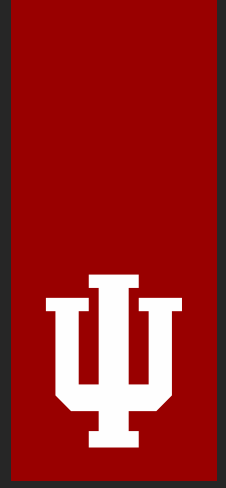


SCHOOL OF PUBLIC HEALTH BLOOMINGTON

Data Sharing: Pros, Cons, and How-to

Stephanie Dickinson, sd3@indiana.edu
Biostatistics Consulting Center, IU SPH,
NSC Comparative Data Analytics Core at UAB

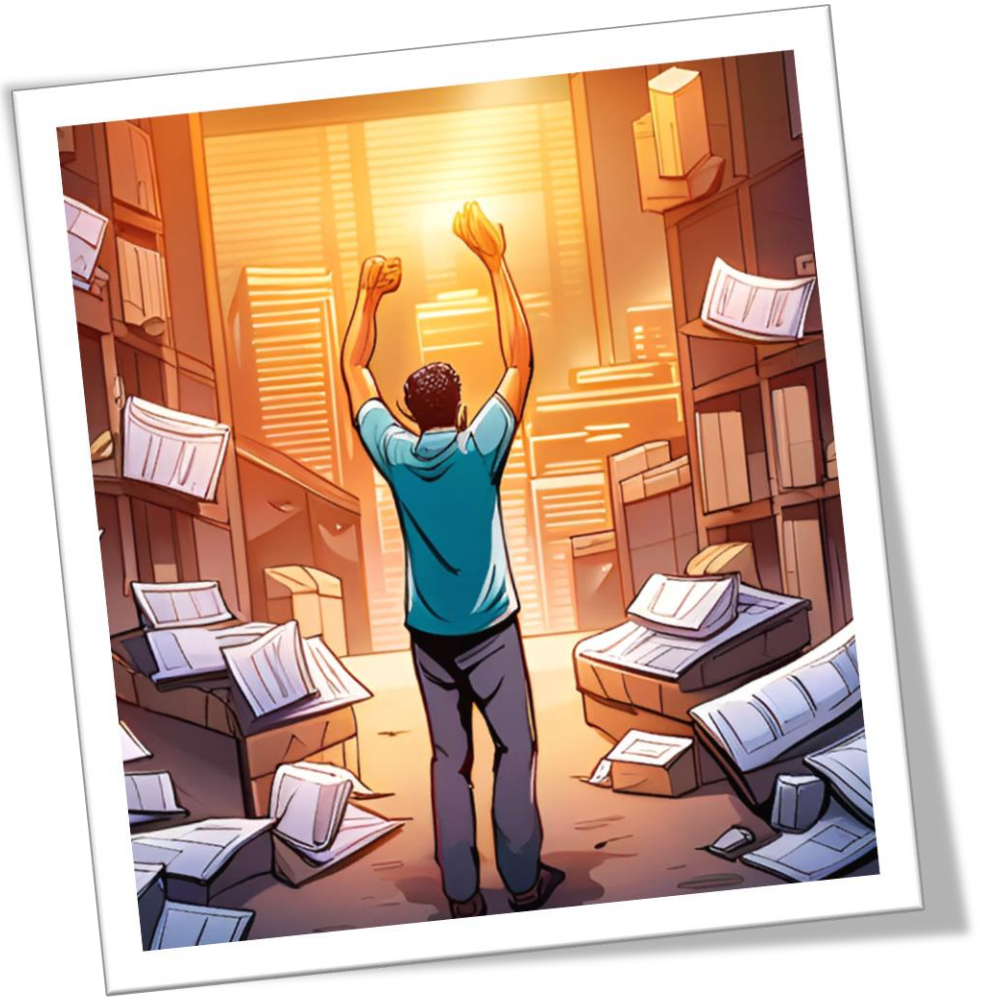
INDIANA UNIVERSITY BLOOMINGTON

Outline

- Challenges
- Advantages
- What & when to share
- How-to

Challenges of Data Sharing

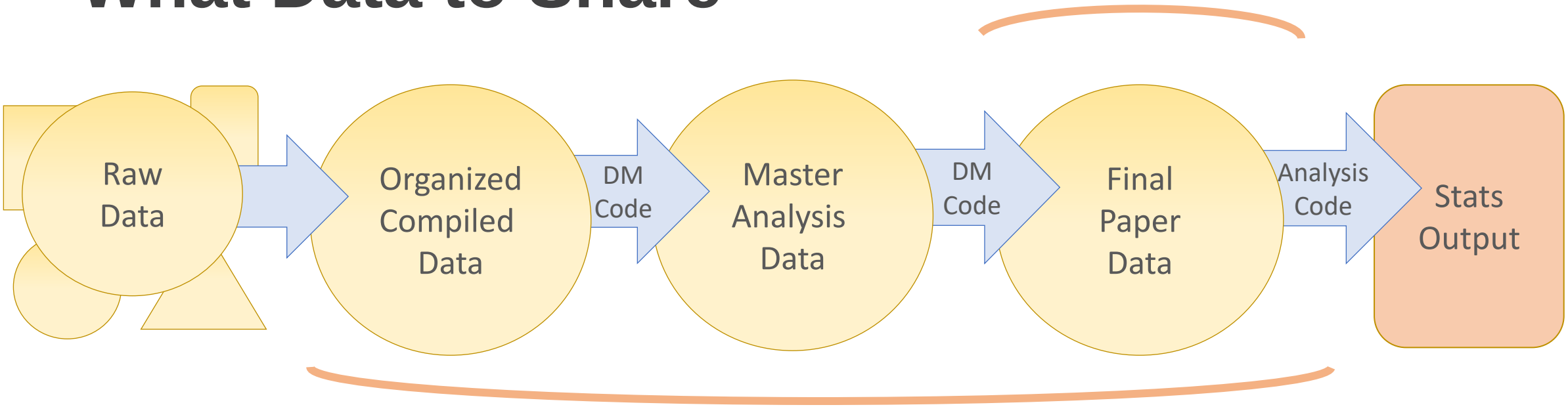
- Researchers are not sure what to share or how
- Time and Money
  - Organizing data, codebooks, repositories takes time
  - One more thing to do for grant proposal (DMS) or journal submission
- Needs strong organization and communication in team
- Potential for errors and inconsistencies in data or analysis
- Ethical and legal questions regarding data ownership and privacy:
  - Questions of data ownership, access rights, and usage control
  - Human subject data privacy



Advantages of Data Sharing

- It's **required** for NIH funding and many journals
- Allows analytic ***Reproducibility***
  - Reproduce all results, running code provided on data provided
- Allows sensitivity or other, ***New analyses***
- Advances science for ***New research*** questions.
- Organized data and clean code **reduces errors** in publications
- **Saves time later** to share with collaborators or scientific community

What Data to Share



- Lab books
- Paper surveys
- REDCap, Qualtrics
- Bio samples

- Excel, CSV, SAS
- Tabulated into rows and columns
- Data dictionary
- SDTM

- Derived variables
- Recodes, Transforms
- Subject level, Longitudinal, etc
- ADaM

- Data necessary and sufficient to reproduce analyses in paper
- Clean and tidy

Note: Syntax/Code (Script) is used for each Data Management (DM) step and Analysis

When to Share

1. “At a minimum, scientific data supporting a publication must be shared **by the time of publication** …”

2. “Other scientific data must be shared **by the end of the research project** or protocol.”

https://oir.nih.gov/sourcebook/intramural-program-oversight/intramural-data-sharing/2023-nih-data-management-sharing-policy

How to Share Data

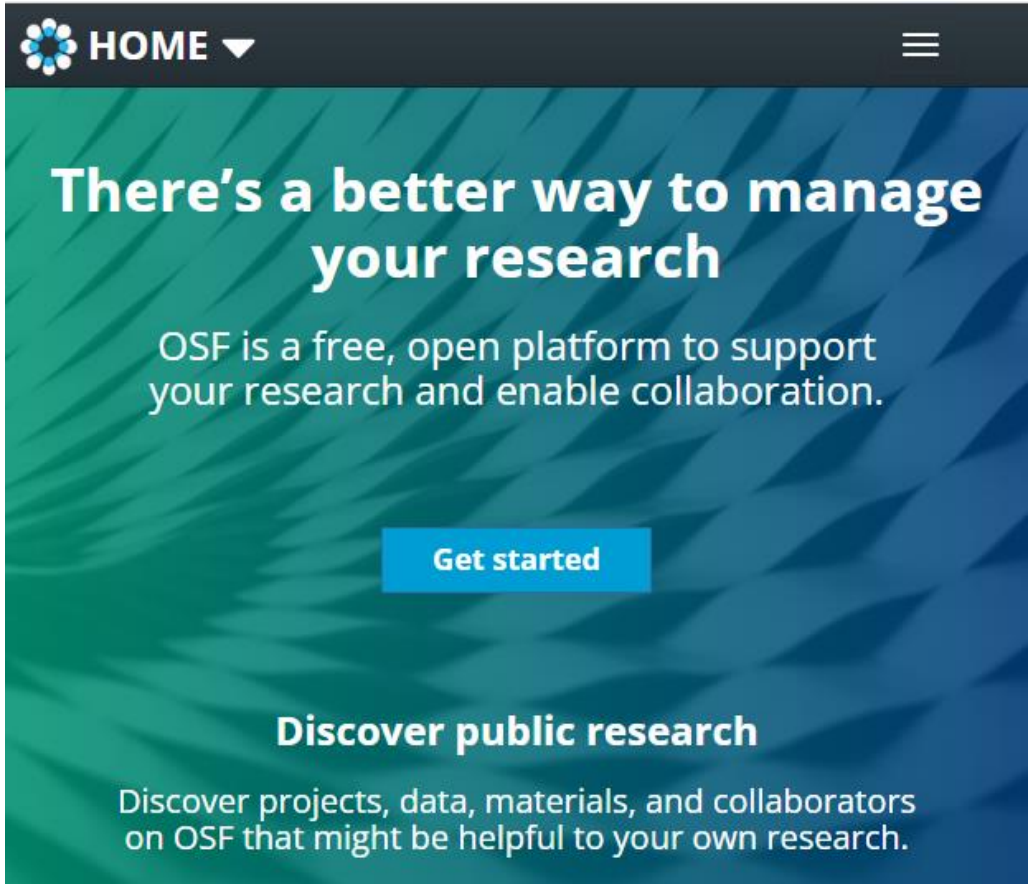
Include **Codebook/Data** Dictionary and File **Directory**

Codebook: variables, labels, definition, **and coding** (e.g. 1=Trt, 0=Ctrl).

File Directory: Clearly organize and label files.

FAIR (Findable, Accessible, Interoperable, Reusable)

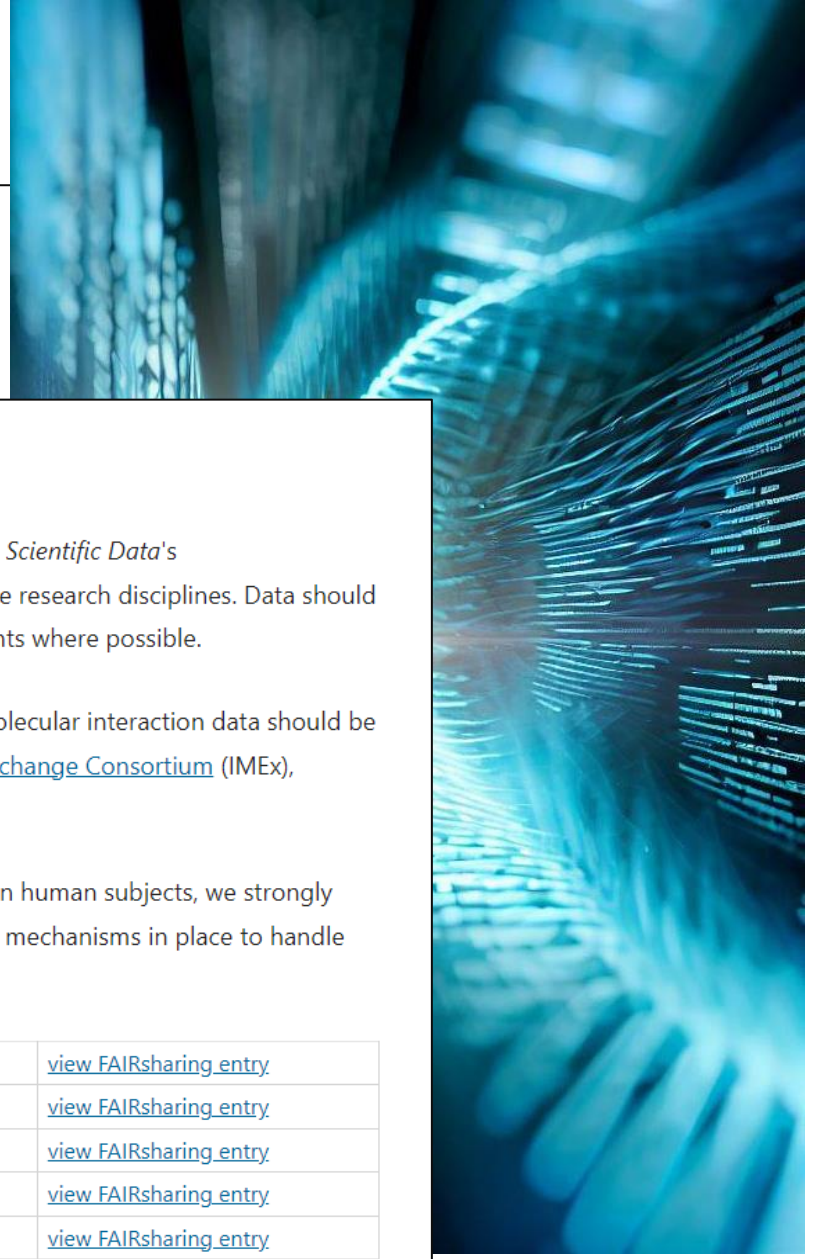
Repositories



- We like **OSF.io** (Open Science Framework) for easy public file sharing
- Support effective data discovery and reuse.
- Consider repositories specific to the discipline or data type.
- See "Selecting a Repository" and "Repositories for Sharing Scientific Data" from NIH
- See "Data Repository Guidance" from Nature
- Whatever you choose, put the **LINK** in your paper

View data repositories

- **Biological sciences:** Nucleic acid sequence; Protein sequence; Molecular & supramolecular structure; Neuroscience; Omics; Taxonomy & species diversity; Mathematical & modelling resources; Cyt
  focused resources
- **Health sciences**
- **Chemistry and Chemical biology**
- **Earth, Environmental and Space scienc**
  sciences; Astronomy & planetary sciences
  sciences; Ecology; Geomagnetism & Palae
  sciences
- **Physics**
- **Materials science**
- **Social sciences**
- **Generalist repositories**



Omics

Functional genomics

Functional genomics is a broad experimental category, and *Scientific Data*'s recommendations in this discipline likewise bridge disparate research disciplines. Data should be deposited following the relevant community requirements where possible.

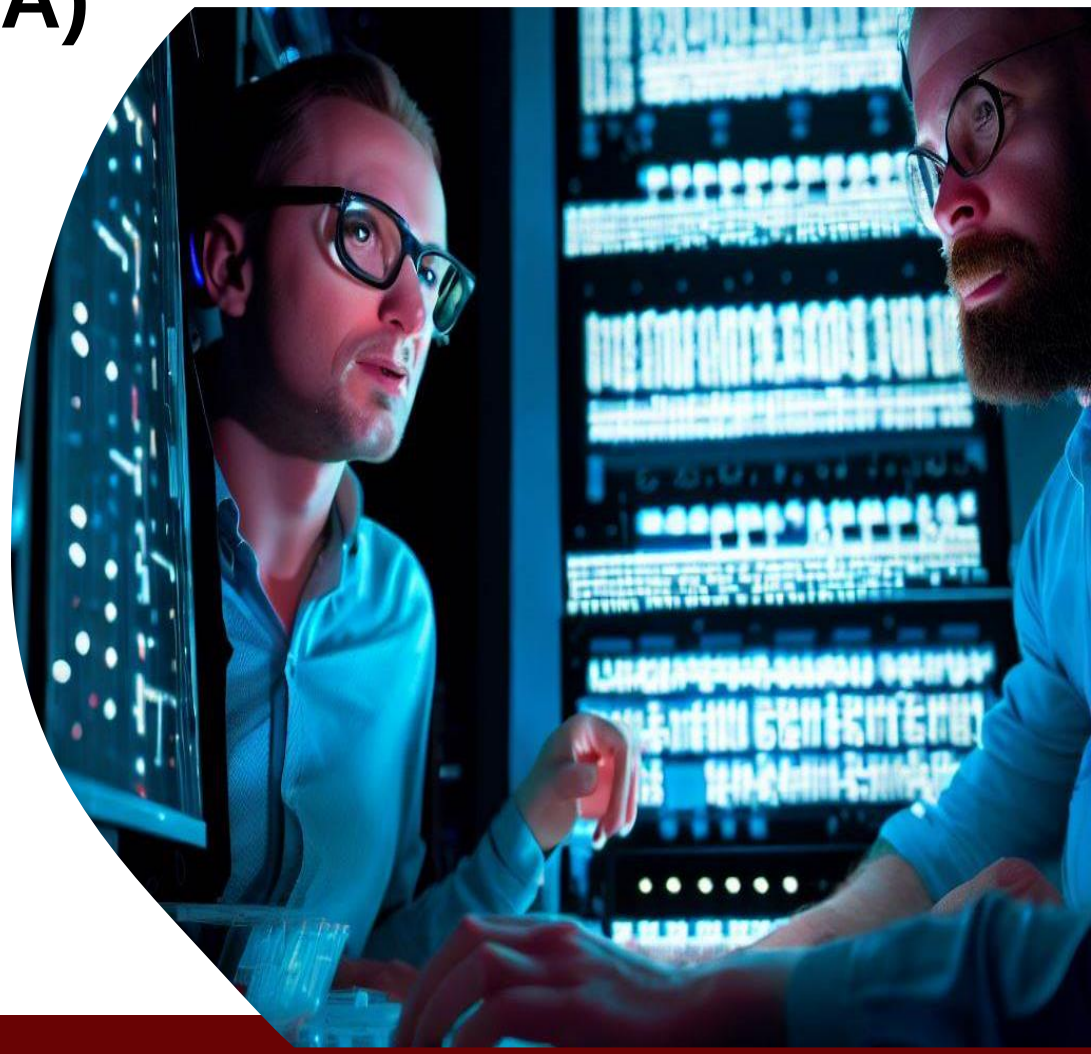
Please refer to the MIAME standard for microarray data. Molecular interaction data should be deposited with a member of the International Molecular Exchange Consortium (IMEx), following the MIMIx recommendations.

For data linking genotyping and phenotyping information in human subjects, we strongly recommend submission to dbGAP, EGA or JGA, which have mechanisms in place to handle sensitive data.

ArrayExpress	view FAIRsharing entry
Gene Expression Omnibus (GEO)	view FAIRsharing entry
GenomeRNAi	view FAIRsharing entry
dbGAP	view FAIRsharing entry
The European Genome-phenome Archive (EGA)	view FAIRsharing entry
Database of Interacting Proteins (DIP)	view FAIRsharing entry

One example:
The Sequence Read Archive (SRA)

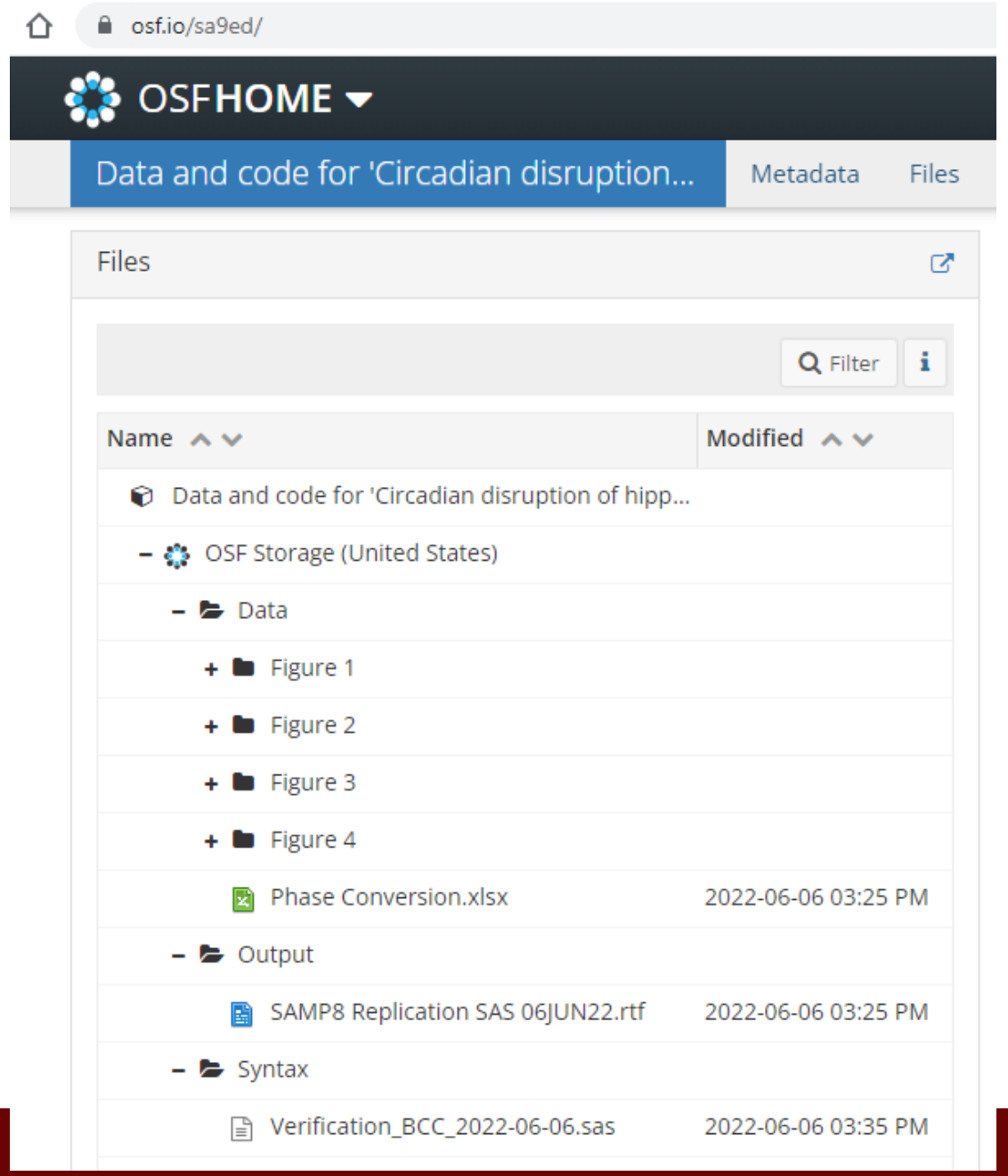
- NIH/NCBI archive of all types of *sequencing data*.
- Largest publicly available repository of high-throughput sequencing data.
- Accepts data from all branches of life, metagenomics, and environmental surveys
- Including those involving human subjects (de-identified).
- Stores raw sequencing data and alignment information.



Examples

- For papers we verify & support through NSC, we share data and code on OSF.io, e.g. https://osf.io/sa9ed/

- For two papers we reviewed, data were public, and conclusions were upheld, but there were challenges.
  - Iram, T., Kern, F., Kaur, A. et al. Young CSF restores oligodendrogenesis and memory in aged mice via Fgf17. Nature 605, 509–515 (2022). https://doi.org/10.1038/s41586-022-04722-0
  - Li, V.L., He, Y., Contrepois, K. et al. An exercise-inducible metabolite that suppresses feeding and obesity. Nature 606, 785–790 (2022). https://doi.org/10.1038/s41586-022-04828-5



The End


AFRAID TO SHARE
RESEARCH DATA
imgflip.com


SHOW ME THE DATA!!!!
imgflip.com


YOUR RESEARCH DATA
SHARE YOU MUST
imgflip.com

RRT (Rigor, Reproducibility, Transparency)

Analysis Verification Checklist

1. Verify that you can ***reproduce*** every numeric result with final data and code shared.

2. Verify that the analyses are described ***transparently*** in text and tables.

3. Verify that the analyses were done ***rigorously***.