Research Data Management

Best Practices

GPS Skills Course
Winter 2018

Mary Linn Bergstrom
Reid Otsuji

The Library
UC SAN DIEGO

Workshop Objectives

Students will be aware of the research data management life cycle and best practices in managing research data.

Students will be aware of the Teaching Integrity in Empirical Research [TIER] protocol

Students will be familiar with the Open Science Framework resource and how it's components align with best practices in research data management.

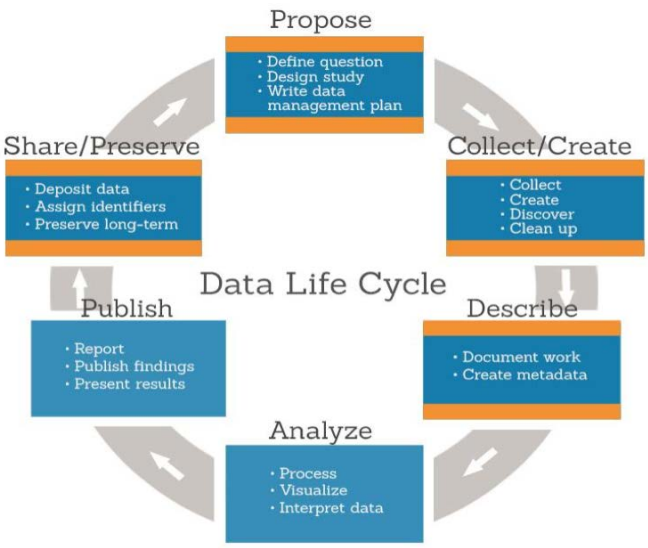
The Library UC SAN DIEGO

Introduction

Part 1:
A. Why is data management important?
B. Best practices to Consider

Part 2:
Open Science Framework

Research Data Life Cycle



Managing data in the Data Life Cycle:

- Data management planning
- Documenting project/file details
- Choosing file formats
- File organization & naming conventions
- Access control & security
- Backup & Storage
- Sharing and Preservation

The Library
UC SAN DIEGO

Data Management Strategy

- Establish best practices for your data management
- Plan to share well-documented data
- Create a concise data management plan for your grant proposal
- Standardize data management practices and policies in your research lab.

Ensure that your data will be available to colleagues, peers & future generations to enable reproducible research

The Library
UC SAN DIEGO

Benefits

- Promotes successful data collection techniques
- Improves ease of using and sharing data
- Saves time, effort and resources during the research project
- Increases research impact and visibility.
- Reduces cost of creating, protecting and storing data
- Ensures that data will be available to colleagues, peers & future generations to enable reproducible research

The Library
UC SAN DIEGO

The Library
UC SAN DIEGO

Data Sharing and Management Snafu in 3 Short Acts

https://www.youtube.com/watch?v=66oNv_DJuPc

RIGOR AND REPRODUCIBILITY

Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings. The application of rigor ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results. When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research. This is especially important for clinical trials in humans, which are built on studies that have demonstrated a particular effect or outcome.

Johns Hopkins University students in a laboratory. *Johns Hopkins University*

In recent years, however, there has been a growing awareness of the need for rigorously designed published preclinical studies, to ensure that such studies can be reproduced. This webpage provides information about the efforts underway by NIH to enhance rigor and reproducibility in scientific research.

Email Updates

Sign up to receive email updates about rigor and reproducibility.

Sign up for updates

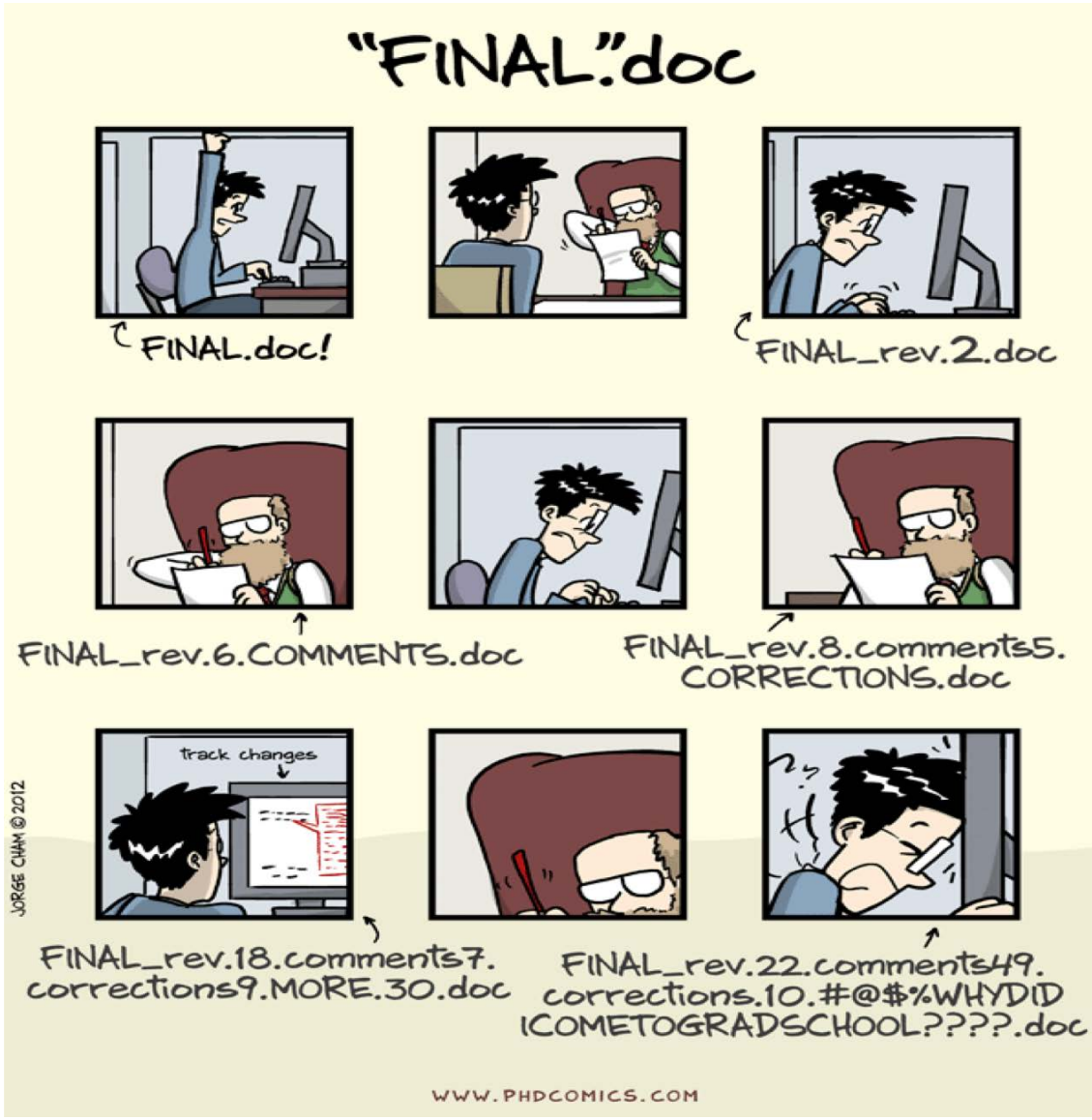
Contact Us

Please send email to NIHReprodEfforts@od.nih.gov .

Best Practices

- Organization
- TIER Protocol v3
- Documentation and Description
- Metadata
- Data Clean-up
- Basic Storage
- Backup
- Preservation

The Library
UC SAN DIEGO


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

Organization

File and Folder organization
Choose a consistent filing system that will make sense to you or someone else five years from now.

Choose a logical directory hierarchy. For example: TIER Documentation Protocol.
http://www.projecttier.org/tier-protocol/specifications/

Assign descriptive file names. E.g. DOLInterview_DoeJane_20061207

//Project001/SiteB/SiteB_2010_rawdata.txt

Is better than . . .

//Project001/SiteB/2010/rawdata.txt

The Library
UC SAN DIEGO

TIER Protocol

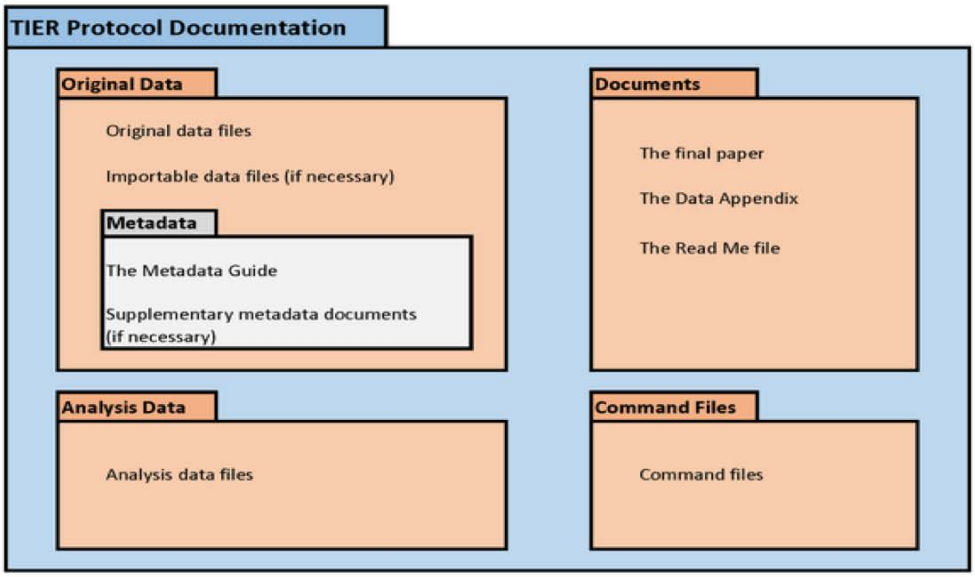
Developed at Haverford College –
Teaching Integrity in Empirical Research [TIER] protocol

- a recommended protocol for comprehensively documenting all the steps of data management and analysis that go into an empirical research paper.
- All documentation, do-files, scripts, raw data, metadata, and a copy of the final paper are organized in a specific file structure.
- This file structure keeps your data organized and supports easy replication of results.

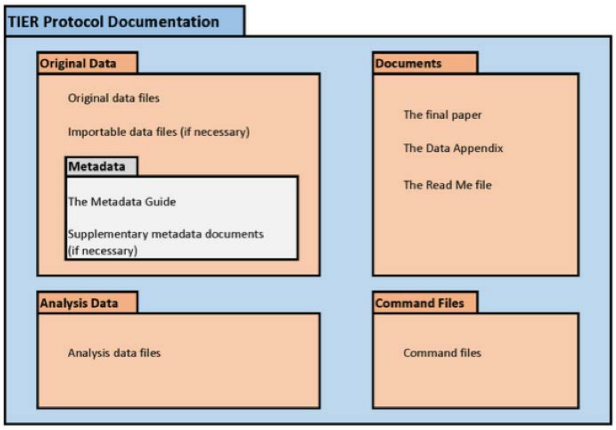
http://www.projecttier.org/tier-protocol/specifications/

The Library
UC SAN DIEGO

TIER File Structure



Folder Contents



Original Data folder: all original data, importable data files

Metadata sub folder: metadata guide, supplementary metadata documentation

Documentation folder: Readme file, data appendix, copy of final paper

Analysis Data: analysis data files

Command files folder: do-files or scripts used for data processing and analysis to reproduce results

The Library UC SAN DIEGO

http://www.projecttier.org/tier-protocol/specifications/

Documentation & Description

- Describe the method used to create derived data products.
- Consider creating templates for data collection.
- At the file level: Take consistent notes on file changes, name changes, dates of changes, etc.
- Include critical information, such as date or location, in the data table, not just as metadata embedded in the file name.

The Library
UC SAN DIEGO

Identifiers

- Personal identifiers ORCID
  - Create your ORCID
  - https://orcid.org/register

ORCID
Connecting Research and Researchers

- Digital Object Identifiers - DOIs
  - EZID https://ezid.cdlib.org/

The Library
UC SAN DIEGO

Metadata

Metadata is data about your data.

Creating metadata, i.e., information about your data's contents, structure, and permissions, makes it possible for others to find and use your data properly.

Without good metadata, you might not be able to reuse your own data five years from now!

The Library
UC SAN DIEGO

Data Clean-up

OpenRefine (http://openrefine.org/), for making sure records and variables are consistently coded, filling in known blanks, replacing text selectively, transforming data, and more.

The Library
UC SAN DIEGO

Who, me?

The Library
UC SAN DIEGO

Basic Storage

- Computers and shared servers can be good places for **temporary** storage of your working files.
- Store copies of data in open, **stable formats** (e.g., ascii, .txt, .csv, .pdf) for long term accessibility. . .
- Use flash drives **only for file transfer.**
- Cloud storage can be a convenient way to store and share temporary working files.
- For long-term storage, data should be put into well-managed **preservation system.**

The Library
UC SAN DIEGO

The Library
UC SAN DIEGO

Backup

- Rule of 3: Keep 2 copies onsite, 1 offsite.

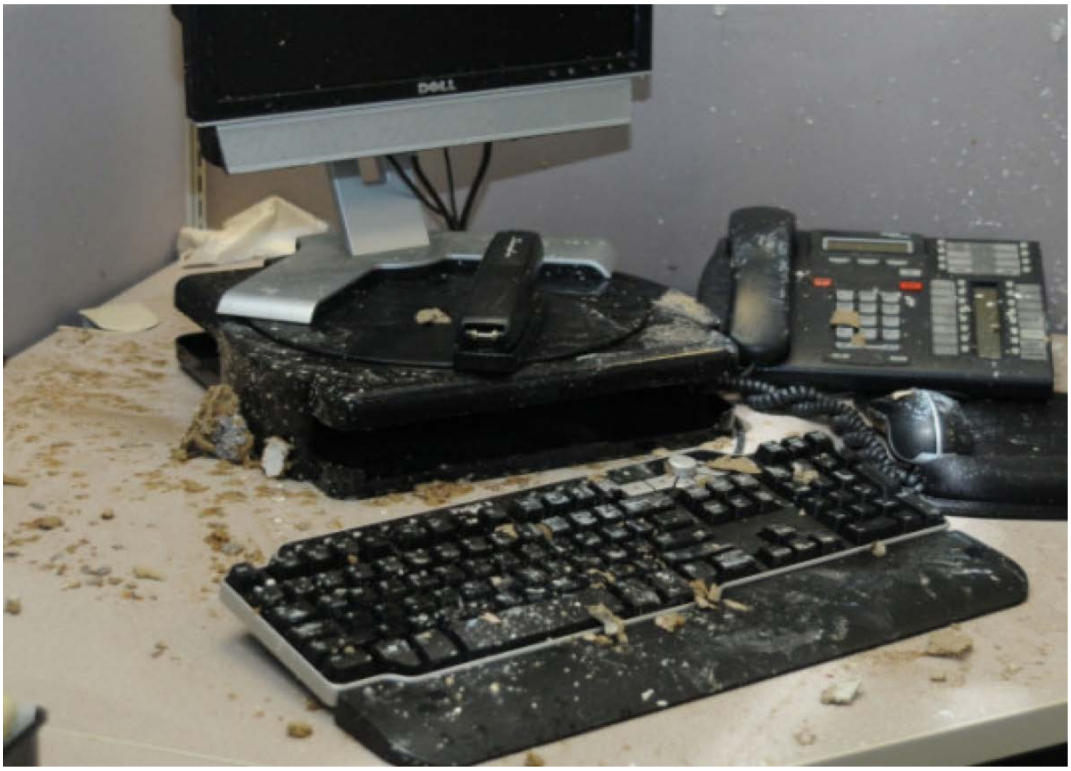
  LOCKSS concept – Lots Of Copies Keeps Stuff Safe

- Backup regularly and frequently - automate the process if possible.

Have a backup plan!

The Library
UC SAN DIEGO

The Library
UC SAN DIEGO



Preservation

- Preservation is the act of making sure your data are secure and accessible for future generations.
- Long-term preservation is not merely storage or backing up of your data.
- Identify data with long-term value. Preserve the raw data and any intermediate/derived/time consuming products that are expensive to reproduce or can be directly used for analysis.
- Preserve any scripted code and data that was used to clean and transform the raw data.
- Example:

  Save tabular data in a delimited text format.

  Save data in uncompressed and unencrypted formats, where possible.

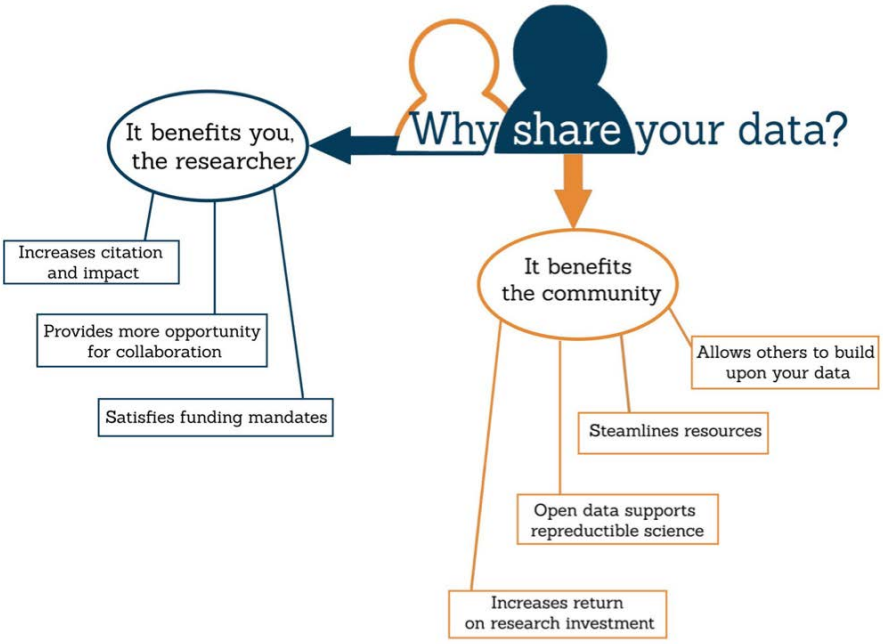
The Library
UC SAN DIEGO

The Library
UC SAN DIEGO

Data Sharing

Data sharing allows for reproducibility, transparency, and data re-use in research.

Sharing is easier if data are managed well from the start of a project.



The Library
UC SAN DIEGO

Research Data Curation Program

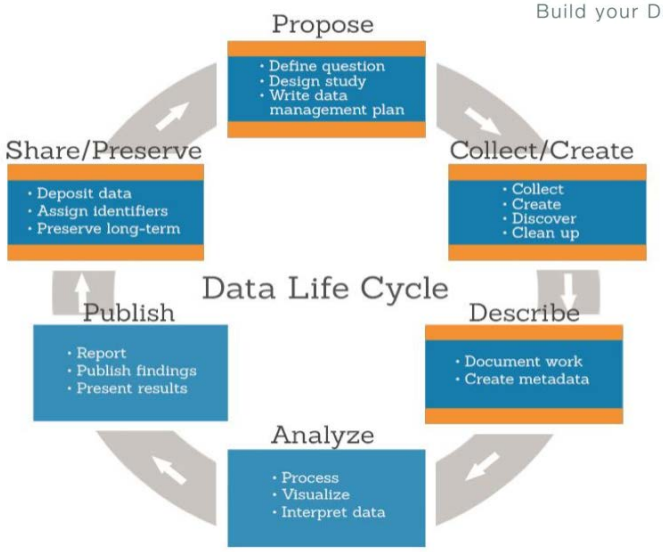
The Library UC SAN DIEGO

Support for the data life cycle:

DMPTool
Build your Data Management Plan

DIGITAL COLLECTIONS

CHRONOPOLIS



Working and Staging Storage

Large File Transfer Service
RDL-share.ucsd.edu

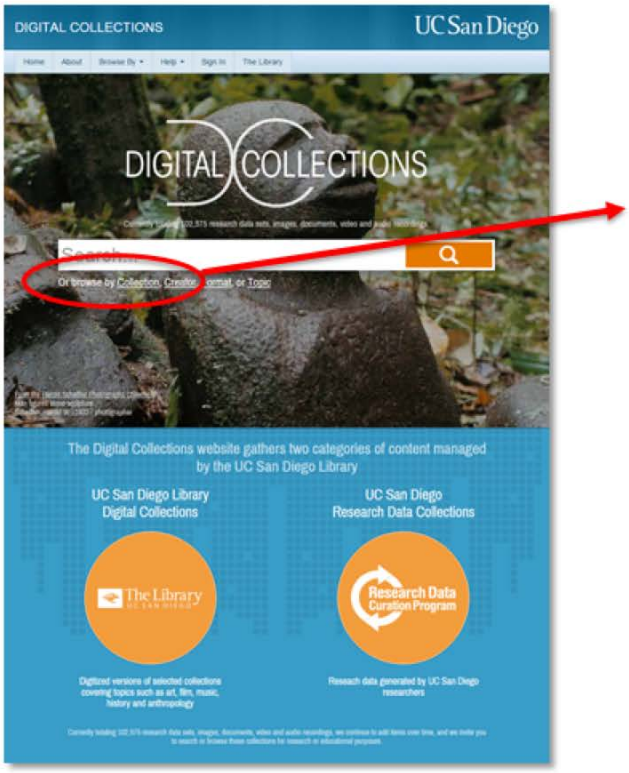
Open Science Framework

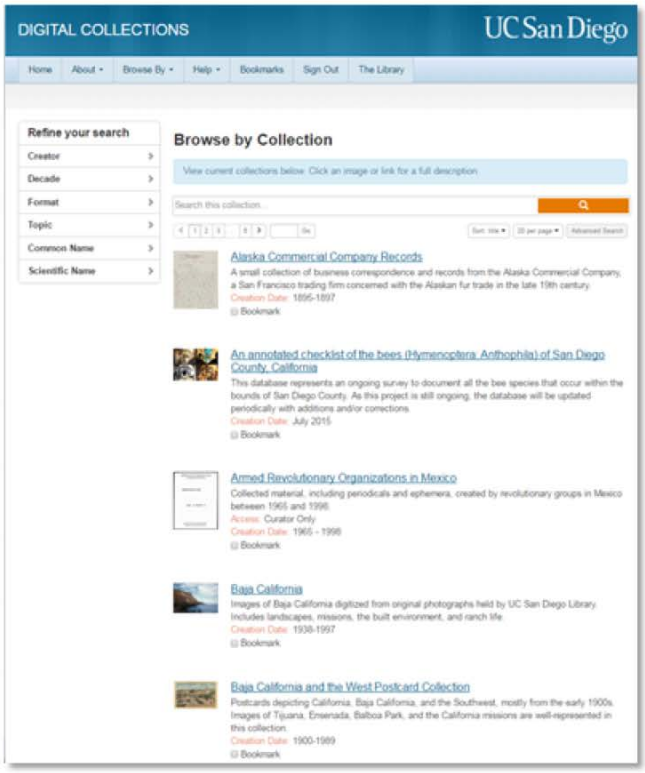
software carpentry

Data repository services at the Library

The Library UC SAN DIEGO

UC San Diego Library Digital Collections: http://library.ucsd.edu/dc

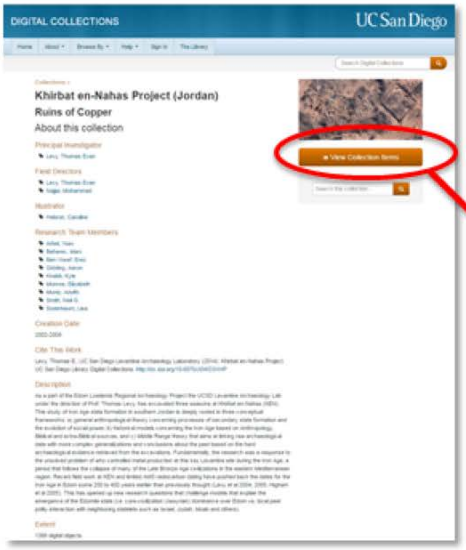




Data repository services at the Library

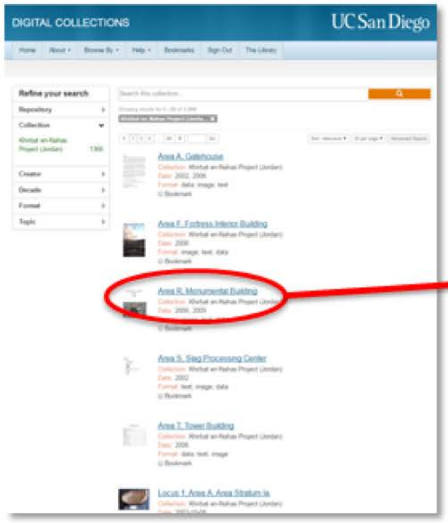
The Library UC SAN DIEGO

UC San Diego Library Digital Collections: http://library.ucsd.edu/dc

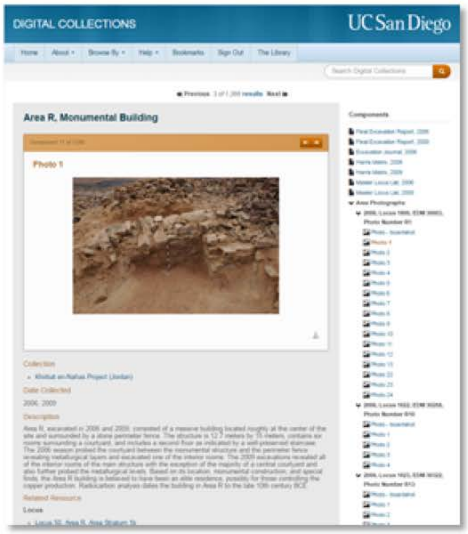


Collection page

List of objects in collection



Object page



Part 2: Center for Open Science

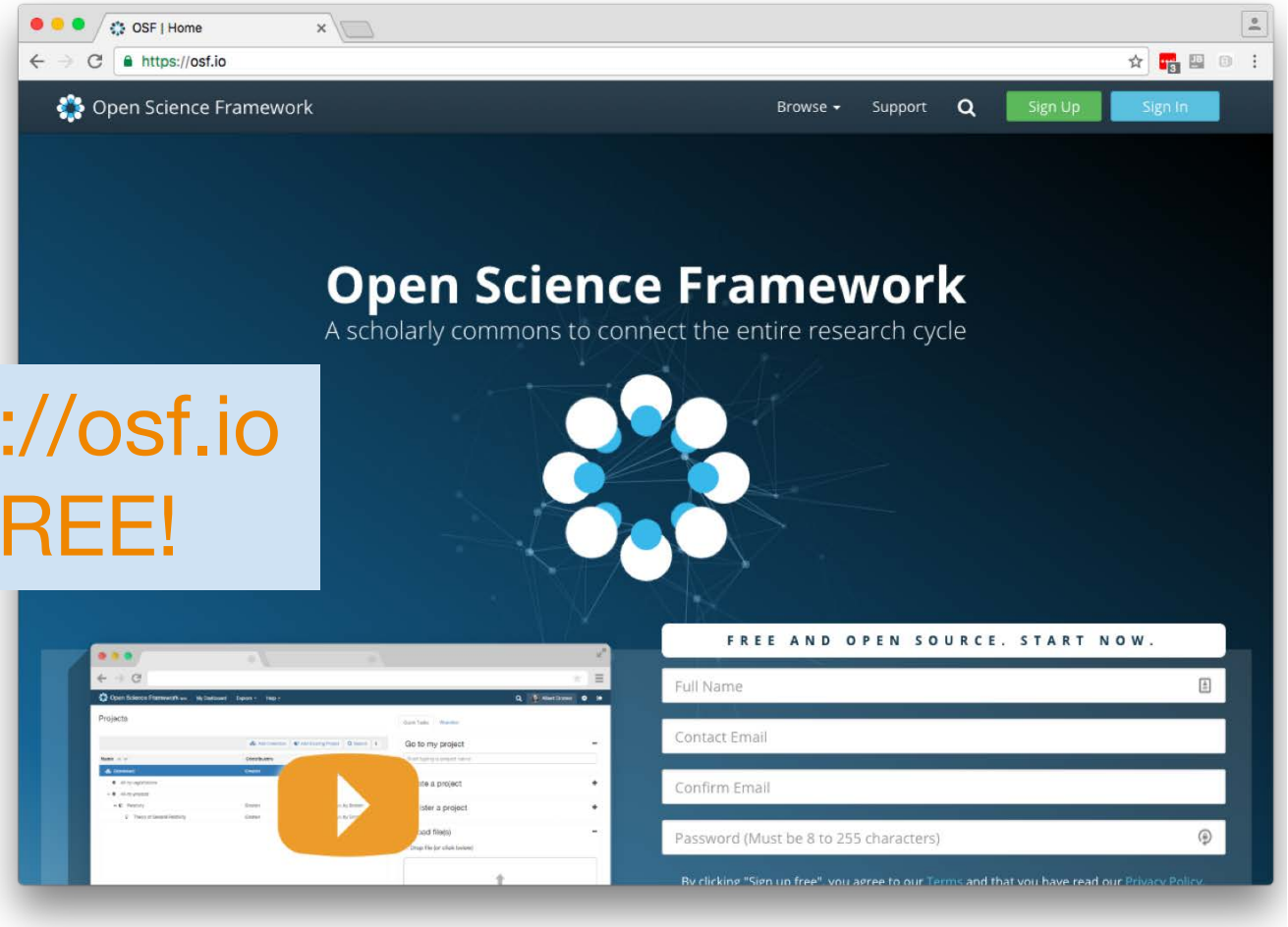
Open Science Framework

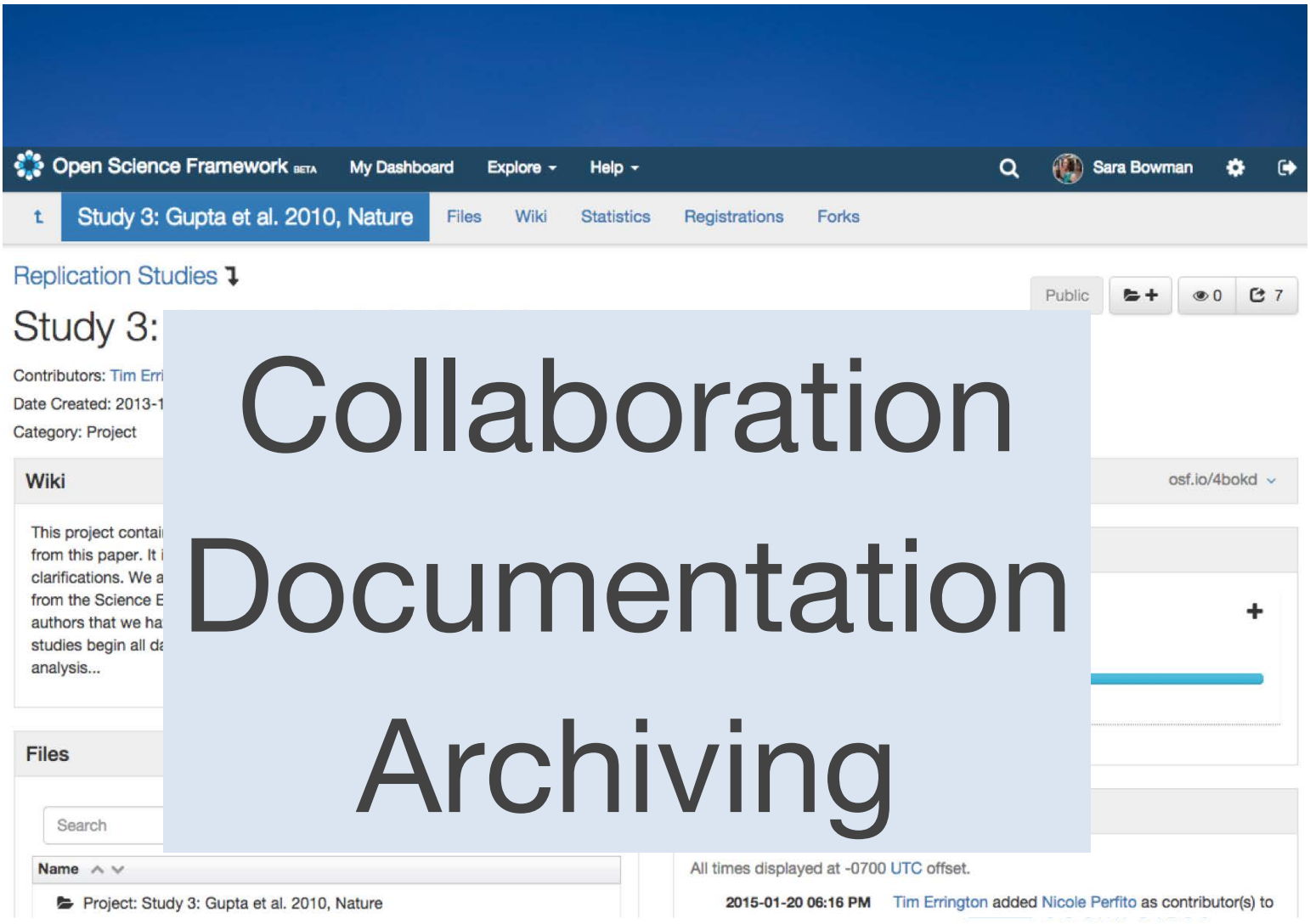
http://cos.io/ | http://osf.io

The Library
UC SAN DIEGO

Open Science Framework

http://osf.io
FREE!





Collaboration

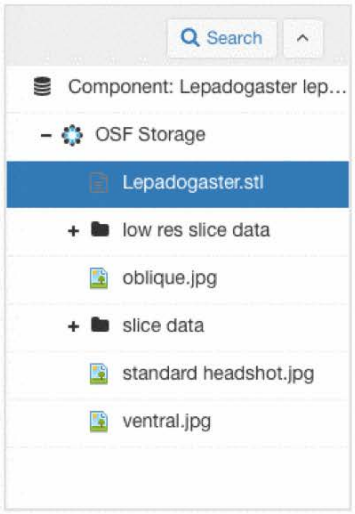
Documentation

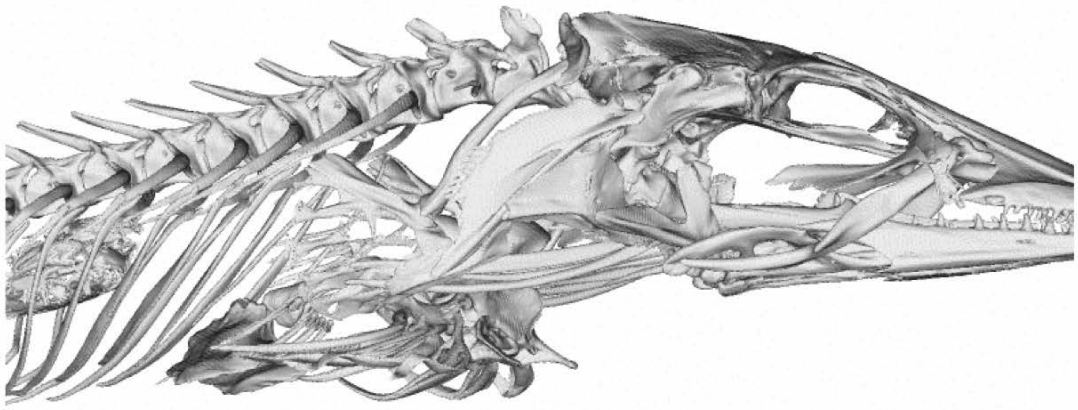
Archiving

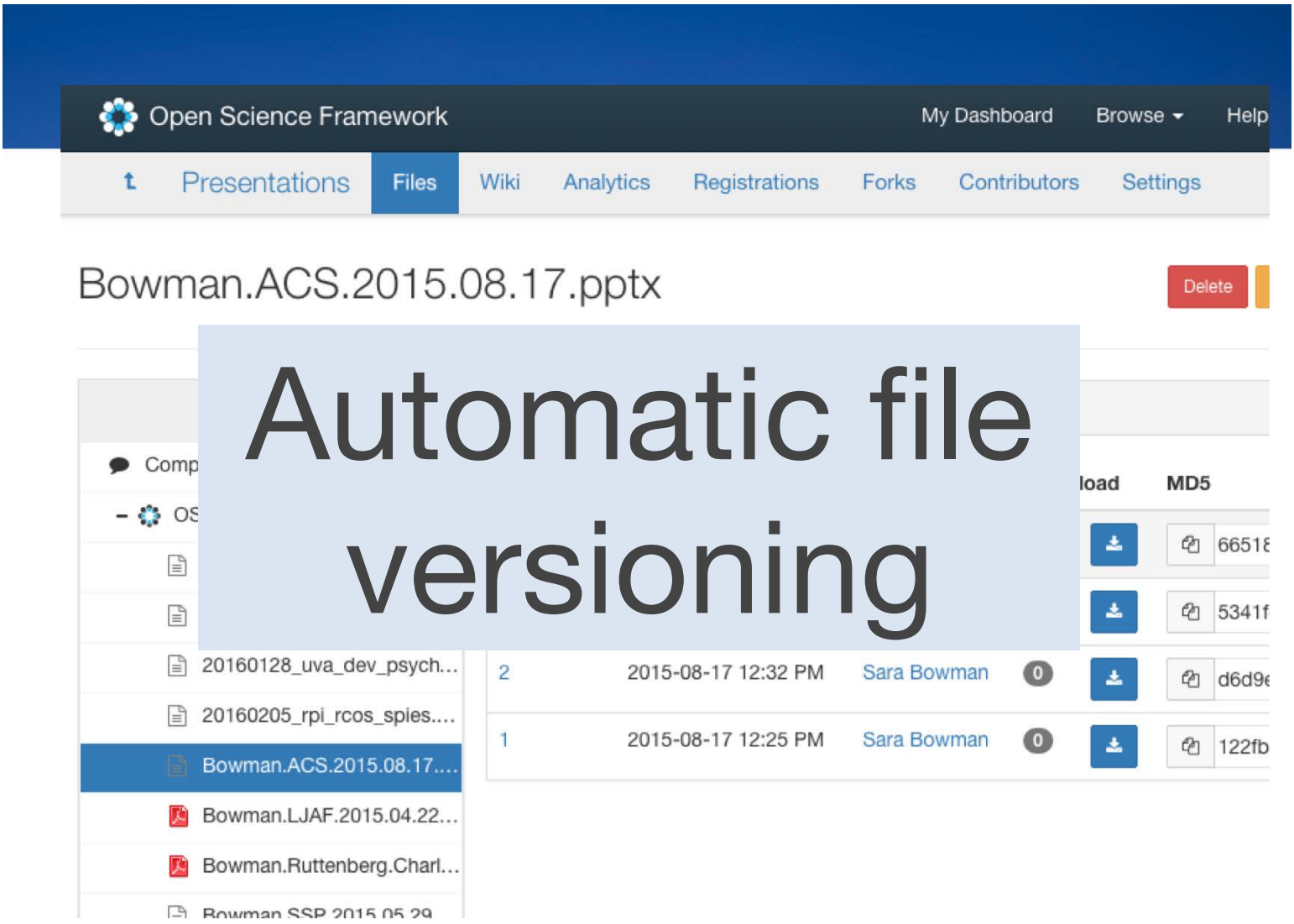
Put data, materials, and code on the OSF

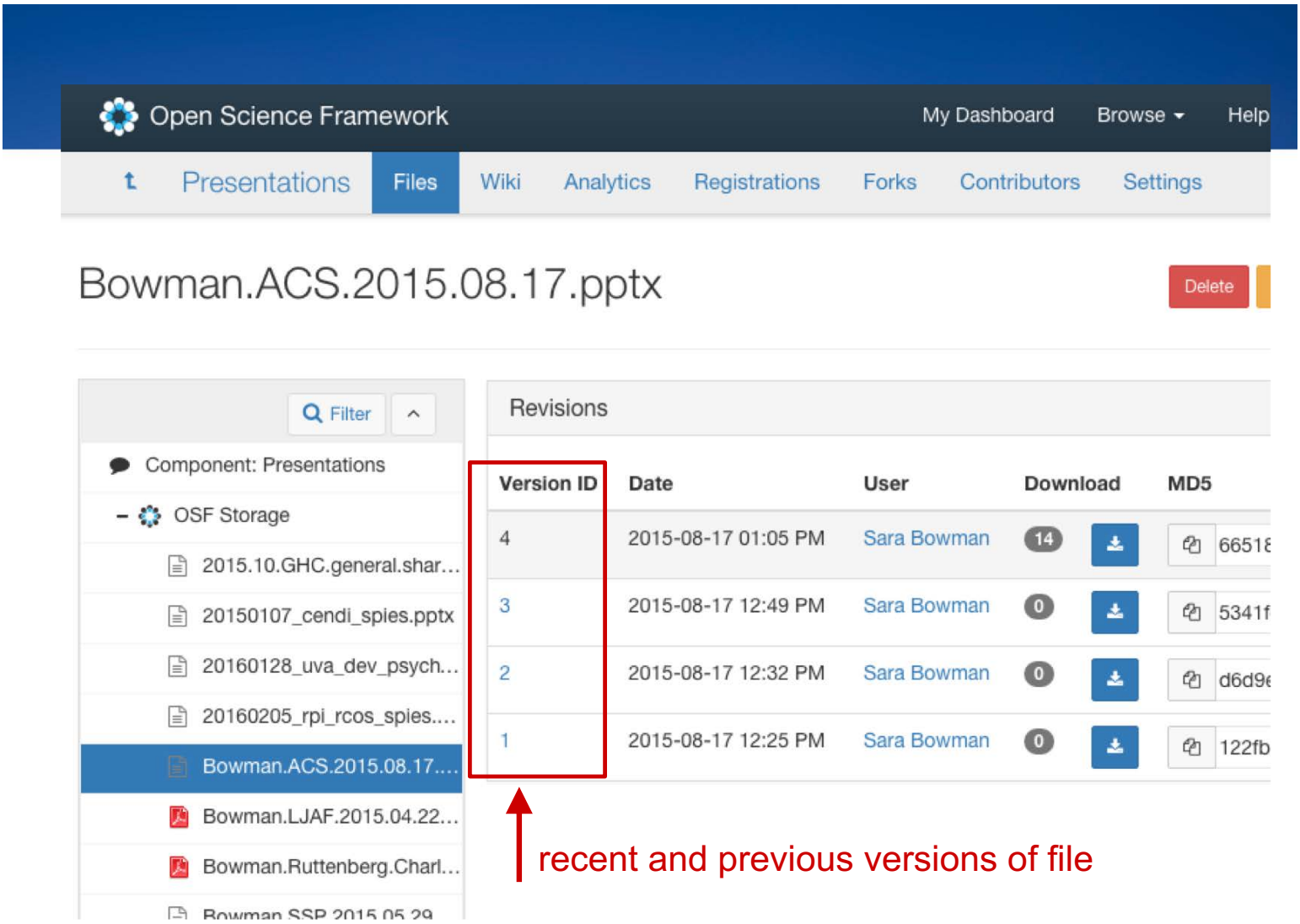
Lepadogaster.stl

Share | Download | View | Revisions






Open Science Framework
My Dashboard
Browse
Help
Presentations
Files
Wiki
Analytics
Registrations
Forks
Contributors
Settings
Bowman.ACS.2015.08.17.pptx
Delete
Automatic file versioning
MD5
20160128_uva_dev_psych...
20160205_rpi_rcos_spies....
Bowman.ACS.2015.08.17....
Bowman.LJAF.2015.04.22...
Bowman.Ruttenberg.Charl...
2
2015-08-17 12:32 PM
Sara Bowman
1
2015-08-17 12:25 PM
Sara Bowman


Open Science Framework
My Dashboard
Browse
Help
Presentations
Files
Wiki
Analytics
Registrations
Forks
Contributors
Settings
Bowman.ACS.2015.08.17.pptx
Delete
Filter
Component: Presentations
OSF Storage
2015.10.GHC.general.shar...
20150107_cendi_spies.pptx
20160128_uva_dev_psych...
20160205_rpi_rcos_spies....
Bowman.ACS.2015.08.17....
Bowman.LJAF.2015.04.22...
Bowman.Ruttenberg.Charl...
Revisions
Version ID
Date
User
Download
MD5
4
2015-08-17 01:05 PM
Sara Bowman
14
66518
3
2015-08-17 12:49 PM
Sara Bowman
0
5341f
2
2015-08-17 12:32 PM
Sara Bowman
0
d6d9e
1
2015-08-17 12:25 PM
Sara Bowman
0
122fb
recent and previous versions of file

Persistent Citable Identifiers

Citation: osf.io/

APA

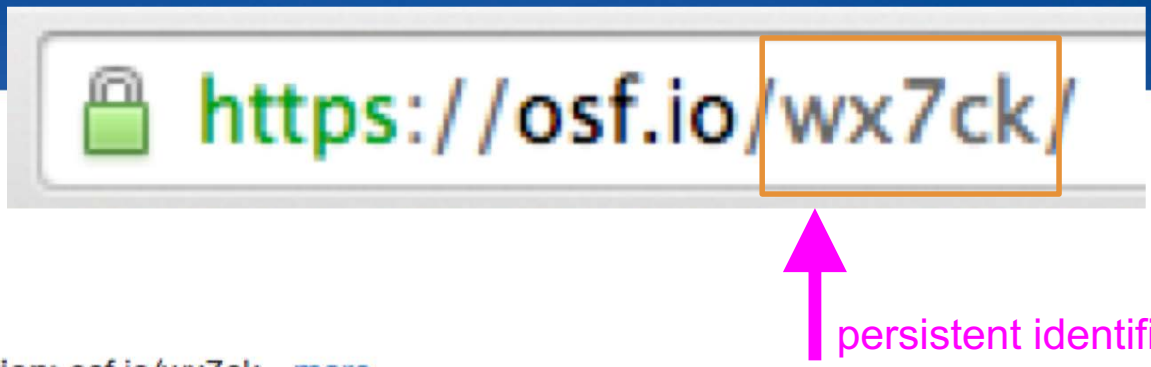
Klein, R. A., Rat… M. J., Bocian, K., et al. …abs" Replication Project. I…

MLA

Klein, R. A., Rat… M. J., Bocian, K., et al. …eplication Project."…

Chicago

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, , Bernstein, M. J., Bocian, K., et al. "Investigating Variation in Replicability: A "Many Labs" Replication Project." Open Science Framework (2014). osf.io/wx7ck



Citation: osf.io/wx7ck more

APA

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, , Bernstein, M. J., Bocian, K., et al. (2014). Investigating Variation in Replicability: A "Many Labs" Replication Project. Retrieved from Open Science Framework, osf.io/wx7ck

used in a citation

MLA

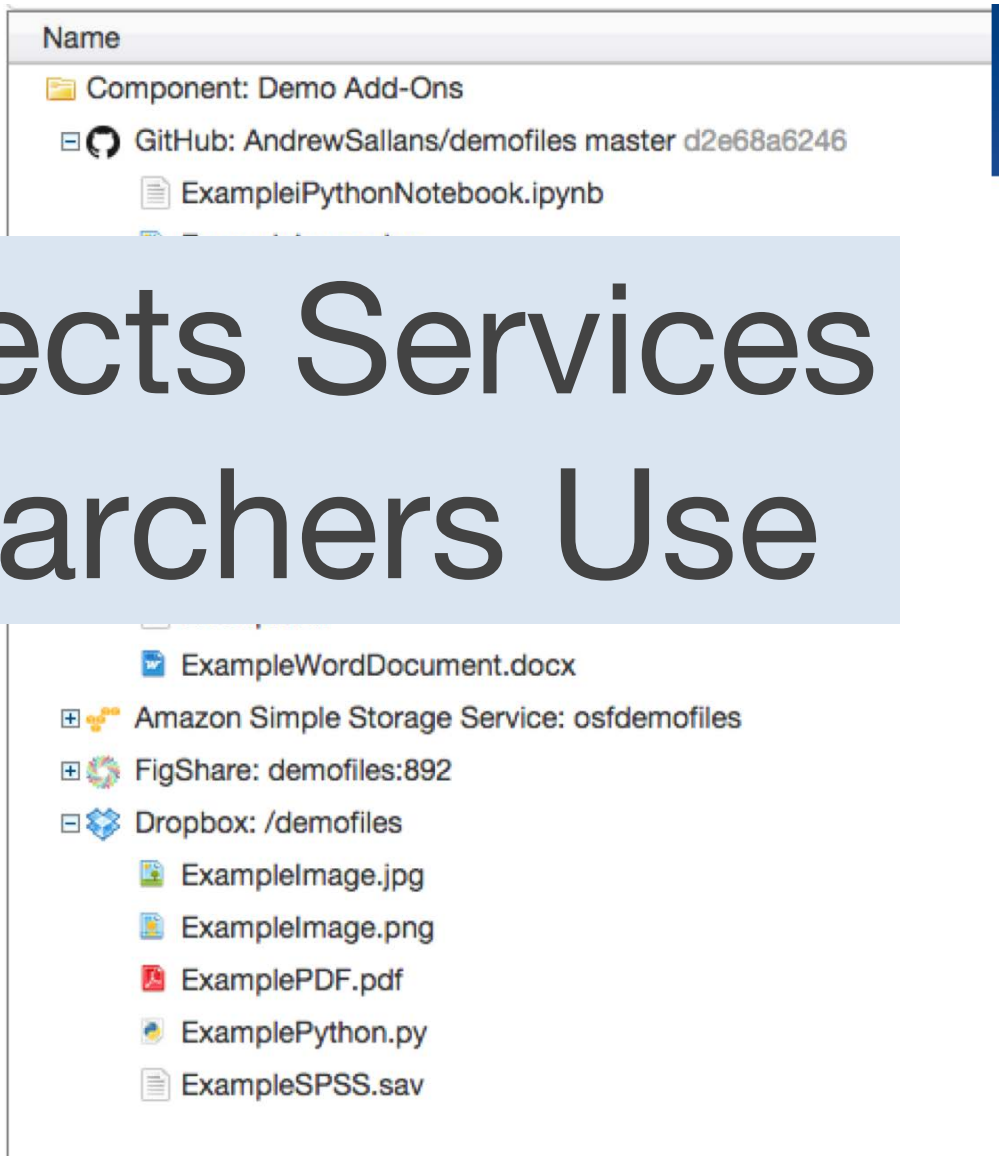
Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, , Bernstein, M. J., Bocian, K., et al. "Investigating Variation in Replicability: A "Many Labs" Replication Project." Open Science Framework, 2014. osf.io/wx7ck

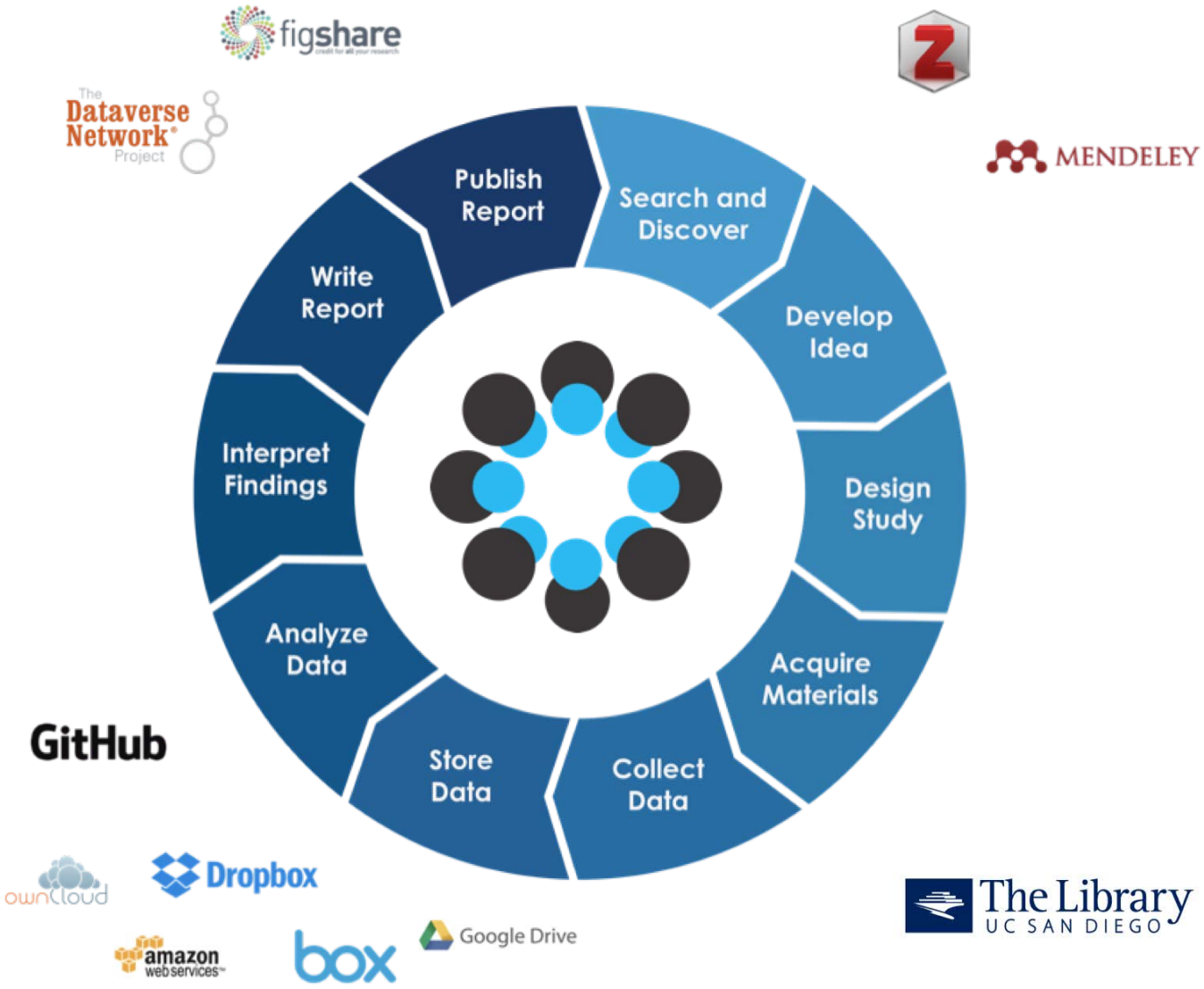
Chicago

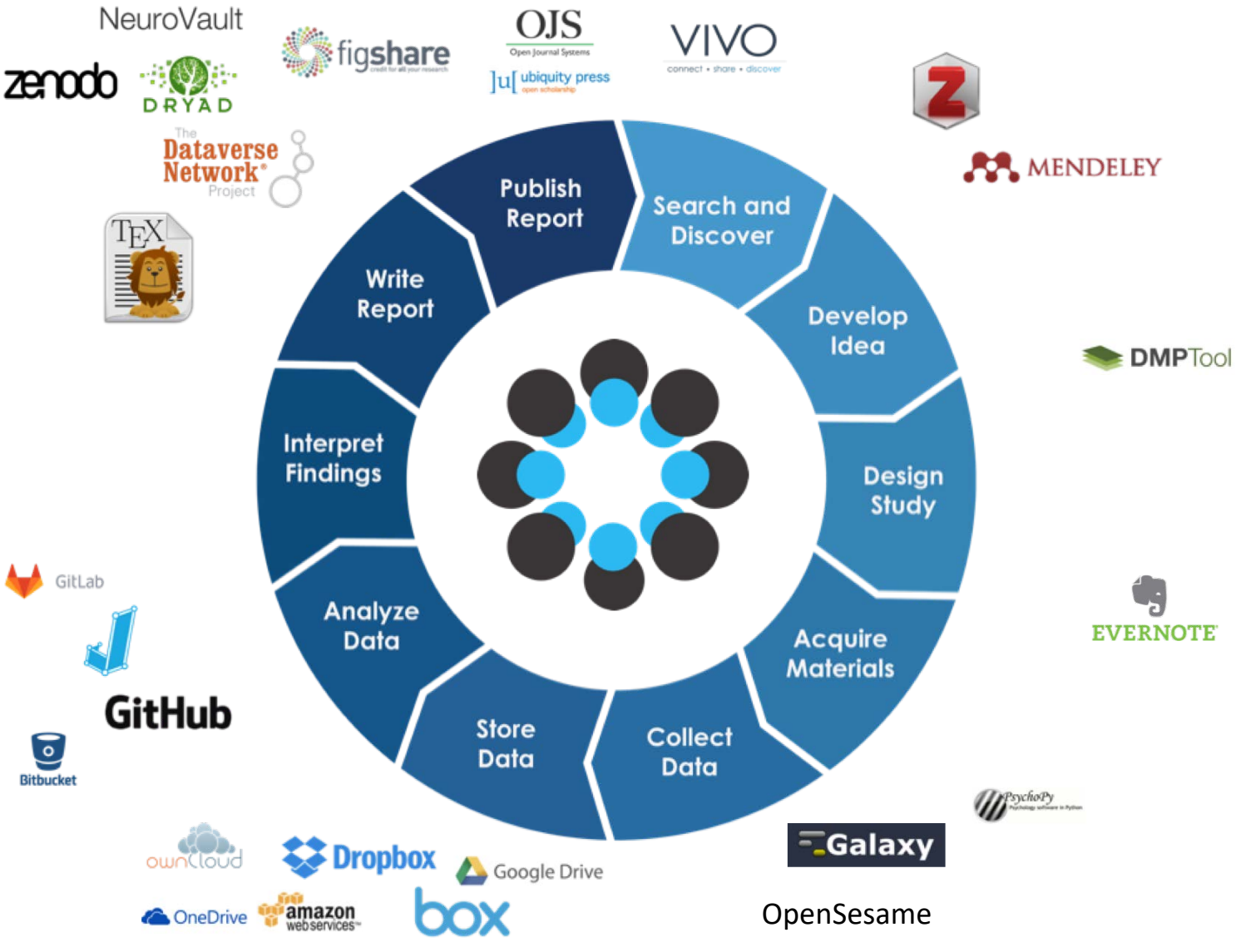
Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, , Bernstein, M. J., Bocian, K., et al. "Investigating Variation in Replicability: A "Many Labs" Replication Project." Open Science Framework (2014). osf.io/wx7ck


https://osf.io/wx7ck/
https://osf.io/c97pd/
Register
Is data collection for this project underway or complete?
all of the project materials, but
ntent and files cannot be deleted
lete and comprehensive for what
ue
Registration

Connects Services Researchers Use

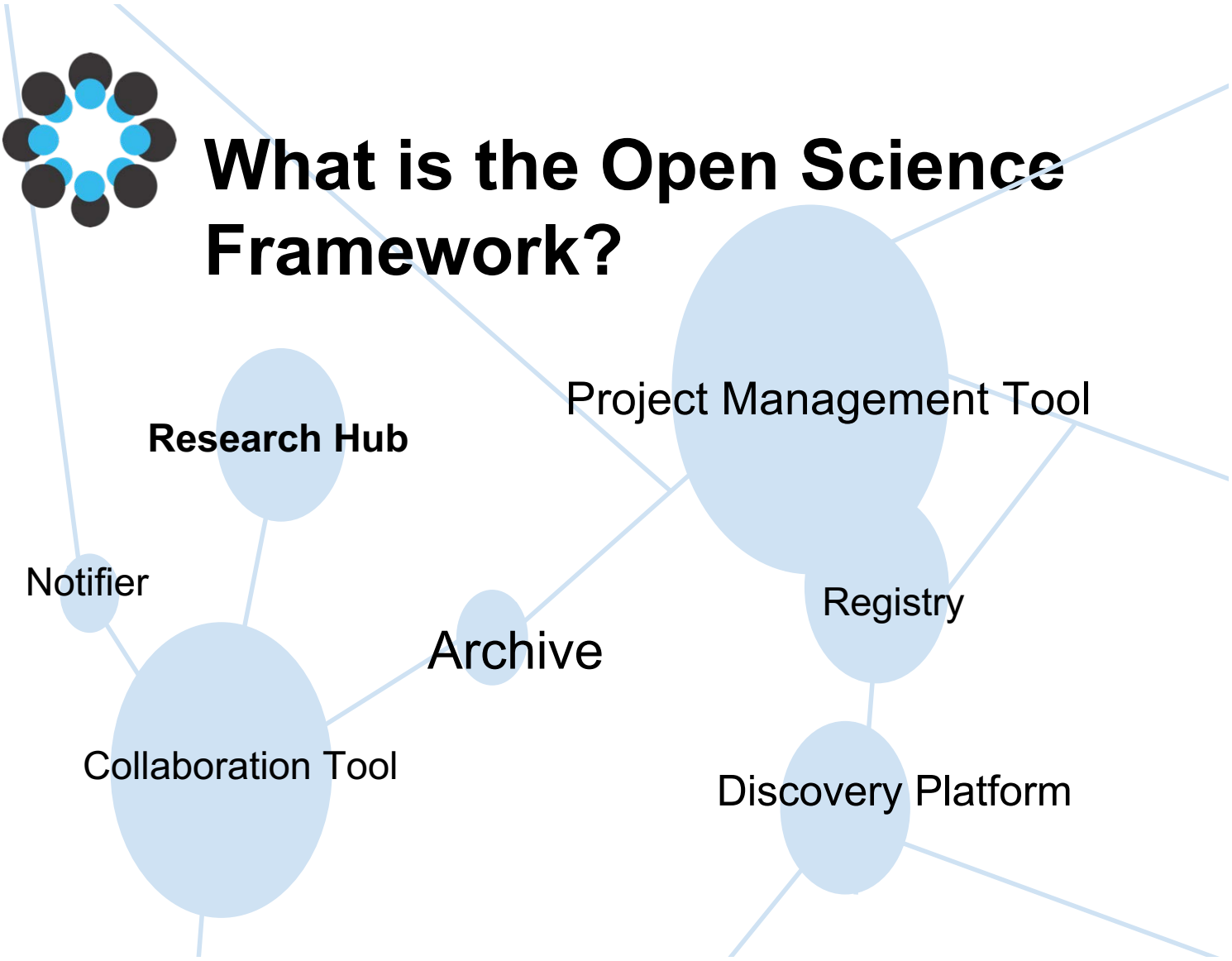



figshare
The Dataverse Network Project
MENDELEY
Publish Report
Search and Discover
Write Report
Develop Idea
Interpret Findings
Design Study
Analyze Data
Acquire Materials
Store Data
Collect Data
GitHub
owncloud
Dropbox
amazon web services
box
Google Drive
The Library UC SAN DIEGO


NeuroVault
zenodo
DRYAD
figshare
OJS
Open Journal Systems
ubiquity press
VIVO
connect • share • discover
The Dataverse Network Project
MENDELEY
TEX
Publish Report
Search and Discover
Develop Idea
Design Study
Acquire Materials
Collect Data
Store Data
Analyze Data
Interpret Findings
Write Report
DMPTool
GitLab
GitHub
Bitbucket
EVERNOTE
PsychoPy
Galaxy
ownCloud
Dropbox
Google Drive
OneDrive
amazon web services
box

OpenSesame

What is the Open Science Framework?



Questions?

The Library
UC SAN DIEGO

Hands-On with the OSF

Let's take a look at the OSF:

https://osf.io/

The Library
UC SAN DIEGO

Resources

UC San Diego Library
Research Data Curation Program
https://library.ucsd.edu/research-and-collections/data-curation/index.html

Digital Collections
http://library.ucsd.edu/dc

TIER Protocol Specification
http://www.projecttier.org/tier-protocol/specifications/

The Library
UC SAN DIEGO