Data Management & The Open Science Framework

Reid Otsuji
RRROBOTS Course
Scripps Institution of Oceanography
May 16, 2017

The Library
UC SAN DIEGO

Introduction

Part 1:

A. Why is data management important?

B. Best practices to Consider

Part 2:

Open Science Framework Demo

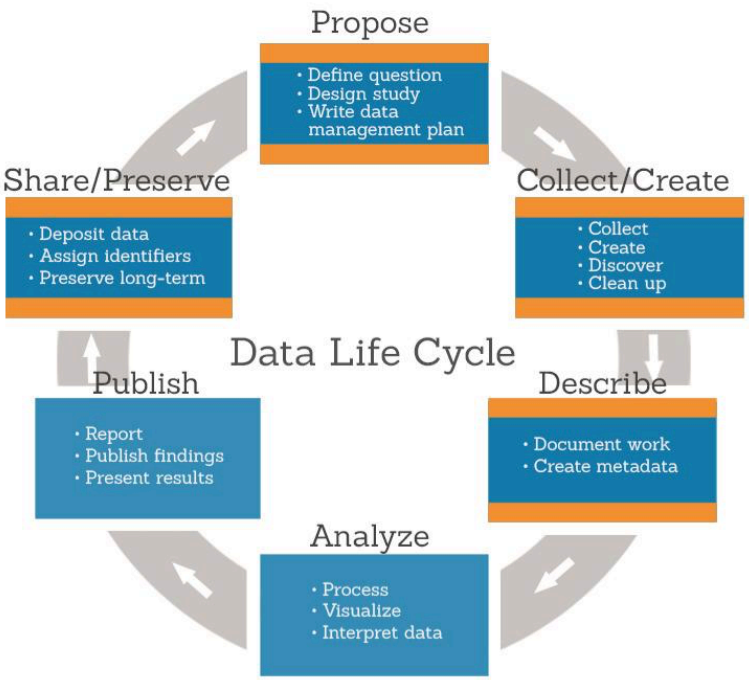
The Library
UC SAN DIEGO

Part 1a

Why Manage Data?

The Library
UC SAN DIEGO

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

A good data management strategy:

- Establish best practices for your data management
- Plan to share well-documented data
- Well prepared data saves time
- Create a concise data management plan for your grant proposal
- Reduces cost of creating, protecting and storing data

Ensures your data will be available to future generations to enable reproducible research

The Library
UC SAN DIEGO

Benefits

Benefits of good data management:

- Promotes successful data collection.
- Ease of using and sharing data.
- Helps to increase research impact and visibility.
- Standardize data management practices and policies in your research lab.

Saves you time, effort and resources during the research project.

The Library
UC SAN DIEGO

Best Practices

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

The Library
UC SAN DIEGO

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 by Haverford College –

The **Teaching Integrity in Empirical Research** or TIER protocol, is a recommended protocol for comprehensively documenting all the steps of data management and analysis that go into an empirical research paper.

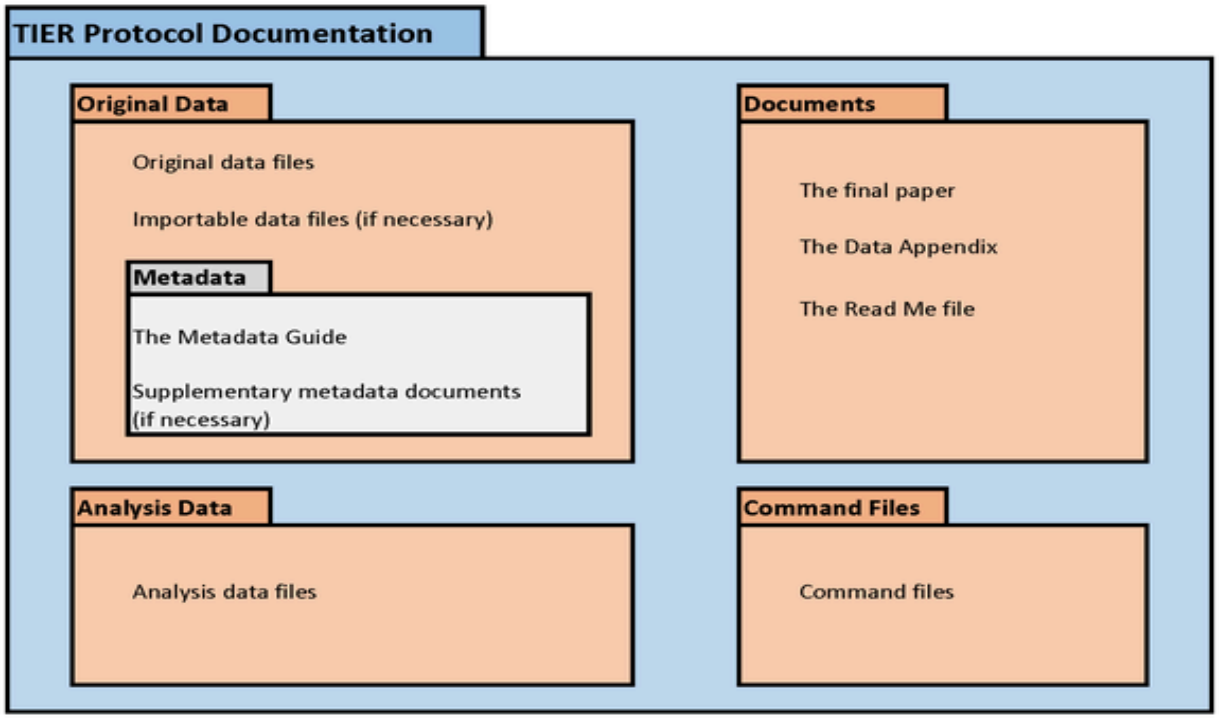
All documentation, do-files, scripts, raw data, metadata, that are presented in a paper are organized in a specific file structure.

This file structure keeps your data organized and offers easy replication of results reported in a paper.

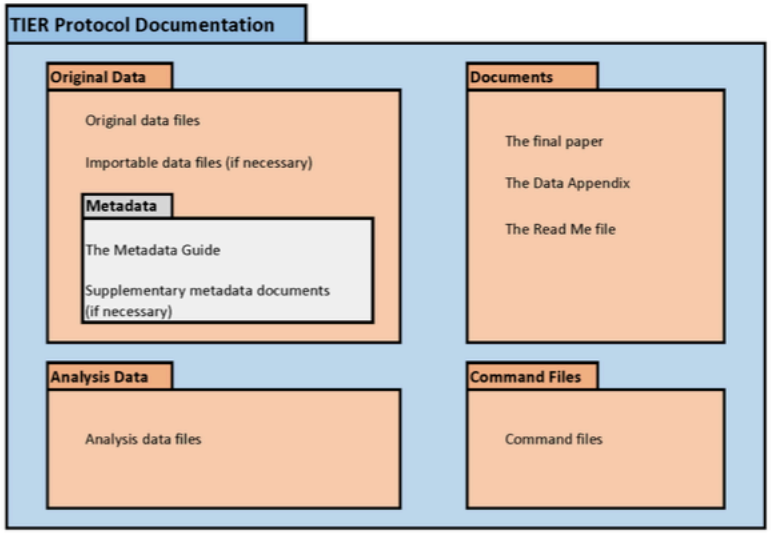
http://www.projecttier.org/tier-protocol/specifications/

The Library
UC SAN DIEGO

TIER File Structure



Folder Contents



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

Original Data folder: all original data, importable data files

Metadata sub folder: metadata guide, supplementary metadata documentation

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

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.

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

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

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.

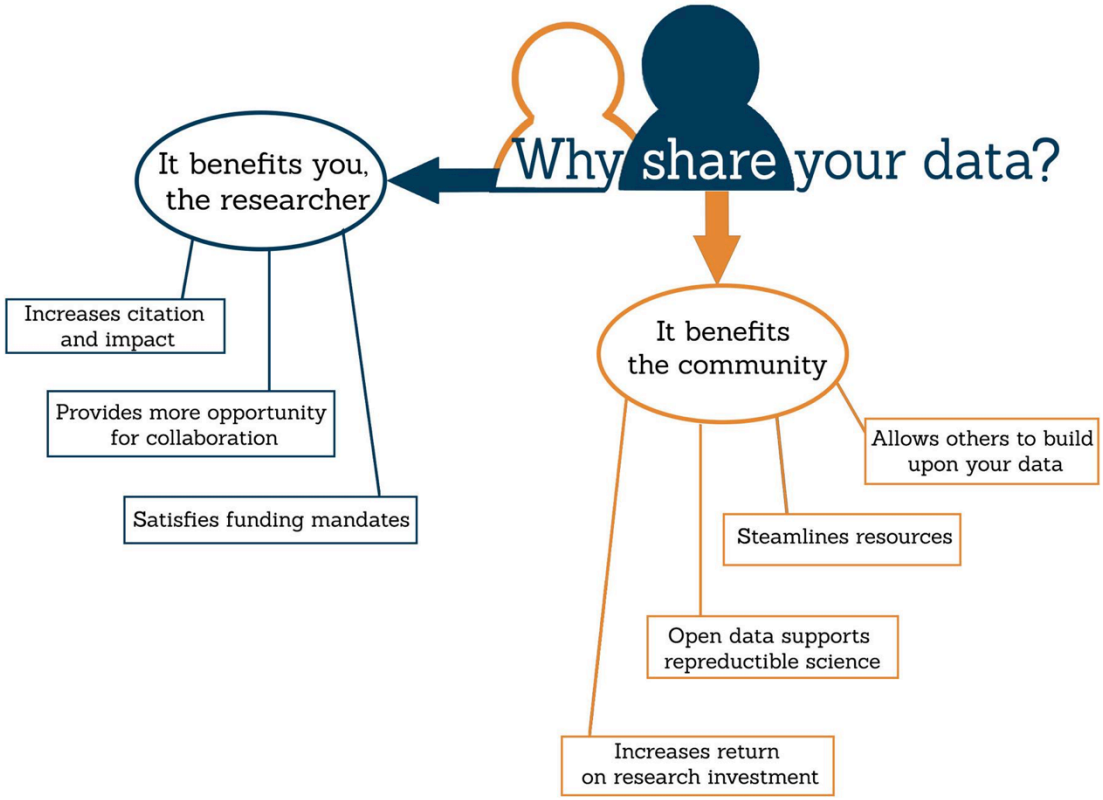
Save data in uncompressed and unencrypted formats, where possible.

The Library
UC SAN DIEGO

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

Part 2: Center for Open Science

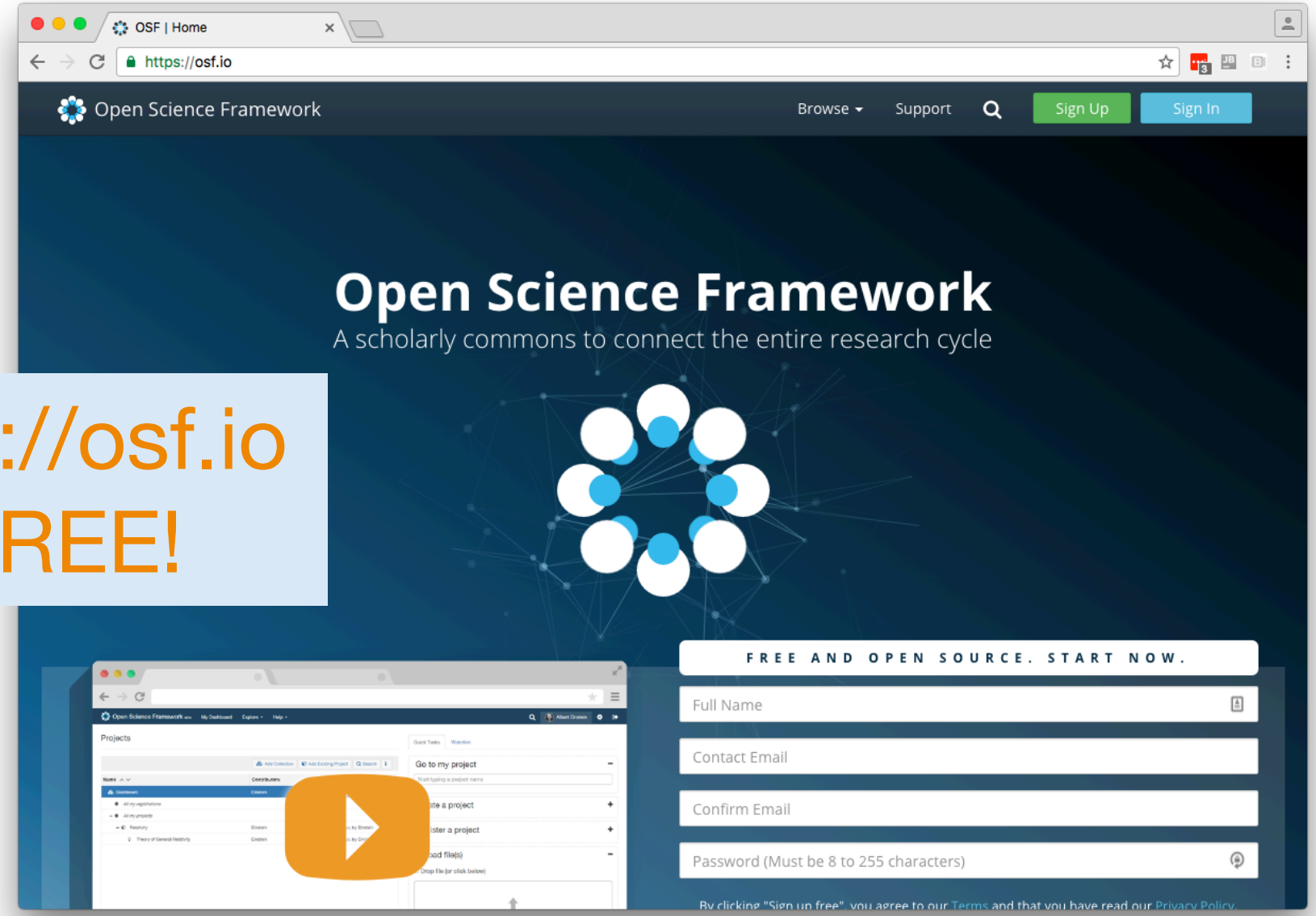
Open Science Framework

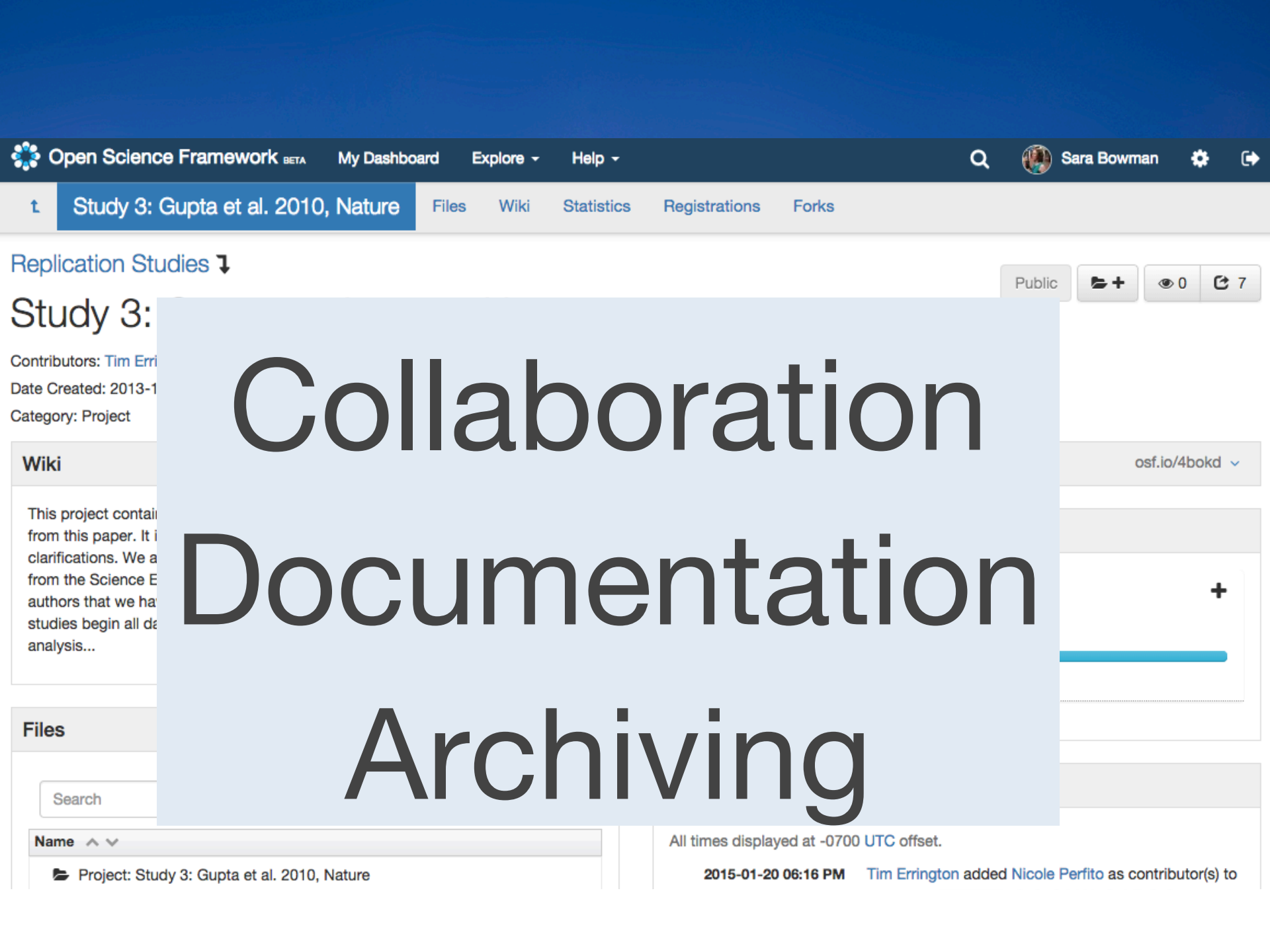
http://cos.io/ | http://osf.io

The Library
UC SAN DIEGO

Open Science Framework

http://osf.io
FREE!




Open Science Framework BETA
My Dashboard
Explore
Help
Sara Bowman
Study 3: Gupta et al. 2010, Nature
Files
Wiki
Statistics
Registrations
Forks
Replication Studies
Public
0
7
Study 3:
Contributors: Tim Erri
Date Created: 2013-1
Category: Project
Wiki
osf.io/4bokd
This project contai
from this paper. It i
clarifications. We a
from the Science E
authors that we ha
studies begin all da
analysis...
Collaboration
Documentation
Archiving
Files
Search
Name
Project: Study 3: Gupta et al. 2010, Nature
All times displayed at -0700 UTC offset.
2015-01-20 06:16 PM
Tim Errington added Nicole Perfito as contributor(s) to

Put data, materials, and code on the OSF

Lepadogaster.stl

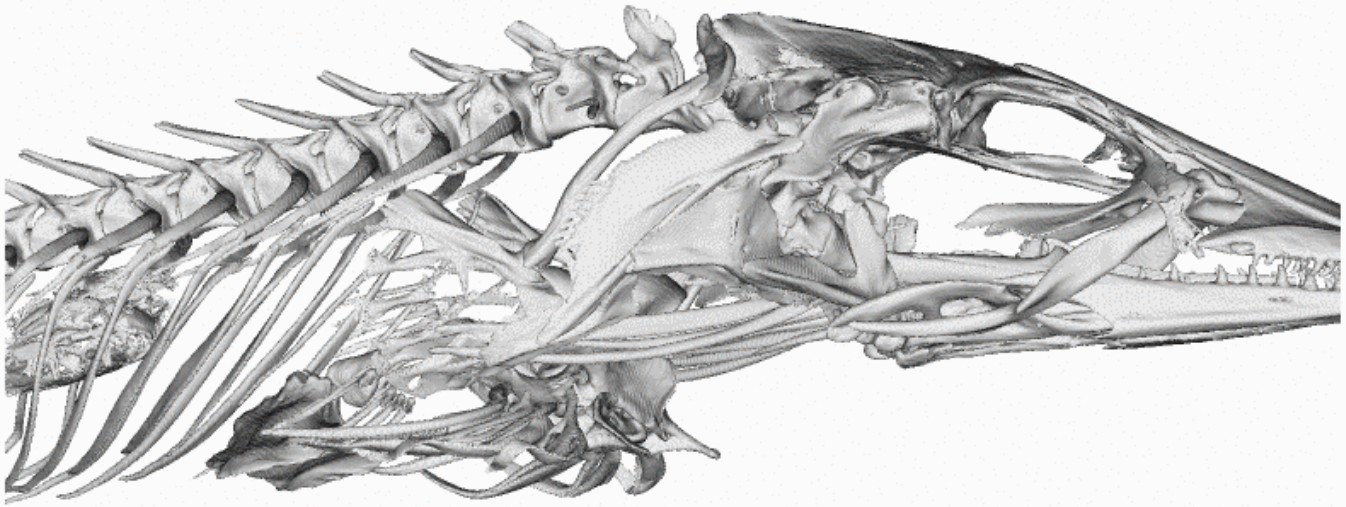
Share | Download | View | Revisions

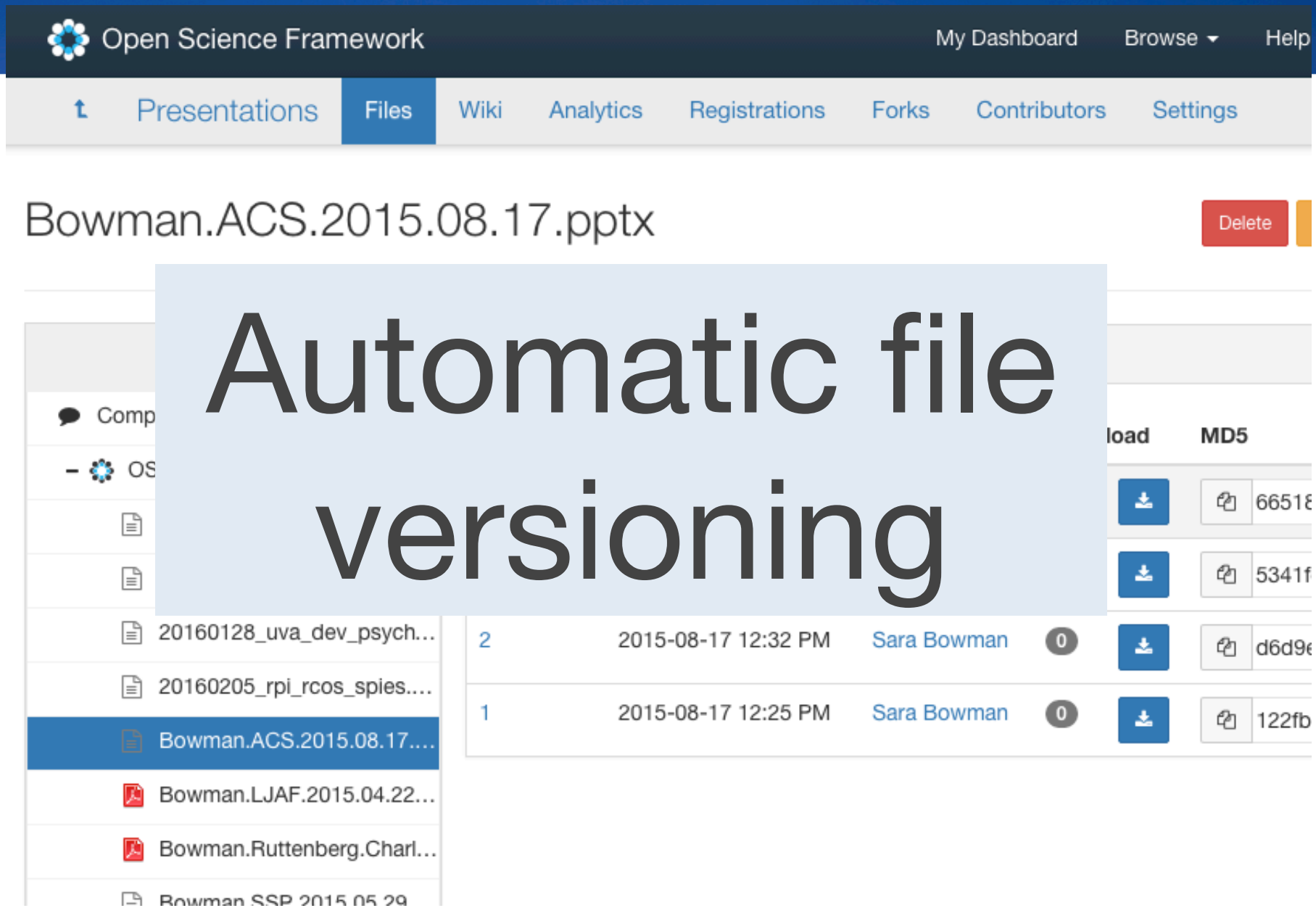
Search

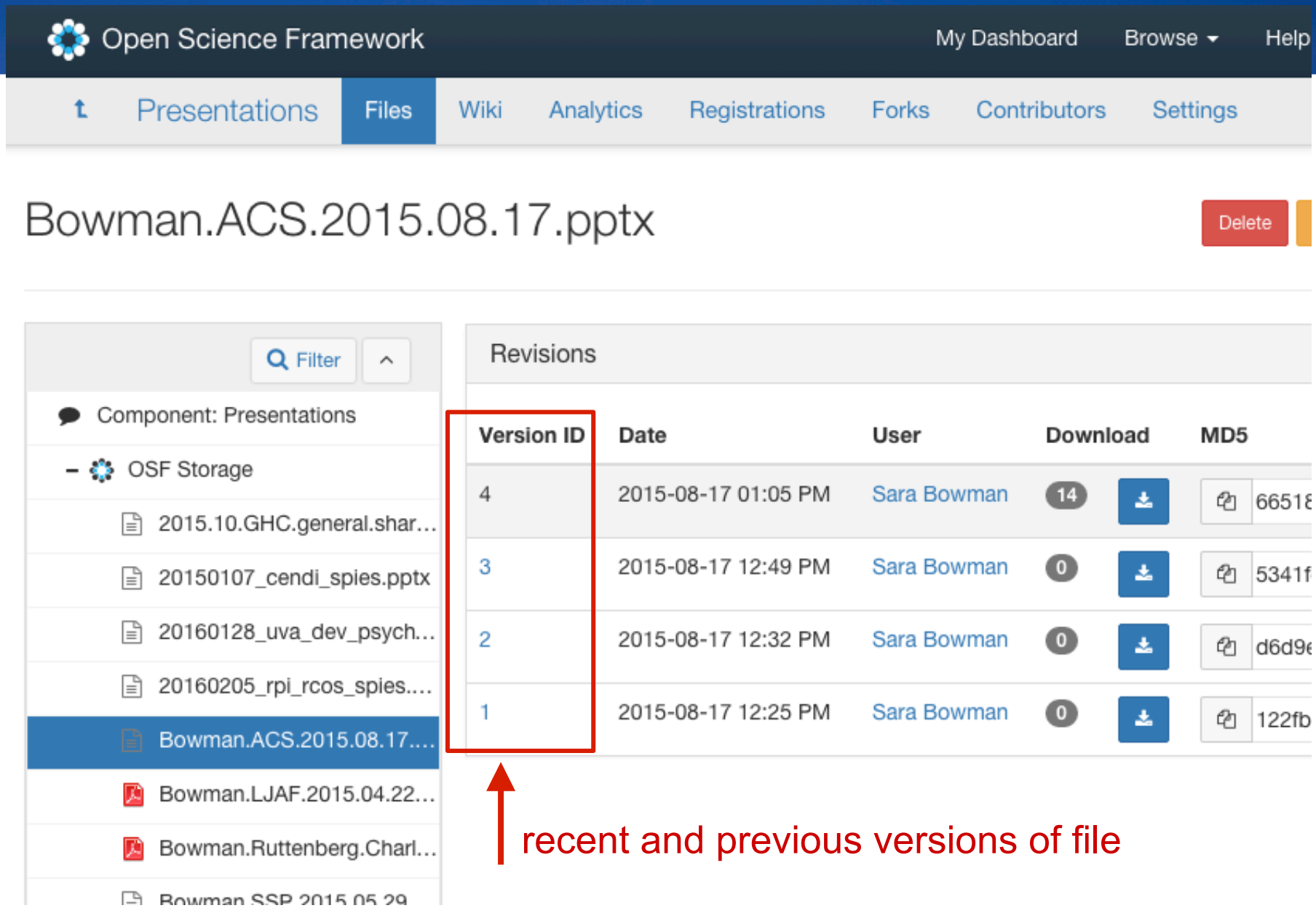
Component: Lepadogaster lep…

OSF Storage

- Lepadogaster.stl
- low res slice data
- oblique.jpg
- slice data
- standard headshot.jpg
- ventral.jpg




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
Comp
OS
20160128_uva_dev_psych...
20160205_rpi_rcos_spies....
Bowman.ACS.2015.08.17....
Bowman.LJAF.2015.04.22...
Bowman.Ruttenberg.Charl...
load
MD5
2
2015-08-17 12:32 PM
Sara Bowman
0
d6d9e
1
2015-08-17 12:25 PM
Sara Bowman
0
122fb


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

persistent identifier

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

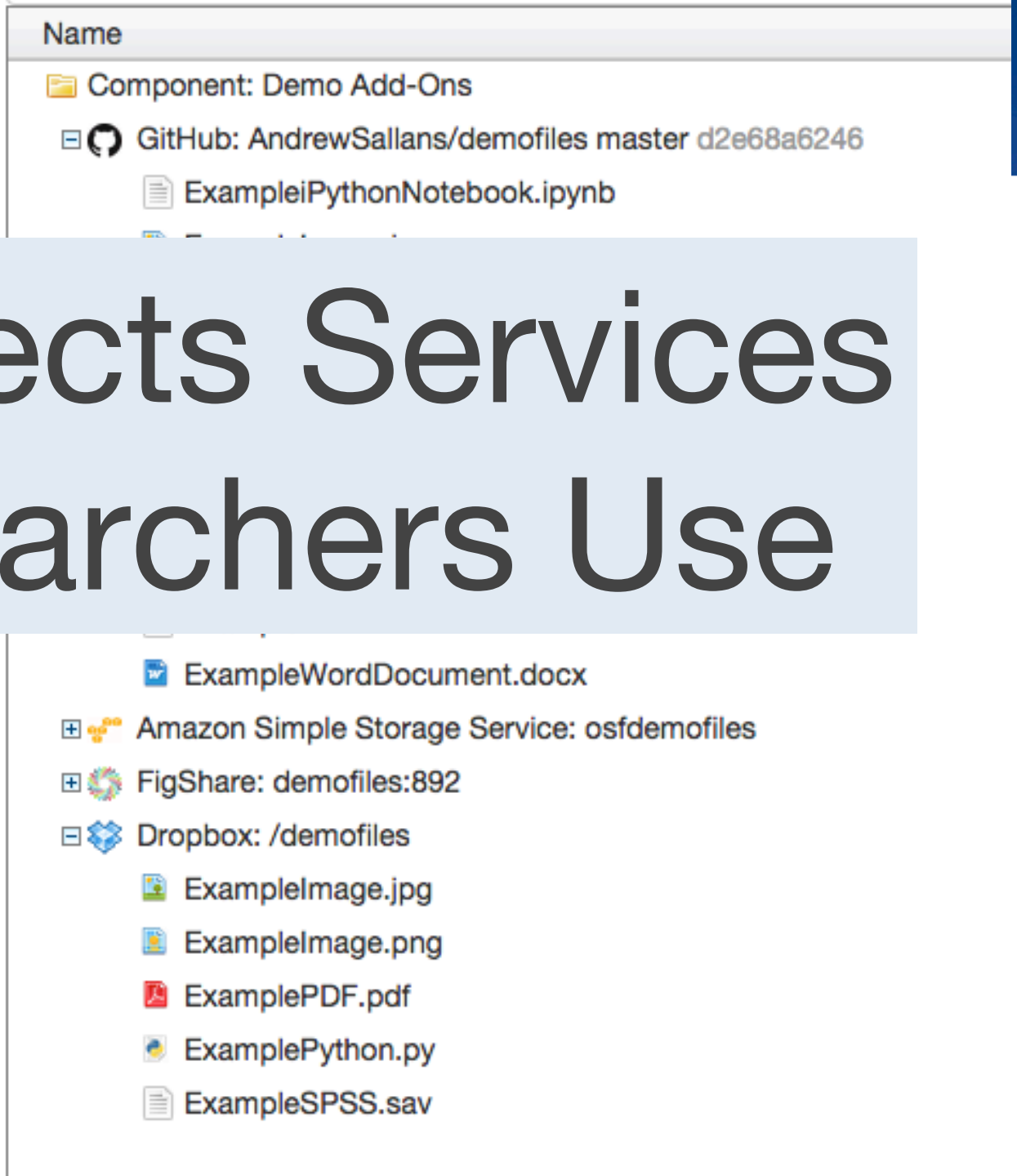
Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, , Bernstein, M. J., Bocian, K., et al. "Investigating Variation in Replicability: A "Many Labs" Replication Project." Open Science Framework, 2014. osf.io/wx7ck

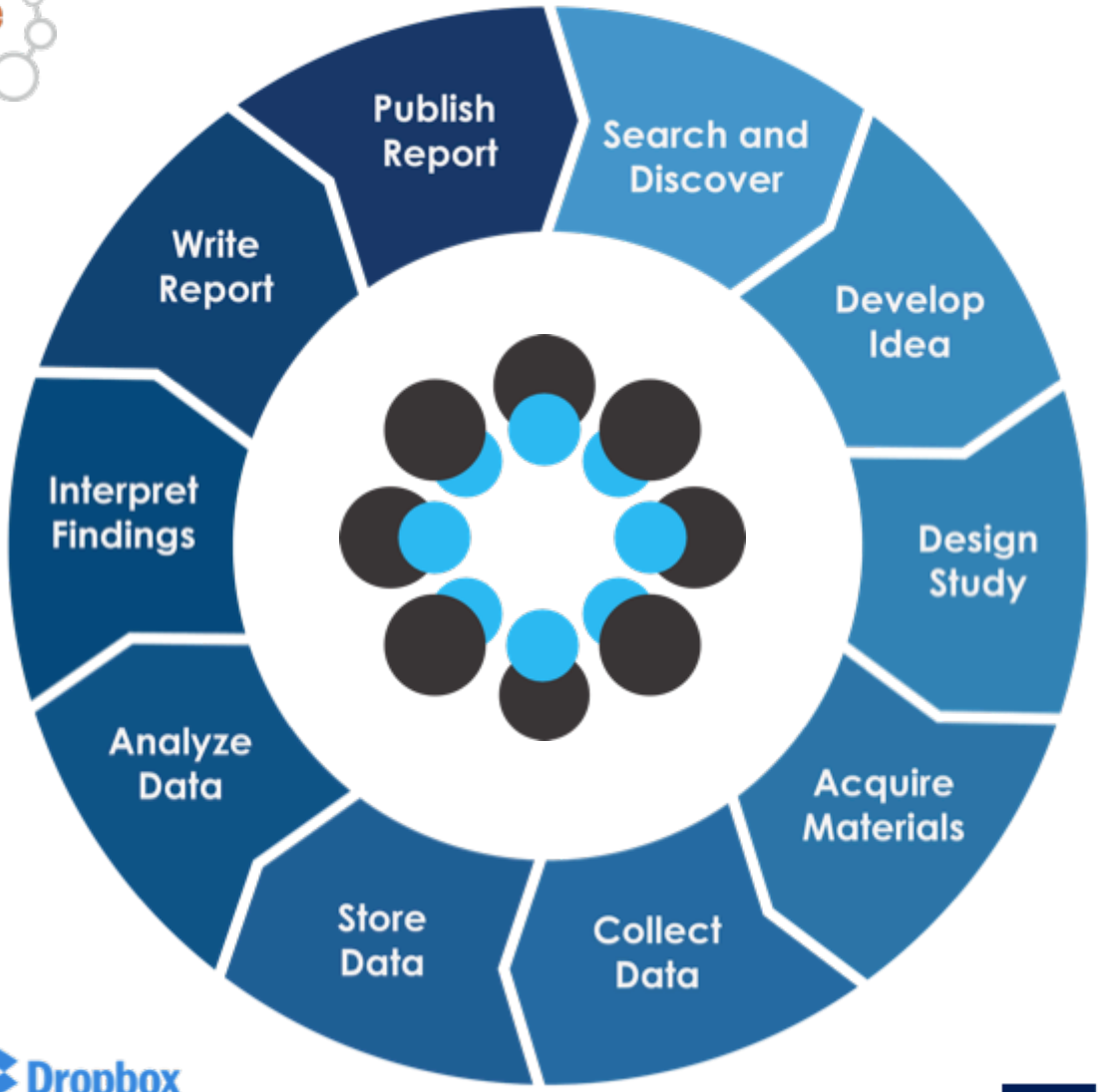
Chicago

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

GitHub

ownCloud

Dropbox

amazon web services

box

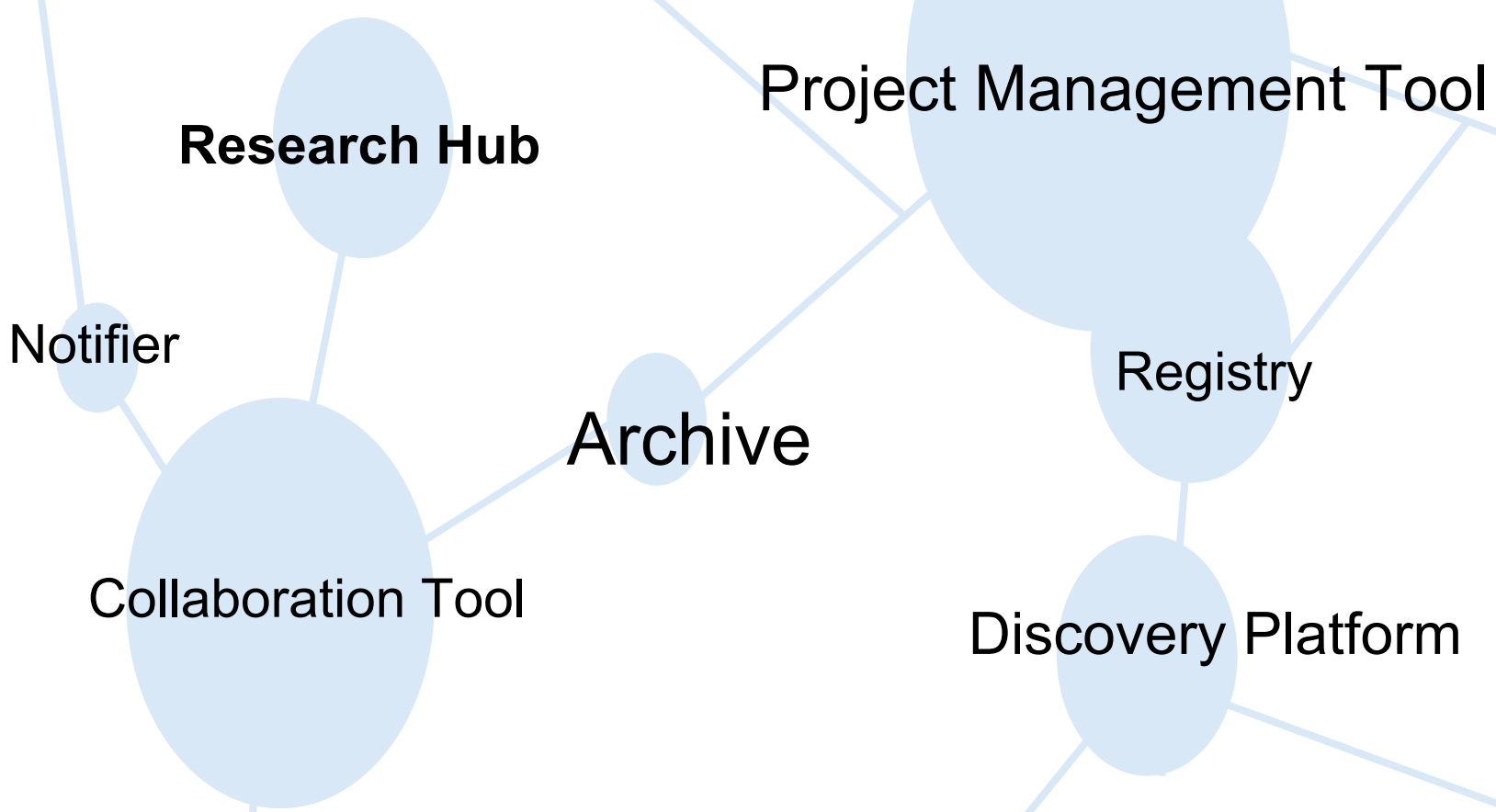
Google Drive

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
DMPTool
GitLab
GitHub
Bitbucket
EVERNOTE
owncloud
Dropbox
Google Drive
OneDrive
amazon web services
box
PsychoPy
Galaxy

OpenSesame

What is the Open Science Framework?



Hands-On with the OSF

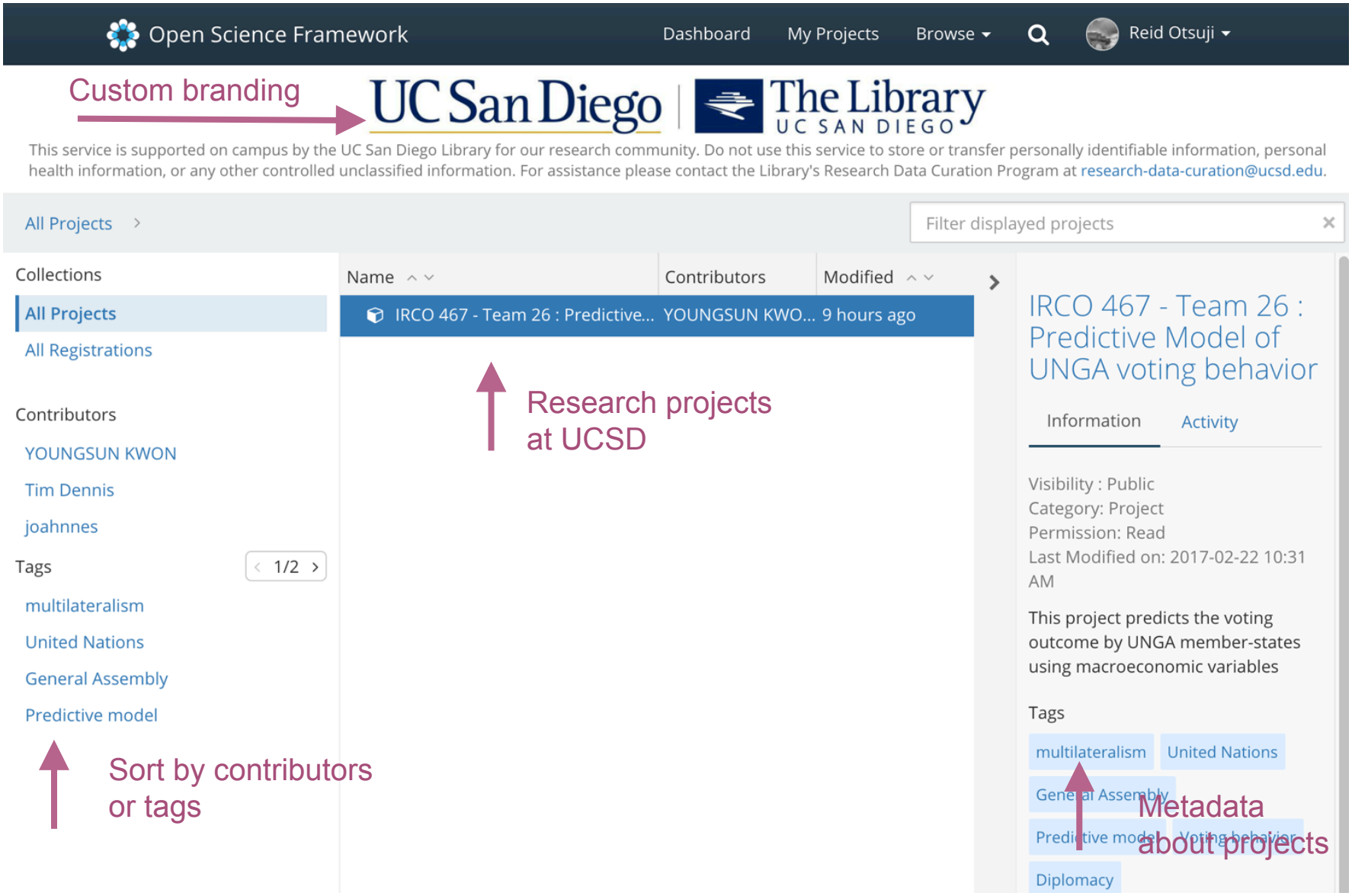
Let's take a look at the:
https://osf.io/

The Library
UC SAN DIEGO

OSF for Institutions

The Library
UC SAN DIEGO

OSF for Institutions



OSF for Institutions