

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

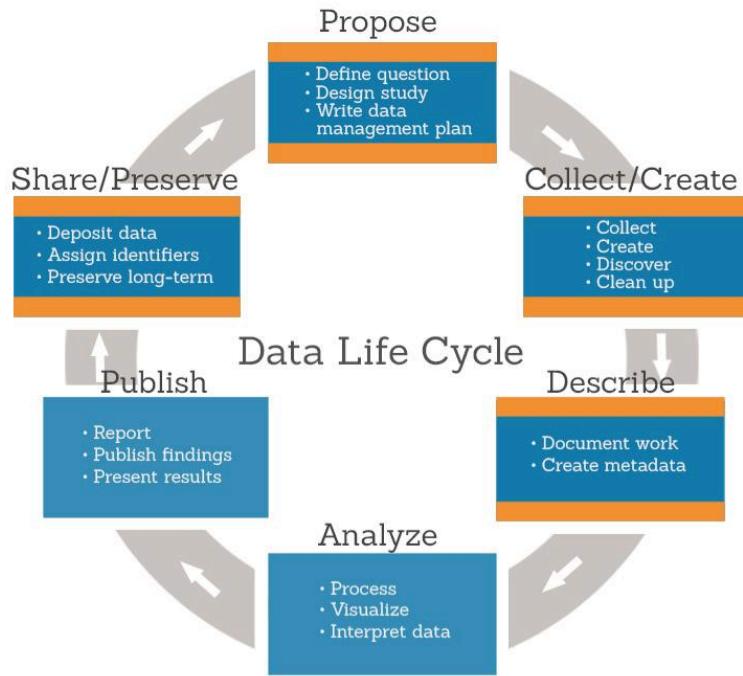
Part 2: Open Science Framework Demo



Part 1a

Why Manage Data?

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

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

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.

Best Practices

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

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

TIER Protocol

Developed by Haverford College –

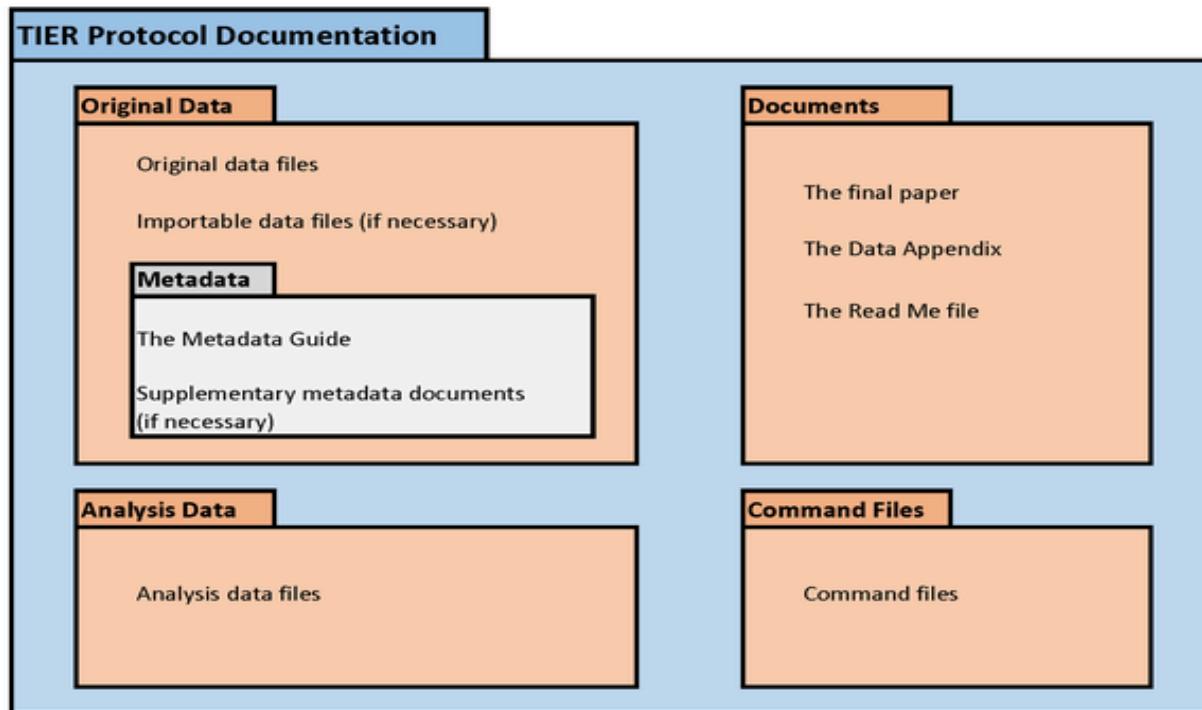
The **Teaching Integrity in Empirical Research** or TIER protocol, is a recommended protocol for comprehensively documenting all the steps of data management and analysis that go into an empirical research paper.

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

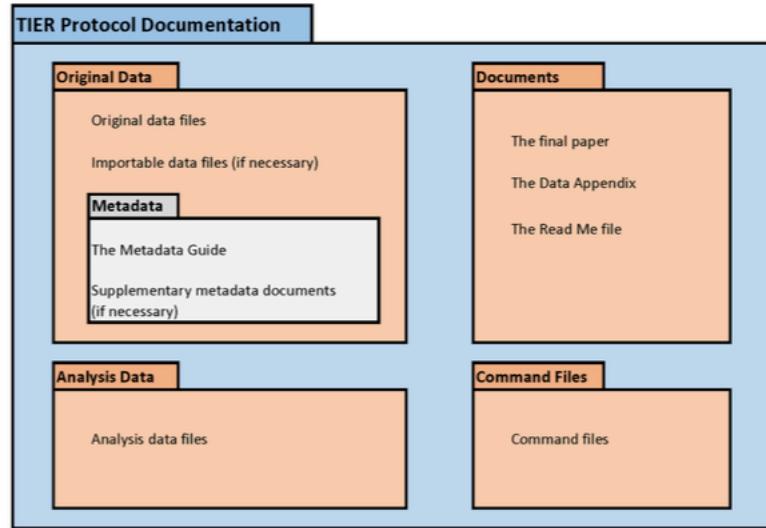
This file structure keeps your data organized and offers easy replication of results reported in a paper.

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

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



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.

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!

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.**

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!

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:

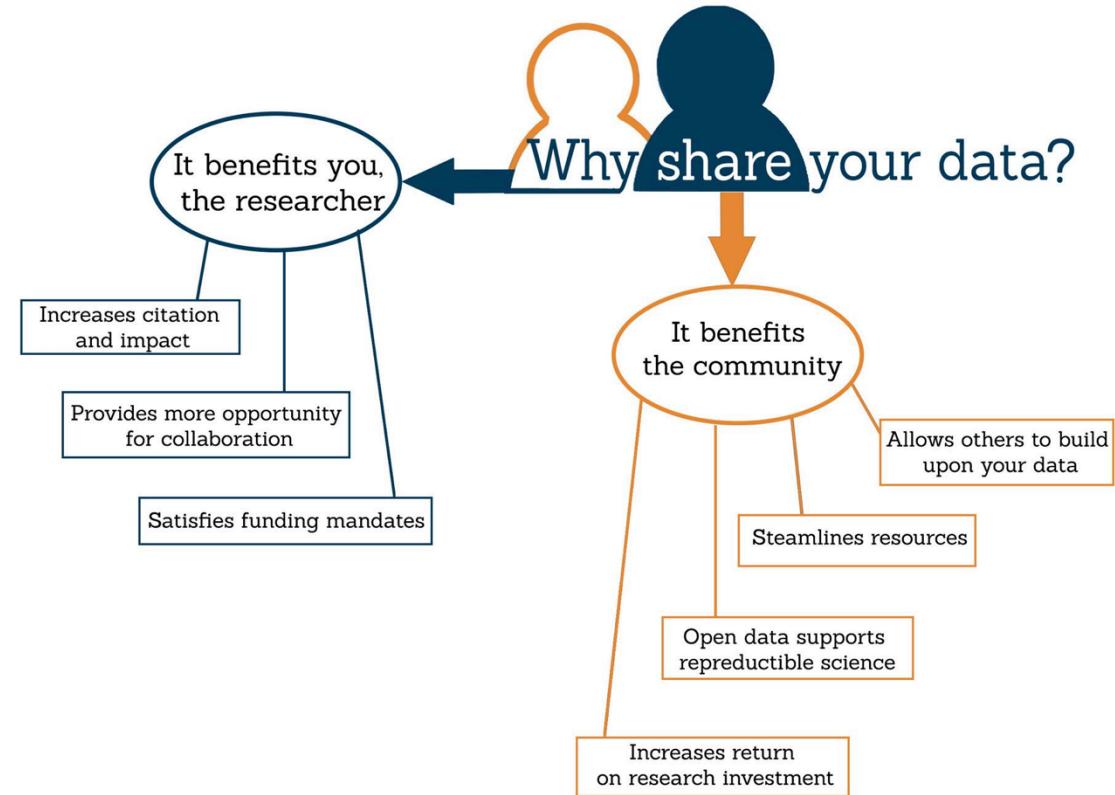
Save tabular data in a delimited text format.

Save data in uncompressed and unencrypted formats, where possible.

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.



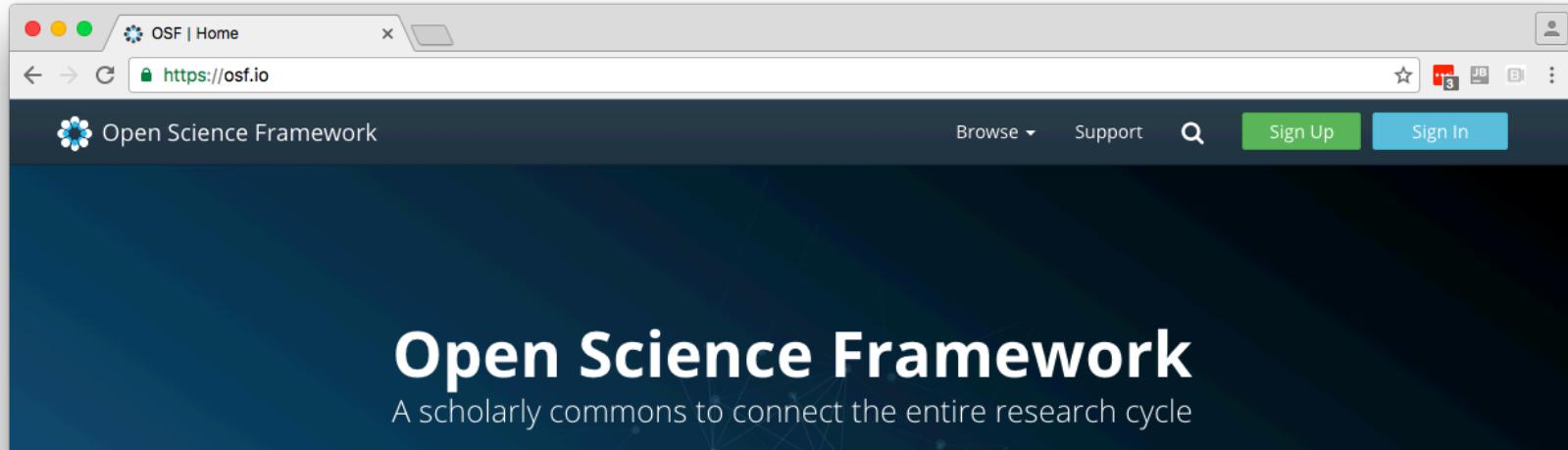
Part 2: Center for Open Science



Open Science Framework

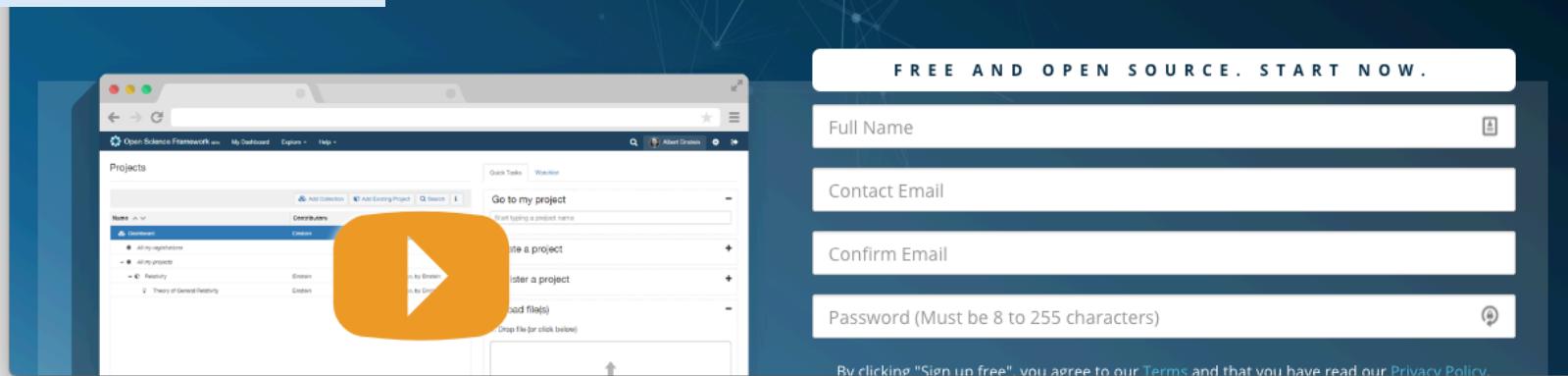
<http://cos.io/> | <http://osf.io>

Open Science Framework



The screenshot shows the OSF homepage in a web browser. The URL https://osf.io is visible in the address bar. The page features a dark blue header with the OSF logo, navigation links for 'Browse', 'Support', 'Sign Up', and 'Sign In'. Below the header, the main title 'Open Science Framework' is displayed in large white text, followed by the subtitle 'A scholarly commons to connect the entire research cycle'. A network graph graphic is visible in the background.

<http://osf.io>
FREE!



The screenshot shows the sign-up process on the Open Science Framework website. It includes a screenshot of the OSF dashboard with a play button overlay, a sign-up form with fields for 'Full Name', 'Contact Email', 'Confirm Email', and 'Password (Must be 8 to 255 characters)', and a legal notice at the bottom.

FREE AND OPEN SOURCE. START NOW.

Full Name

Contact Email

Confirm Email

Password (Must be 8 to 255 characters)

By clicking "Sign up free", you agree to our [Terms](#) and that you have read our [Privacy Policy](#).

Replication Studies ▼

Public



Study 3:

Contributors: Tim Errington

Date Created: 2013-1

Category: Project

Wiki

This project contains documentation from this paper. It includes clarifications. We also include notes from the Science Editors and authors that we have received as studies begin all data analysis...

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

Collaboration Documentation Archiving

osf.io/4bokd ▼

+

Put data, materials, and code on the OSF

Lepadogaster.stl

Share Download View Revisions

Search ^

Component: Lepadogaster lep...

- OSF Storage

Lepadogaster.stl

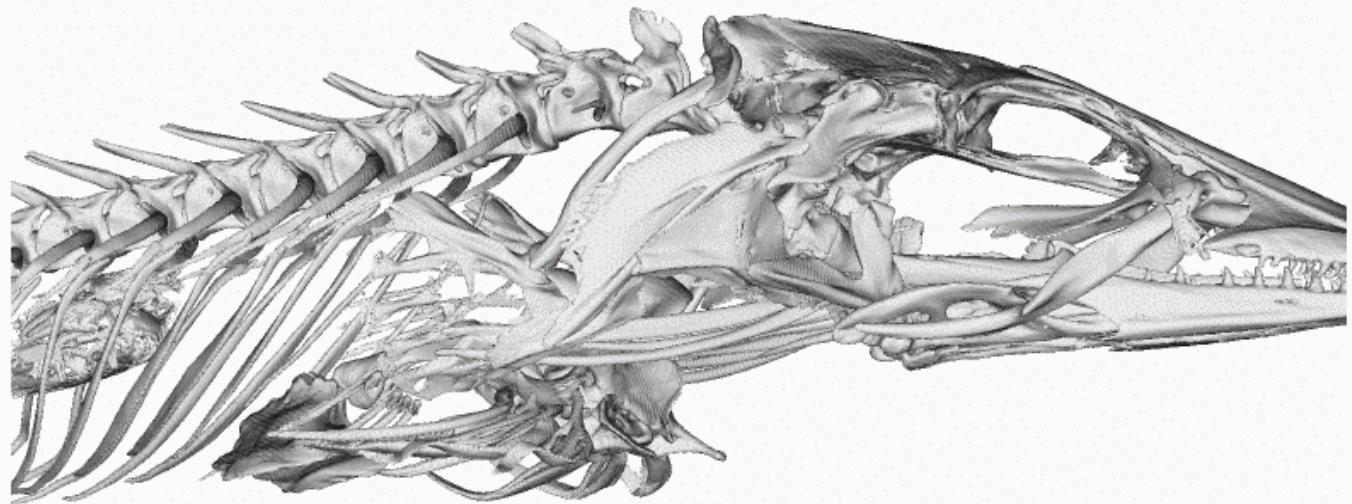
+ low res slice data

oblique.jpg

+ slice data

standard headshot.jpg

ventral.jpg





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...

Bowman SSP 2015 05 29

load

MD5



66518



5341f



d6d9e



122fb

2

2015-08-17 12:32 PM

Sara Bowman

0

1

2015-08-17 12:25 PM

Sara Bowman

0

[Presentations](#)[Files](#)[Wiki](#)[Analytics](#)[Registrations](#)[Forks](#)[Contributors](#)[Settings](#)

Bowman.ACS.2015.08.17.pptx

[Delete](#)

		Filter	^		
Component: Presentations		Revisions			
		Version ID	Date	User	Download
-	OSF Storage	4	2015-08-17 01:05 PM	Sara Bowman	14 66518
	2015.10.GHC.general.share...	3	2015-08-17 12:49 PM	Sara Bowman	0 5341f
	20150107_cendi_spies.pptx	2	2015-08-17 12:32 PM	Sara Bowman	0 d6d9e
	20160128_uva_dev_psych...	1	2015-08-17 12:25 PM	Sara Bowman	0 122fb
	Bowman.ACS.2015.08.17....				
	Bowman.LJAF.2015.04.22...				
	Bowman.Ruttenberg.Charl...				
	Bowman.SSP.2015.05.29				



recent and previous versions of file



<https://osf.io/wx7ck/>

Citation: osf.io/wx7ck

APA

Klein, R. A., Ratliff, K., et al. (2014). "Investigating Variation in Replicability: A "Many Labs" Replication Project." Open Science Framework (2014). osf.io/wx7ck

MLA

Klein, R. A., Ratliff, K., et al. (2014). "Investigating Variation in Replicability: A "Many Labs" Replication Project." Open Science Framework (2014). osf.io/wx7ck

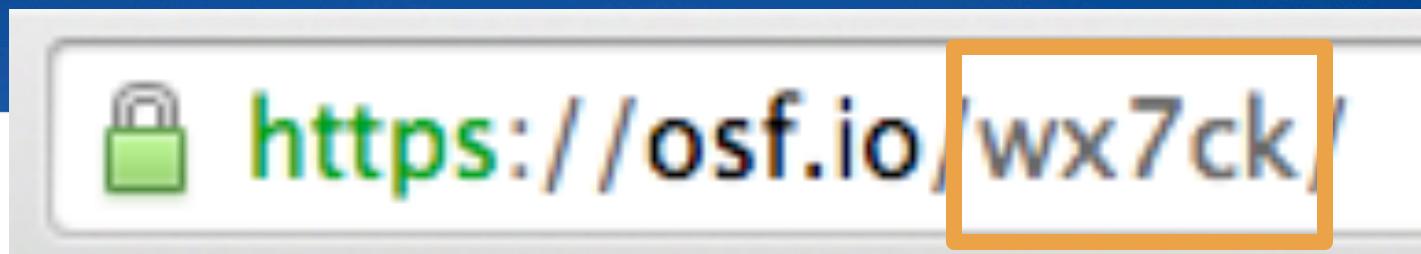
Chicago

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

Persistent Citable Identifiers

M. J., Bocian,
abs" Replication

M. J., Bocian,
eplication



https://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/

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

used in a citation



<https://osf.io/wx7ck/>



<https://osf.io/c97pd/>

Register

Is data collection for this project underway or complete?

Registration

all of the project materials, but
|

content and files cannot be deleted
plete and comprehensive for what

Type register if you are sure you want to continue

Name

-  Component: Demo Add-Ons
-  GitHub: AndrewSallans/demofiles master d2e68a6246
 -  ExampleIPythonNotebook.ipynb

Connects Services Researchers Use

-  ExampleWordDocument.docx
-  Amazon Simple Storage Service: osfdemofiles
-  FigShare: demofiles:892
-  Dropbox: /demofiles
 -  ExampleImage.jpg
 -  ExampleImage.png
 -  ExamplePDF.pdf
 -  ExamplePython.py
 -  ExampleSPSS.sav



GitHub



NeuroVault

zenodo



figshare
credit for all your research

OJS
Open Journal Systems
Ju[ubiquity press
open scholarship

VIVO
connect + share + discover



The
Dataverse
Network®
Project



MENDELEY

DMPTool

GitLab



GitHub



ownCloud

Dropbox

Google Drive

OneDrive

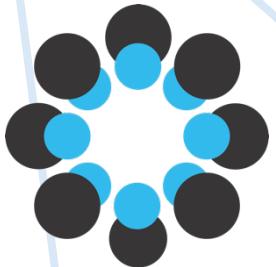
Amazon
web services™

box

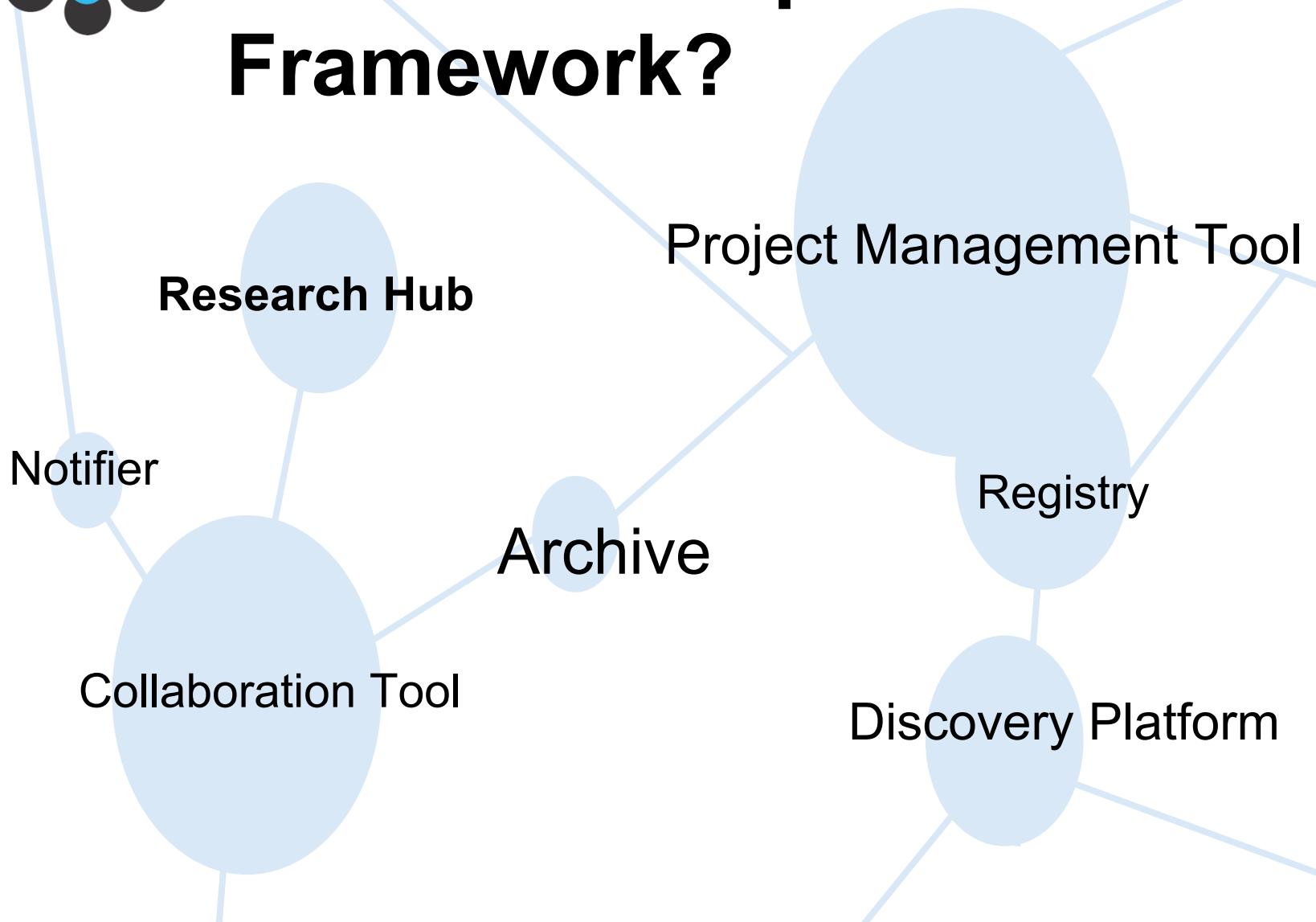
OpenSesame

PsychoPy
Psychology software in Python

Galaxy



What is the Open Science Framework?



Hands-On with the OSF

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



OSF for Institutions

OSF for Institutions

Open Science Framework Dashboard My Projects Browse Reid Otsuji

Custom branding → UC San Diego | The Library UC SAN DIEGO

This service is supported on campus by the UC San Diego Library for our research community. Do not use this service to store or transfer personally identifiable information, personal health information, or any other controlled unclassified information. For assistance please contact the Library's Research Data Curation Program at research-data-curation@ucsd.edu.

All Projects > Filter displayed projects

Collections

- All Projects
- All Registrations

Contributors

- YOUNGSUN KWON
- Tim Dennis
- joahnnes

Tags

- < 1/2 >
- multilateralism
- United Nations
- General Assembly
- Predictive model

↑ Sort by contributors or tags

↑ Research projects at UCSD

IRC0 467 - Team 26 : Predictive Model of UNGA voting behavior

Information Activity

Visibility : Public
Category: Project
Permission: Read
Last Modified on: 2017-02-22 10:31 AM

This project predicts the voting outcome by UNGA member-states using macroeconomic variables

Tags

- multilateralism
- United Nations
- General Assembly
- Predictive model
- Voting behavior
- Diplomacy

↑ Metadata about projects

OSF for Institutions

Open Science Framework Dashboard My Projects Browse Reid Otsuji

UC San Diego | The Library UC SAN DIEGO

This service is supported on campus by the UC San Diego Library for our research community. Do not use this service to store or transfer personally identifiable information, personal health information, or any other controlled unclassified information. For assistance please contact the Library's Research Data Curation Program at research-data-curation@ucsd.edu.

All Projects > Filter displayed projects

Collections	Name	Contributors	Modified
All Projects	IRCO 467 - Team 26 : Predictive... YOUNGSUN KWO... 9 hours ago		
All Registrations			

IRC0 467 - Team 26 : Predictive Model of UNGA voting behavior

Information Activity

Visibility : Public
Category: Project
Permission: Read
Last Modified on: 2017-02-22 10:31 AM

This project predicts the voting outcome by UNGA member-states using macroeconomic variables

Tags

multilateralism United Nations General Assembly Predictive model Voting behavior Diplomacy

YOUNGSUN KWON

Tim Dennis

joahnnes

Tags 1/2

multilateralism United Nations General Assembly Predictive model