or experience to use the data appropriately
    - Risk of data being used incorrectly or without context
    - People come to the wrong conclusion (by accident or on purpose)

- Lack of resources
    - Staff labor - time it takes it takes to upload information
    - Heightens data resource needs (time, money, software need, etc.)
    - Wasted Resources (time, money)
    - Not productive to science
    - Lack of resources to maintain (money, time)
    - Gain vs effort
    - Extra step to an already tight and timely process
    - No additional time or resources needed, or if so, minimal
    - Need to provide resources or minimal ask for staff to contribute or to get buy-in

- Competition & career risk
    - Staff feel their career is at risk
    - Decreased competitiveness
    - Competitive field (someone taking your information and using in a way that marginalizes the data generator)
    - To generate and share data that is clear and reproducible competition must be eliminated
    - Others taking work and enhancing it or taking it farther
    - Inability to get follow-on funding

- Storage and access
    - Storage is a barrier
    - Long term storage issues
    - Tape Drives or thumb drives are problematic
    - Current difficulty with storage systems
    - Inferior network connections

  - Inability to easily access data

- Standards for data and workflows
  - Creating a standardized workflow will be challenging
  - Organizing the data will be difficult
  - Data standards need to be in place prior to anything else

- Meta data and data structure
  - Lack of information provided for the data to be useful
  - Lack of narrative or relevant information (*e.g.,* meta data)

Neutrals

- Meta data details
  - All details that are relevant must be included
  - Contextual information must be included
  - The details are important to make sure you can reproduce the data the same or learn from their mistakes as to why it was done that way

- General
  - Sample preparation information must be included
  - Creating standardization for everything practicable
  - File names structure must be agreed upon
  - Data structure must be agreed upon
  - User friendly interfaces and processes
  - Need stability and reliability within the tool itself

Discussion

Thematic and Empirical Landscape

The data delineates a multifaceted tableau of staff perceptions concerning FAIR data principles. A preponderance of concerns overshadows the positive elements, demonstrating an existential dissonance within the research cohort engaged in this effort vis-à-vis FAIR data adoption. Issues span the gamut from data reliability and control, resource limitations, to career and competitiveness risks. While neutrality prevails on procedural aspects like metadata and standardized workflows, the overarching sentiment reflects a cautious hesitancy.

Overarching Concerns and Nuanced Strategies

The recurring thread of administrative burden without corresponding resource allocation, and the perceived jeopardy to career advancement, resonate uniformly across career stages and specialized domains. These concerns underscore the imperative for a nuanced, stakeholder-oriented approach to implement FAIR data principles, an approach that navigates the intricate web of operational exigencies and academic valuations.

Recommendations

1. Working Group Formation: Given the heterogeneous landscape of perceptions and needs, the constitution of a diverse working group, as recommended, could act as a linchpin for synthesizing the multiple axes of concerns, needs, and possible solutions. Comprising staff from varying disciplines, this ensemble will facilitate a cross-pollination of ideas and strategies aimed at a more universalized adoption of FAIR principles.

2. Technological Adaptions: The recommendation to instantiate NEMO—an open-source software from NIST—underscores the need for a robust technological infrastructure to ameliorate the myriad challenges encountered. Its flexibility for adaptation to different data-type cohorts and capacity for granular instrument usage analytics makes it an apposite choice for integrating into the proposed framework.

3. Metadata and Community Practices: The importance of metadata—a recurrent theme—should be given impetus, guided by established community practices. This ensures that data is not just FAIR, but also contextually enriched, making it a robust scientific artefact.

4. Documentation Workflow: The working group will also assess options like Electronic Lab Notebooks (ELN) for documenting workflow, potentially catalyzing a Laboratory Directed Research and Development (LDRD) initiative to identify best practices.

5. Cost and Infrastructure: Open dialogues with Principal Investigators and Project Managers on cost implications—especially regarding DataHub—are essential to weigh lab-wide support against project-specific funding. Additionally, the recommendation to appoint a Data Librarian reflects an evolved understanding of the role of data in contemporary scientific research, emphasizing the necessity of specialized human resources.

Concluding Remarks

Efforts to transition to FAIR data principles should not merely be a perfunctory nod to modern data management but rather a thoughtful, multi-pronged initiative that takes into account the nuanced landscape outlined herein. Building on these recommendations, organizations can construct a more pragmatic and adaptive roadmap, where FAIR data not only exists but thrives.

References

1. Toribio-Florez D, Anneser L, deOliveira-Lopes FN, et al. Where Do Early Career Researchers Stand on Open Science Practices? A Survey Within the Max Planck Society. *Front Res Metr Anal.* 2020;5:586992.
2. Pasquetto IV, Borgman CL, Wofford MF. Uses and Reuses of Scientific Data: The Data Creators' Advantage. *Harvard Data Science Review.* 2019;1(2).
3. Frey JG, Bird CL. Scientific and technical data sharing: a trading perspective. *J Comput Aided Mol Des.* 2014;28(10):989-996.
4. <Nature_OpEd_2017.pdf>.
5. Draxl C, Scheffler M. The NOMAD laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials.* 2019;2(3).

Appendix – A

Table 4. Survey Questions in Relation to the Appropriate Figure

Questions	Related Figure
Which of the following best describes your role?	Figure 5
What kind of learner are you when it comes to approaching a new technology or program that you are unfamiliar with?	Figure 6
What are your most common daily tasks while conducting direct scientific research? (Check all that apply)	Figure 7
Rank your willingness to use new technologies and/or new software programs to capture and store data	Figure 8
What are the most common methods/tools you use to store/save the raw data generated by instruments for your projects? (Check all that apply)	Figure 9
What are the most common methods/tools you use to store/save the processed data for your projects? (Check all that apply)	Figure 10
What are the most common methods/tools you use to store/save the data analyses for your projects? (Check all that apply)	Figure 11
What are the most common methods/tools you use to capture the scientific context of your projects? (Check all that apply)	Figure 12

Classification

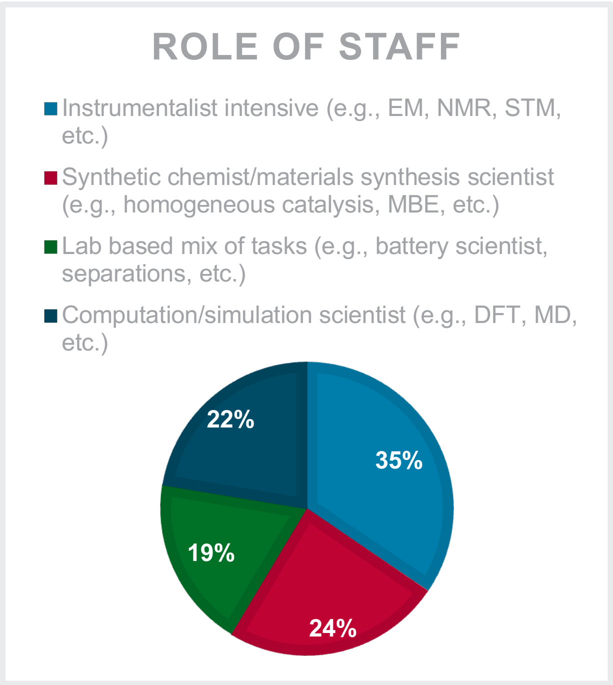

Figure 5. Role of PCSD Staff

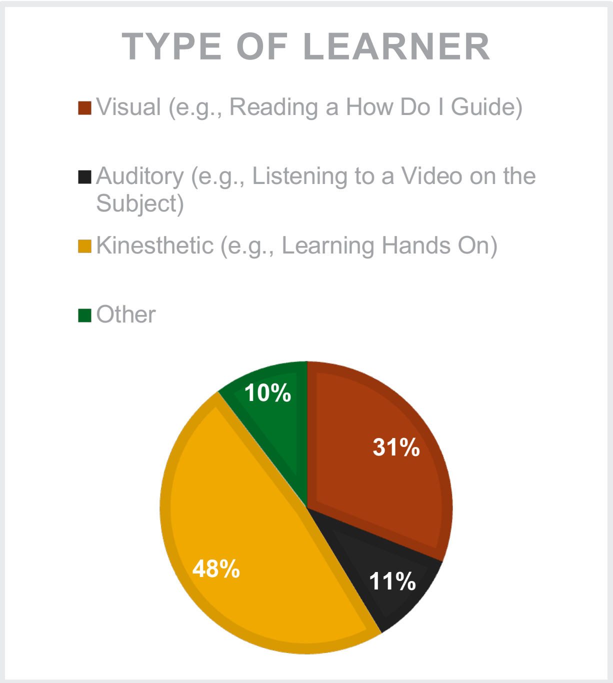

Figure 6. Type of learner for PCSD staff

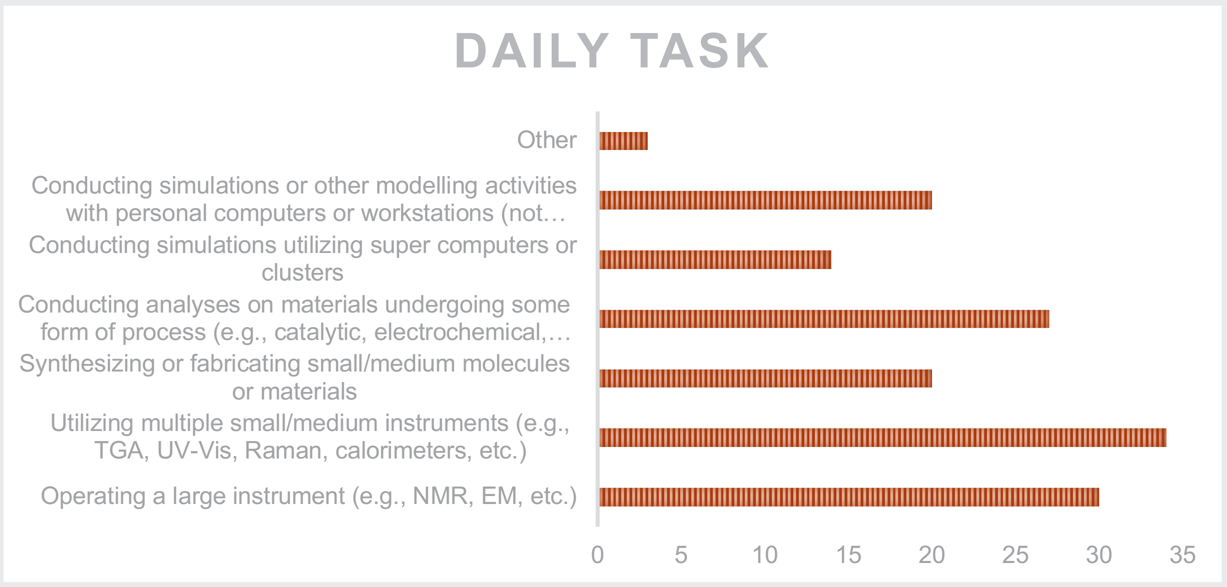

Figure 7. Daily tasks PCSD staff accomplish in their roles.

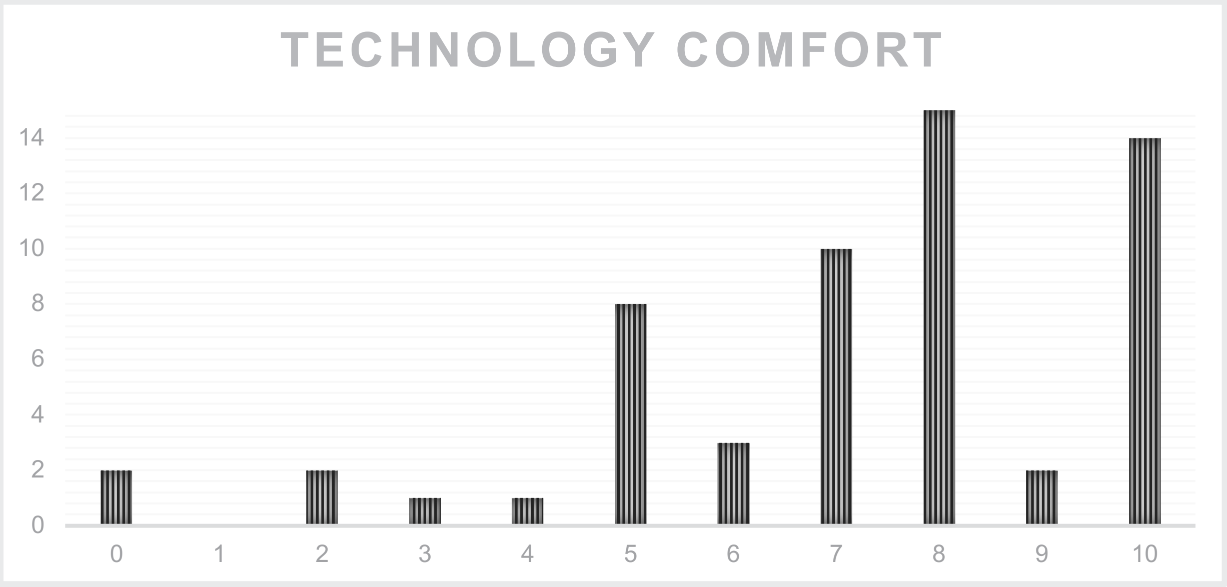

Figure 8. PCSD Technological Adoption Comfort Level

Data Storage

During this survey questions trying to gain an understanding of how staff collect and store their different forms of data were asked.

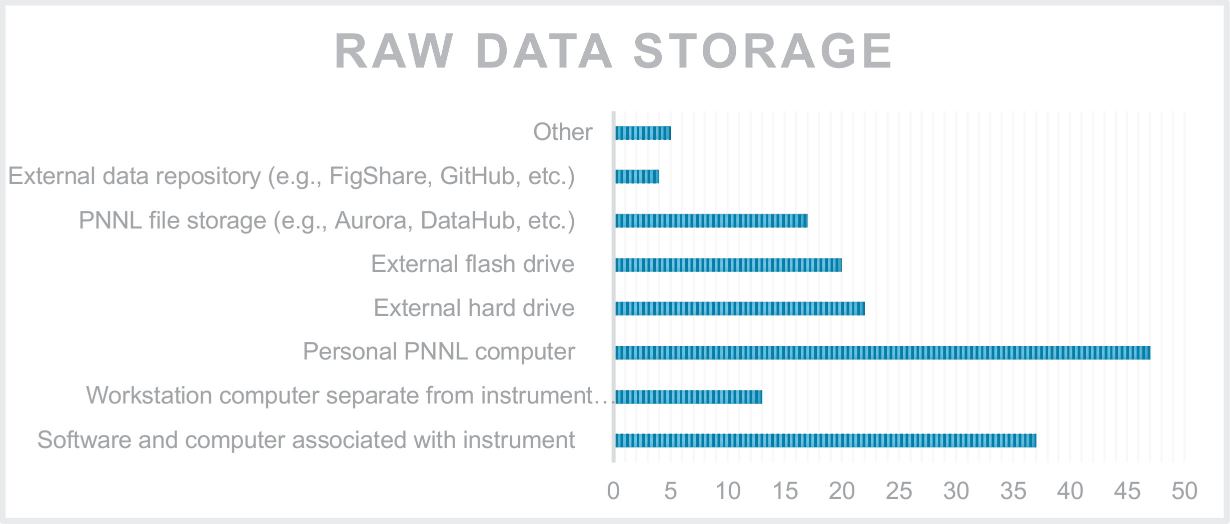

Figure 9. PCSD Staff Raw Data Storage

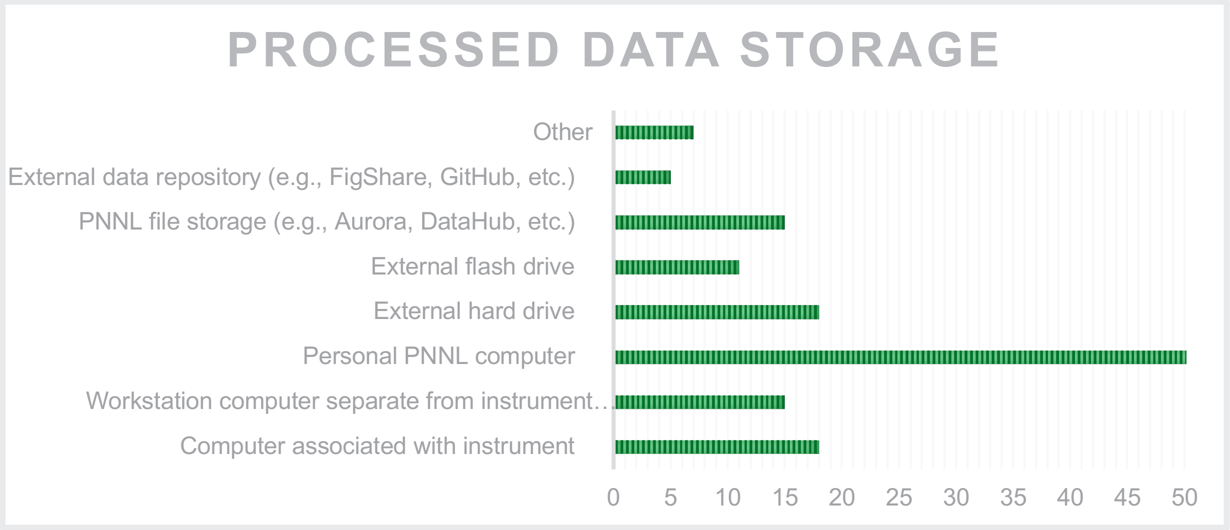

Figure 10. PCSD Staff Processed Data Storage

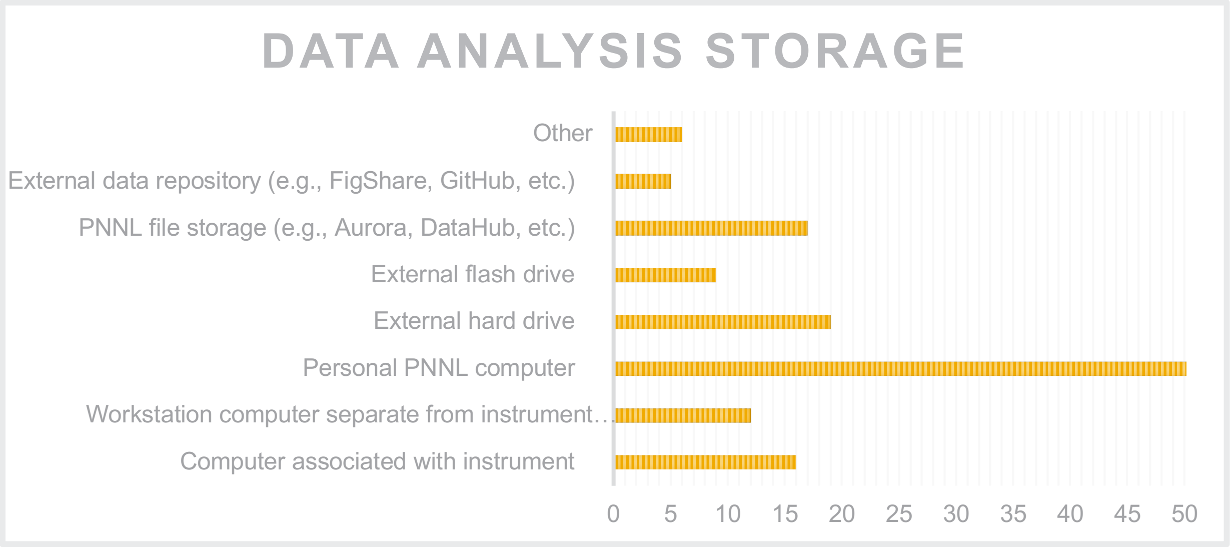

Figure 11. PCSD Staff Data Analysis Storage

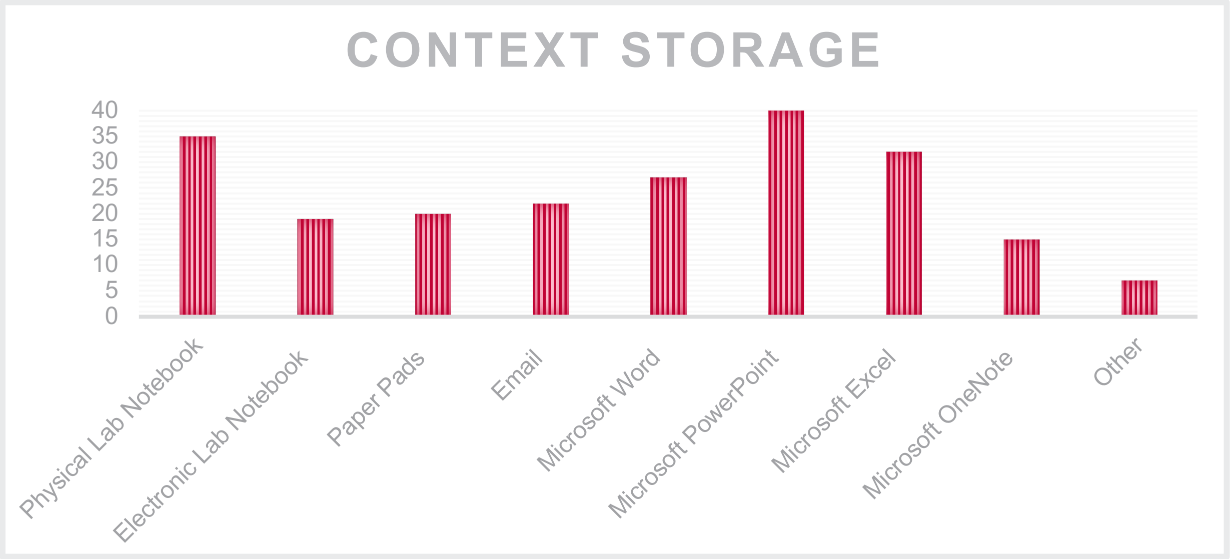

Figure 12. PCSD Staff Contextual Data Storage

Appendix – B

MURAL information from focus groups:

1. Do you agree with this statement: Open data results in efficiencies….
   - Caveats
     - i. Some data cannot be exported to an open source format
     - ii. How do you trust the data and the information that was provided
       - 1. Information is correct
       - 2. All details that are relevant are there
     - iii. Desire to control data usage
     - iv. Assumptions that this will be hard in practice
   - Positives
     - v. Allows for new work to be done or different interpretation on past work
     - vi. Creates a starting point for your work
     - vii. Saves time

2. Career Influence
   - User or Generator
     - i. Positive
       - 1. Allows for access to see others data
       - 2. Saves time and resources
       - 3. Thoughts and Insights from others
         - a. Bugs can be found and fixed
         - b. People can disagree and provide their insights and comments
     - ii. Negative
       - 1. Waste of time/Not Useful
         - a. No one uses it
         - b. Lack of accurate results
         - c. Lack of information provided to be actually useful data
       - 2. Desire to control usage as some users do not have the context or experience to use the data appropriately
       - 3. Inability to easily access data
         - a. Not have to contact original author for support data
       - 4. Lack of Resources
         - a. Staff Labor - Time it takes it takes to upload information
         - b. Lack of tools or storage options
       - 5. Competition
         - a. Staff feel their career is at risk

3. Fully Published Access
   - Caveats
     - i. Risk of data being used incorrectly or without context
       - 1. People come to the wrong conclusion (by accident or on purpose)
       - 2. The details are important to make sure you can reproduce the data the same or learn from their mistakes as to why it was done that way
     - ii. Community is ever changing

    iii. No uses the data
    iv. Heightens data resources (time, money, software need, etc.)
        1. Being able to have someone else reproduce it (readability)
        2. Organize the data
        3. Easily upload the data will encourage use
    v. Competitive field (someone taking your info)
- Positives
    i. Already open sourced if published
        1. Some details are missing though**
    ii. Create a community of failure to learn from one another

4. Access to unpublished data
- Neutral
    i. Career level influences mindset on FAIR data
    ii. Ability to leave unsolved research to the community to finish
    iii. Ease of reproducibility to shift culture to allow scientists to see the importance of details
- Caveats
    i. Wasted Resources (Time, Money)
        1. Non-useful data being out there
            a. Lack of narrative or relevant information
        2. Failed information to shift through
        3. Inability to get follow-on funding
    ii. Competitiveness
        1. Others taking work and enhancing it or taking it farther
    iii. Not productive to science
    iv. Motivation to publish in good journals and have good information to share
- Positives
    i. Usefulness of null results or detailed failed mistakes

5. Given same data as your work, could you make it useful
- Neutral
    i. You can get close to reproducibility but there might always be little differences
    ii. It is nice to know the work is there and done but will normally compare and check the work.
- Caveats
    i. Need clear directions and methods
        1. Dependent on the instrument and software used
        2. All meta data included with context
        3. Sample prep info
        4. Data standards need to be in place
    ii. To have things be clearly reproducible you need to eliminate competition
    iii. Operator skill might be an influencer to reproducibility

6. Process
- Positives
    i. Organization helps

- Caveats
  i. Storage is a barrier
     1. Long term storage issues
        a. Lack of resources to maintain (money, time)
  ii. Gain vs Effort
  iii. Current difficulty with systems
     1. Network connection
     2. Tape Drives or Thumb Drives
- Framework Aspects
  i. Creating a standardized workflow will be challenging
     1. Needs flexibility and adaptability
- Technical Aspects
  i. Dealing with different data types
  ii. Ability to link between raw data, processed data, etc.
  iii. Need stability and reliability within the tool itself
  iv. Need to provide resources or minimal ask for staff to contribute or to get buy-in
  v. Ability to track a sample as it moves from lab to lab or instrument to instrument
- General
  i. Each area in the lab is different. From directorate to even individual person
- Suggestions
  i. Creating standardization
     1. File names
     2. Data Structure
  ii. User Friendly
     1. Ease of use for user interface
     2. No additional time or resources needed, or if so, minimal

7. E-Lab Notebooks
- Positives
  i. Jupyter notebooks for plotting and ideation
  ii. Most is already done electronically
  iii. Ability to insert videos or tutorials
  iv. Searchability on work
- Caveats
  i. One size does not fit all
- Technical Aspects
  i. Notes need refinement and review process that you cannot digitalize
  ii. Lack of ability to physically have laptop in space
  iii. Mac issues with OneNote
  iv. Need ability to type or write formulas and have specialized characters
     1. Stylist or write to text function
     2. Speech to text function

- Framework Aspects
  i. Accessibility for users and collaboration
  ii. Standardization on notes/templates
- General

        i. E-Lab Notebooks have the same risk as open data in general
        ii. Lack of practicality

8. Open-Source Platforms
    - Technical Aspects
        i. Data storage with TBs of data
        ii. Auto upload of data
        iii. Ability to link to IR
        iv. Detailed and specific management plan and system
            1. Dedicated librarians to help
            2. Ability to partner externally
        v. Proper broadcasting of product after release
    - Caveats
        i. Extra step to an already tight and timely process

- Statement on the way things are currently
    - You find a paper that’s relevant, and you can go into the SI and find their “raw” data.
    - I saw an article that said data files are available upon reasonable request. It’s never been requested and I’ve never requested raw data.
    - no one supplies data files in pubs so you must contact authors.
    - the data is entirely absent in papers you need to email the authors.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov