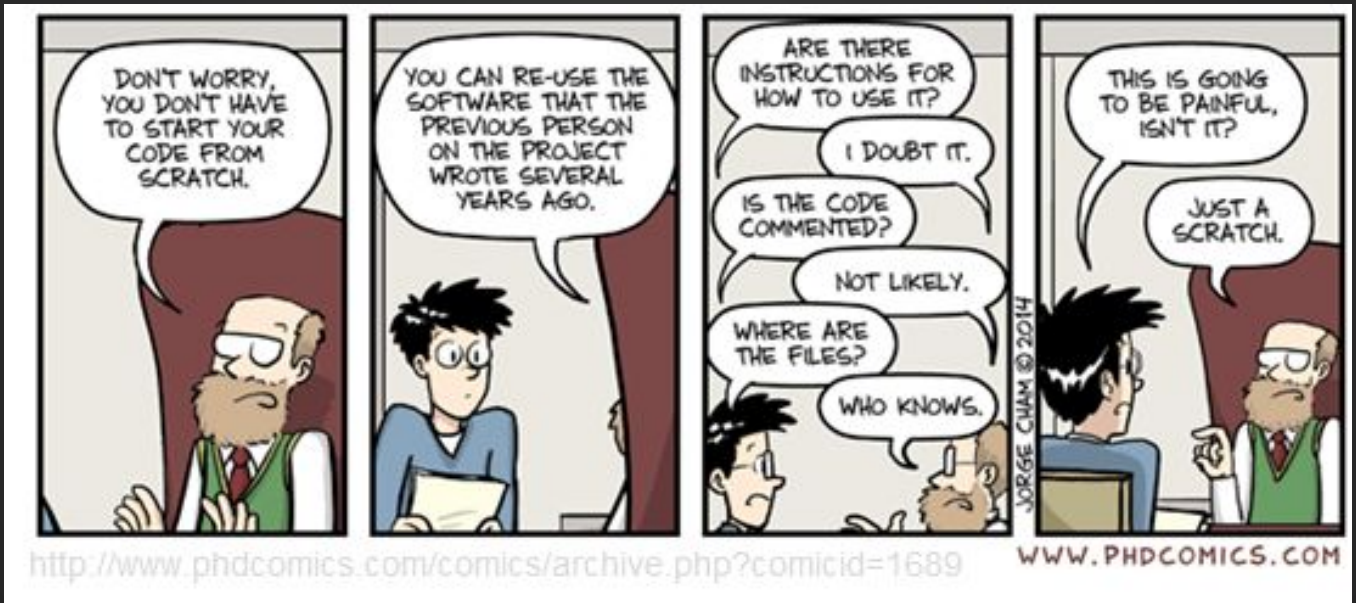
Building Tools and Services to Support Research Software Preservation and Sharing



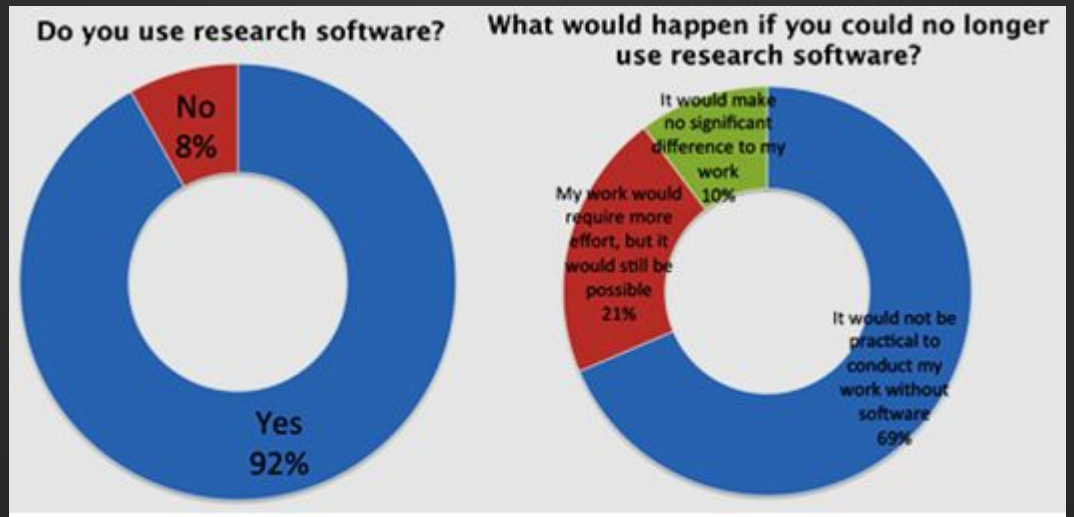
Software in the scholarly record

Nature's top 100 papers:

> "…the vast majority describe experimental methods or software that have become essential in their fields."

At least 5 out of the top 20 describe software

http://www.nature.com/news/the-top-100-papers-1.16224



https://www.software.ac.uk/blog/2014-12-04-its-impossible-conduct-research-without-software-say-7-out-10-uk-researchers

The problem…

CNI Spring 2016 plenary (Victoria Stodden):

> "The Scholarly Record comprises access to, and/or the ability to regenerate, items relied on in the generation of stated results."

> "At present, access to 'items' underlying computational results is limited."

"Access"

- Capture
- Preservation
- Policy
- Incentives
- Community support/advocacy

→ reproducibility, reuse

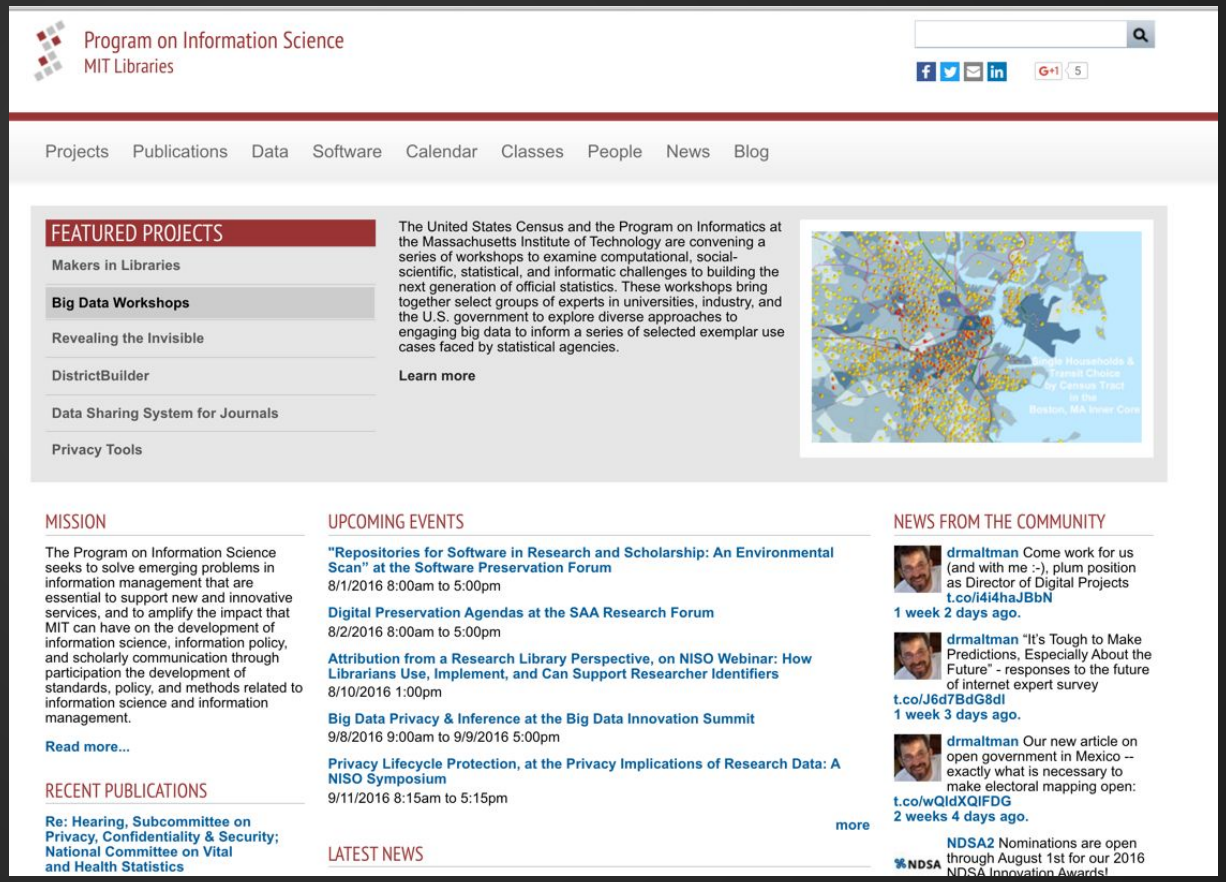
Micah Altman
Program on Information Science
MIT Libraries

Related Publications:

Informatics.mit.edu

Questions:

escience@mit.edu



DISCLAIMER

These opinions are my own, they are not the opinions of MIT, any of the project funders, nor (with the exception of co-authored previously published work) my collaborators

Secondary disclaimer:

"It's tough to make predictions, especially about the future!"

-- Attributed to Woody Allen, Yogi Berra, Niels Bohr, Vint Cerf, Winston Churchill, Confucius, Disreali [sic], Freeman Dyson, Cecil B. Demille, Albert Einstein, Enrico Fermi, Edgar R. Fiedler, Bob Fourer, Sam Goldwyn, Allan Lamport, Groucho Marx, Dan Quayle, George Bernard Shaw, Casey Stengel, Will Rogers, M. Taub, Mark Twain, Kerr L. White, etc.

Collaborators & Co-Conspirators

- Collaborators in research:
  Alex Chassanoff; Nancy McGovern
- Thanks to:
  Lucy Taylor – Graduate Research Intern
- Research Support
  - Supported in part by the Sloan Foundation; CLIR; Massachusetts Institute of Technology

Why Software?

ecma INTERNATIONAL

Standard ECMA-262
7th Edition / June 2016

ECMAScript® 2016 Language Specification

Internet Pioneer Warns Our Era Could Become The 'Digital Dark Ages'

PNAS

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund[a,b,c,1], Thomas E. Nichols[d,e], and Hans Knutsson[a,c]

[a]Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, S-581 85 Linköping, Sweden; [b]Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, S-581 83 Linköping, Sweden; [c]Center for Medical Image Science and Visualization, Linköping University, S-581 83 Linköping, Sweden; [d]Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom; and [e]WMG, University of Warwick, Coventry CV4 7AL, United Kingdom

Edited by Emery N. Brown, Massachusetts General Hospital, Boston, MA, and approved May 17, 2016 (received for review February 12, 2016)

What is Software

Working definition:
"Part of a computer system that consists of encoded information or computer instructions" (wikipedia) that is directly executable within a system.

Corollaries

- Software generally is composed of instantiations of algorithms, heuristics, and fixed information (internal data).
- The behavior and output of software generally depends on the execution context: execution environment (software, hardware, network, networked resources), configuration parameters, and dynamic inputs.

Software curation:

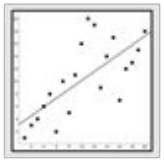
- encompasses/addresses the active practices necessary/related to the creation/acquisition, appraisal/selection, description, transformation, preservation, storage, and dissemination/access/reuse of software over short- and long- periods

Documents*

(compendia, fairy tales)

Assertions about

other entities



x	y
−2	−8
−1	−6.5
0	−5
1	−3.5
2	−2

''We applied a general linear model'
'We conjecture kids will choose candy'
'$\delta = 2.3 * \sqrt{\Omega}$'
'Chewing gum tastes great'

Logical Claims

Theorem 1
....
Lemma 1.1

Speculations, Commentary

Thanks to my dog for his support…

Publication Operations (manuscripts , proofreading , translation, ...)

References, Citation

(Altman, et al. 2013)

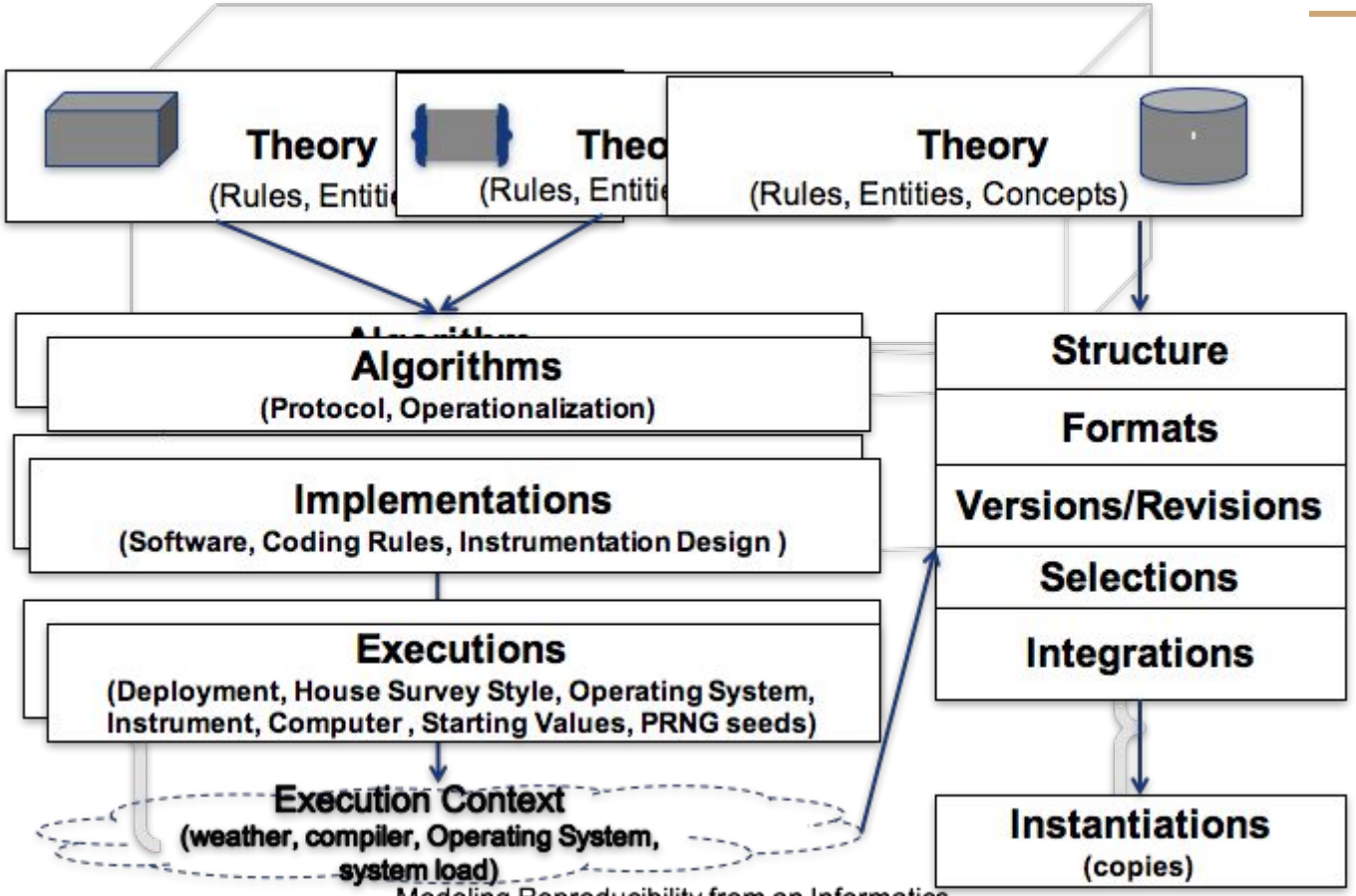
U49845.1 GI:1293613

doi:10.1002/0470841559.ch1
orcid:0000-0001-7382-6960

Internal Meta-Information

Title: XXXX

Modeling Methods, Analysis & Data…



Modeling Reproducibility from an Informatics Perspective

Some Caution About Definitions

"... if they [philosophers] do ask and they want a definition, they do not want the most natural definition, e.g. of 'chair' they do not want the definition 'something to sit on'. Why are they not satisfied with the normal definition of chair, or, to put the question in another way, why do they wish to ask for the definition of a physical object?"

Source: "From the Minutes of the Moral Science Club, 23.2.1939" in *Wittgenstein in Cambridge* (2008)

- Software is often tightly coupled to data
- Boundaries among software objects and systems are fuzzy & permeable
- Usefulness of software is strongly dependent on the intent of the user, knowledge and capabilities of user (documentation matters), and execution context.

Selected software curation use case areas and goals

Historic / cultural / social science	- historical/cultural/social science/organizational scholarship on software development - “” scholarship on use and experience of software by a community - “intrinsic value”/artistic value
Replication and reproducibility	- evaluate evidence claims made in research - reduced deliberate research fraud - check reliability (robustness) of results - check validity accuracy of described methods (accuracy)
Reuse	- increase speed of development - standards compliance - apply methodology to a different corpus - increased quality and dependability
Render other digital objects	- renders other objects meaningful - see digital preservation use cases
Legal	- record of licensing, ownership, copyright - manage legal risks/accountability - compliance with laws/funding mandates - reduce barriers to long-term access for other historic use, replication, reuse, rendering
Citation and attribution	- allocate credit/ evaluate individual academic career - analyze software development/history

Potential Research Questions

- How is software related to research formally disseminated?
- Which “repositories” (points for mid/long term publishing/access of software) are broadly recognized?
- What is the relative prevalence and affordances of “repositories” for software?
- What practices, requirements, or standards for software curation are recognized broadly?
- What models are broadly recognized for describing software, software development, and curation?
- How are researchers using software in their research activities? What are anticipated future use cases?
- How are libraries and archival institutions approaching curation of software?
- How do libraries and archives envision the role of software in their collections?
- What kinds of challenges do libraries and for archival institutions face in curating software?
- What curation practices can library/archival institutions embed into their stewardship activities?
- What do stakeholders (funding institutions, journals) value as metrics for good curation practices?

Preliminary Environmental Scan

Literature Review

- Data Curation, Publication and Citation
- Software research use cases
- Software repositories
- Software & scientific reproducibility
- Software Engineering Methodology

Web Research -- Repository Practice

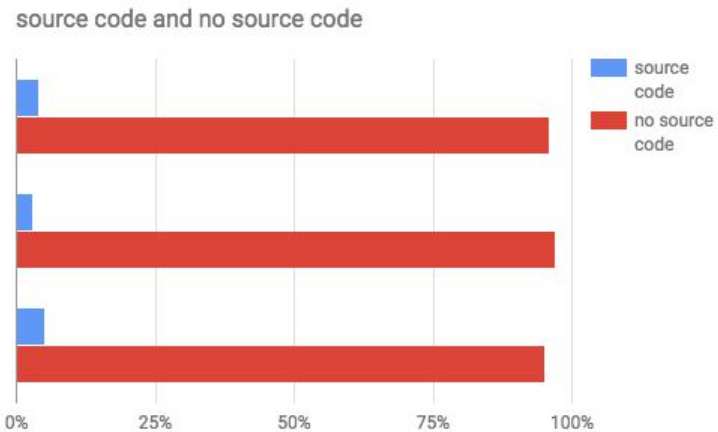
- Review of research repositories & software directories
  - *Sources*: OpenDOAR, Re3Data, Sherpajuliet
  - *Sources*: OpenHub, OSDir, DMOZ
  - *Goals*: Estimate prevalence of repositories that accept research software

Web Research -- Research Software Policies

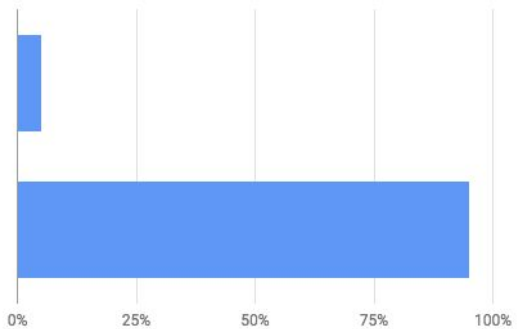
- Review of funder policies
  - Sources: Roarmap; US Federal Agency Websites
- Review of Journal Policies
  - Sources: Google Scholar, WoS, DOAJ, Software Sustainability Institute Index

Prevalence of software in research repositories appears low...

RE3



SherpaJuliet



Most of what is in research repositories appears limited.

Now showing items 1-10 of 29433 Next Page
1 2 3 4 . . . 2944

Communities or Collections matching your query

Software

Items matching your query

Software files to support thesis of Ty Thomson
Thomson, Ty (2008-05-21)

An Object-Oriented Software Reuse Tool
Monegan, Michael D. (1989-04-01)
The Object-oriented Reuse Tool (ORT) supports the reuse of object-oriented software by maintaining a library of reusable classes and recording information about their reusability as well as information associated with their ...

Intellectual Property and Software: The Assumptions are Broken
Davis; Randall (1991-11-01)
In March 1991 the World Intellectual Property Organization held an international symposium attended primarily by lawyers, to discuss the questions that artificial intelligence poses for intellectual property law (i.e., ...

Role of Software Readability on Software Development Cost
Collar, Emilio, Jr.; Valerdi, Ricardo (2006-11-07)
This paper explores the role of software readability on software development cost. We argue that the up front cost of incorporating software readability pays off handsomely at later stages in the life cycle, especially ...

Analysis of Japanese Software Business
Inada, Kenichiro (Massachusetts Institute of Technology, 2010)
Today, our society is surrounded by information system, computers, and software. It is no exaggeration to say that our daily life depends on software and its function. Accordingly, the business of software has made miraculous ...

Software for modeling set perception
Rosenholtz, Ruth; Alvarez, George (2007-10-30)

Computer Aided Evolutionary Design for Software Engineering
Rich, Charles; Shrobe, Howard E.; Waters, Richard C. (1979-01-01)
We report on a partially implemented interactive computer aided design tool for software engineering. A distinguishing characteristic of our project is its concern for the evolutionary character of software

Software files to support thesis of Ty Thomson

Download

Author: Thomson, Ty

Citable URI: http://hdl.handle.net/1721.1/41841
Date Issued: 2008-05-21
URI: http://hdl.handle.net/1721.1/41841

Show full item metadata

Files in this item

Name	Size	Format	Description
Thesis files.zip	56.23Kb	Unknown	

Preliminary findings: State of Software Curation

"Nothing Exists" - Parmenides (ca. 500 BCE)

1. No comprehensive indices of software archives
2. Orders of magnitude fewer software archives than data archives.
   (Corollary: most repositories containing software offer little functionality for discovering, understanding, or reusing it)
3. Very small proportion of funders have policies addressing software curation
4. There is little available advice for those who wish to curate, cite, & preserve software

Contrast with Industry -- Early Impressions

- Enterprise Software Management
  - Mature tools and practices for enterprise license management, deployment, inventory, updating
  - Models & documentation are proprietary, implicit, unstandardized
- Enterprise Software Development
  - Mature tools and practices for revision control; testing; QA
  - Standardized modeling language for modeling software design (e.g. UML) -- with broad tool support for large efforts
  - Less support for model-driven development, testing, deployment
  - Tools and practices focused on product development, QA, and internal reuse -- not on other use cases important to curation and research

Contrast with Data Curation -- Lack of Progress

- Compliance
  - Funder: data management plans, open data
  - Publishers: data access/archiving/citation
- Norms & practices
  - Joint data citation principles
  - Recognition of data in funder biosketches
  - Increased recognition of reproducibility gaps
  - Increased recognition of open data/open science
- Technical infrastructure
  - Data repository directories
  - Data citation indices
  - ORCID researcher identifier and registry
- Recognition
  - Data citation indices
  - Virtual branded archives
  - High-profile data publications

DC1 Data Citation Principles

The Dataverse Project

figshare

DataCite

SCIENTIFIC DATA

Some Exemplars and Promising Initiatives

- Citation and publisher policies
  - FORCE 11 Software Citation Principles www.force11.org/software-citation-principles
  - ACM New Publication Policies on Software Reproducibility and Contributorship www.acm.org/publications/policies
  - PLOS journals.plos.org/plosone/s/materials-and-software-sharing
- Practices for reusable and reproducible software
  - Software carpentry: software-carpentry.org
  - Open software foundation: osf.io
- Long Term Access efforts:

  - www.softwareheritage.org
  - www.softwarepreservationnetwork.org
  - Github/Zenodo: guides.github.com/activities/citable-code
  - Internet Archive: archive.org/details/softwarelibrary
- Formal Software Publication:

  - Software X: www.journals.elsevier.com/softwarex
  - Journal of Statistical Software: www.jstatsoft.org/
  - Open Research Software: openresearchsoftware.metajnl.com

Takeaways...

- Software boundaries and definitions are often contested -- critical for curation efforts to identify targeted use cases
- Some mature practices and tools for software development, maintenance, and deployment exist in industry
- Practices are rapidly emerging in software citation
- There are number of early-stage initiatives in curation and preservations worth close attention
- Software curation for research, history and culture is at early stages -- trails data curation by a decade or more
- Curation efforts should recognize the practices and tools above, although they cover only a portion of the target use cases and lifecycle

Jeff Spies

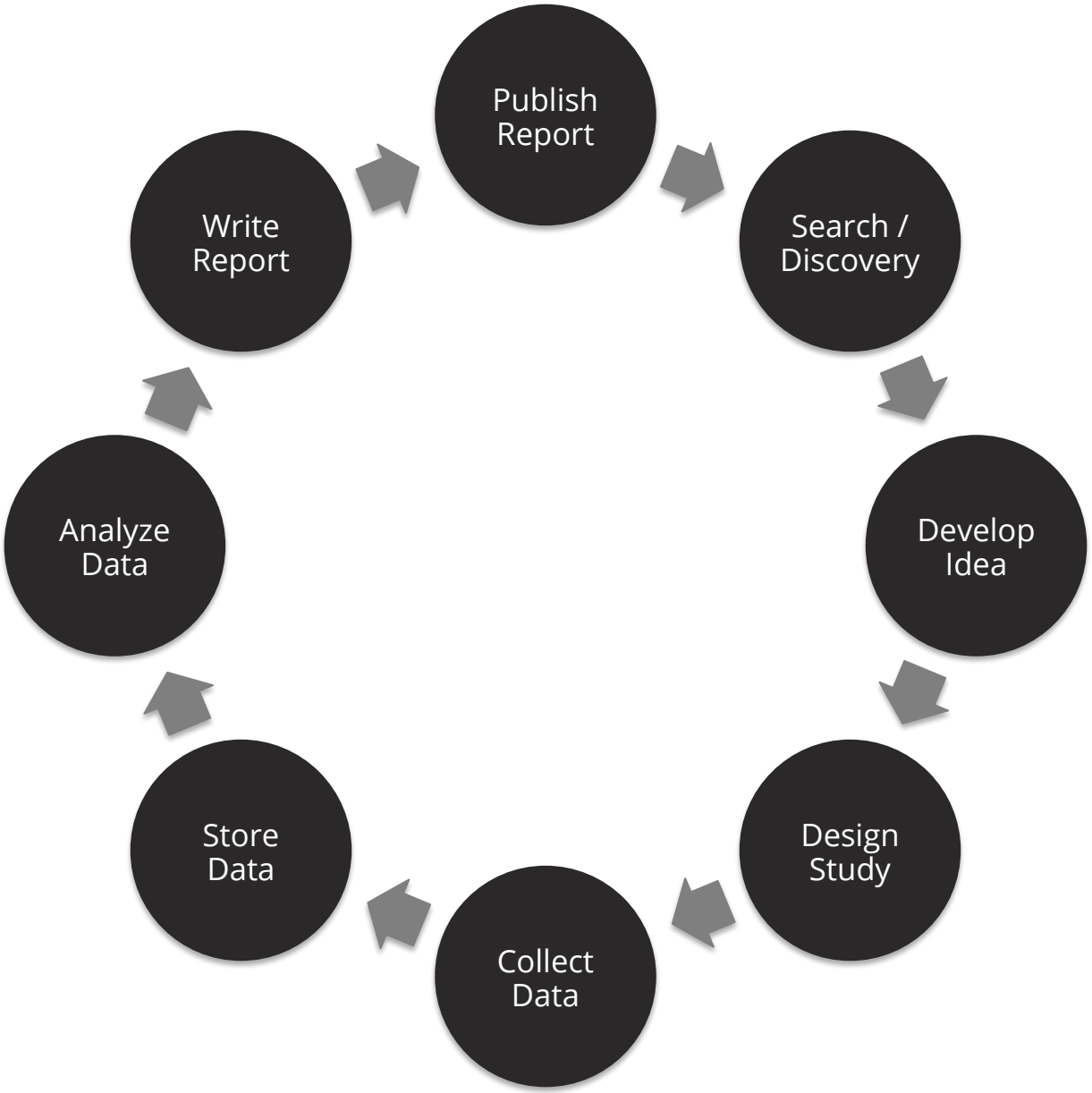
- COS - Co-founder, Chief Technology Officer
- SHARE - Co-lead
- UVA - Asst. Prof., Dept. of Engineering & Society

jeff@cos.io

@jeffspies

Incentives for individual success are focused on **getting it published**, **not getting it right**.

Nosek, Spies, & Motyl, 2012


Publish Report
Search / Discovery
Develop Idea
Design Study
Collect Data
Store Data
Analyze Data
Write Report

Context is important

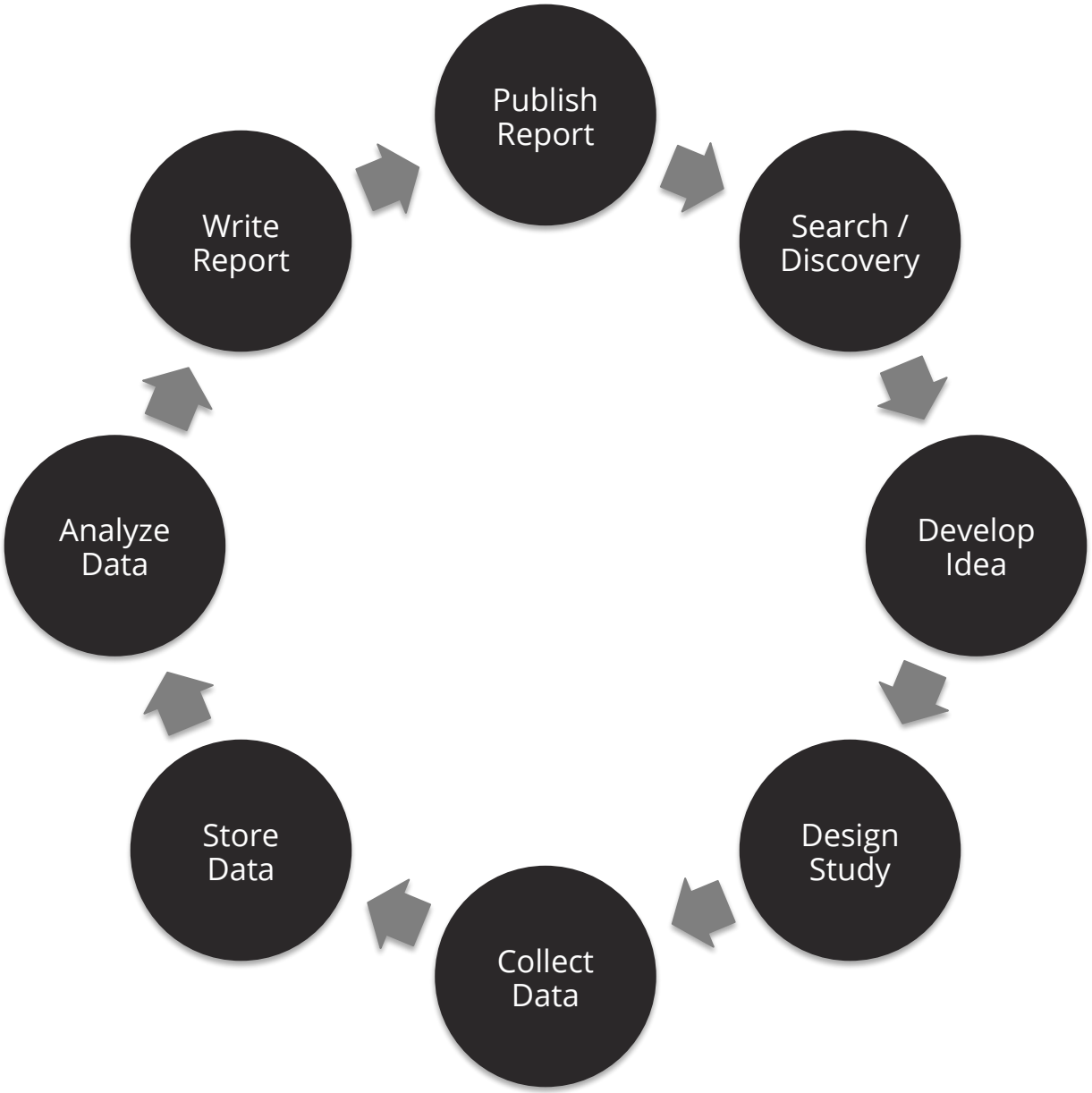
- The problems we deal with are challenging
- We are human
  - Biased
  - Make decisions and interpret results based upon unique experiences
  - Influenced by incentives
  - Forgetful
  - Busy

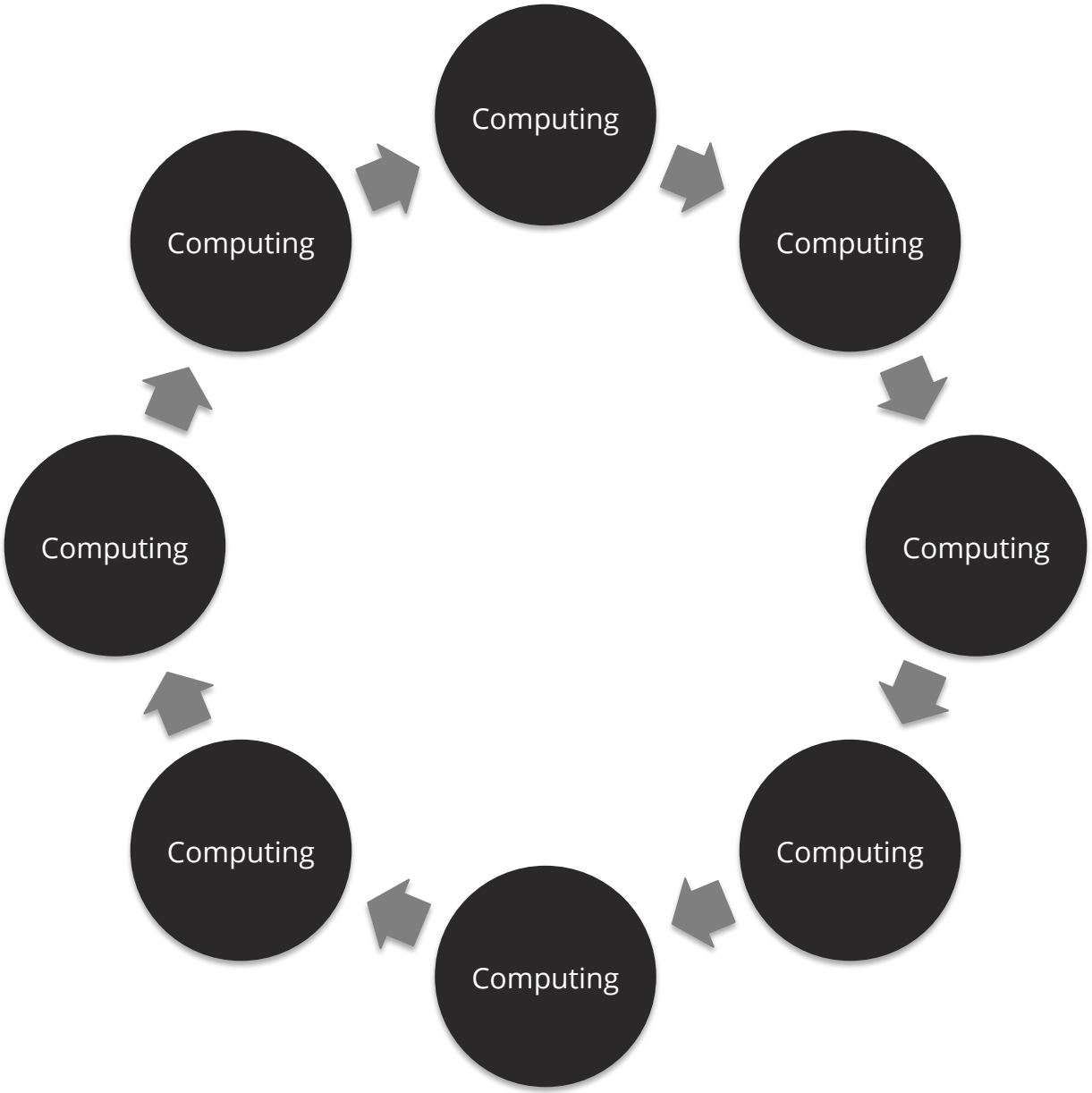
We need context for . . .

- Reproducibility
- Replicability
- Extensibility
- Meta-scholarship

in order to increase research . . .

- Efficiency
- Quality
- Accessibility
- Diversity


Publish Report
Search / Discovery
Develop Idea
Design Study
Collect Data
Store Data
Analyze Data
Write Report


Computing
Computing
Computing
Computing
Computing
Computing
Computing
Computing

Computing is pervasive, and we must capture it as context of the scholarly workflow.

But this isn’t easy. Luckily, we have a partner:

DevOps

DevOps [emphasizes] **collaboration** and **communication**...while **automating the process of software delivery and infrastructure changes**...[such that] building, testing, and releasing software can happen **rapidly**, **frequently**, and **more reliably**.

Wikipedia

We need context for . . .

- Reproducibility
- Replicability
- Extensibility
- Meta-scholarship

in order to increase research . . .

- Efficiency
- Quality
- Accessibility
- Diversity

DevOps needs context for **software** . . .

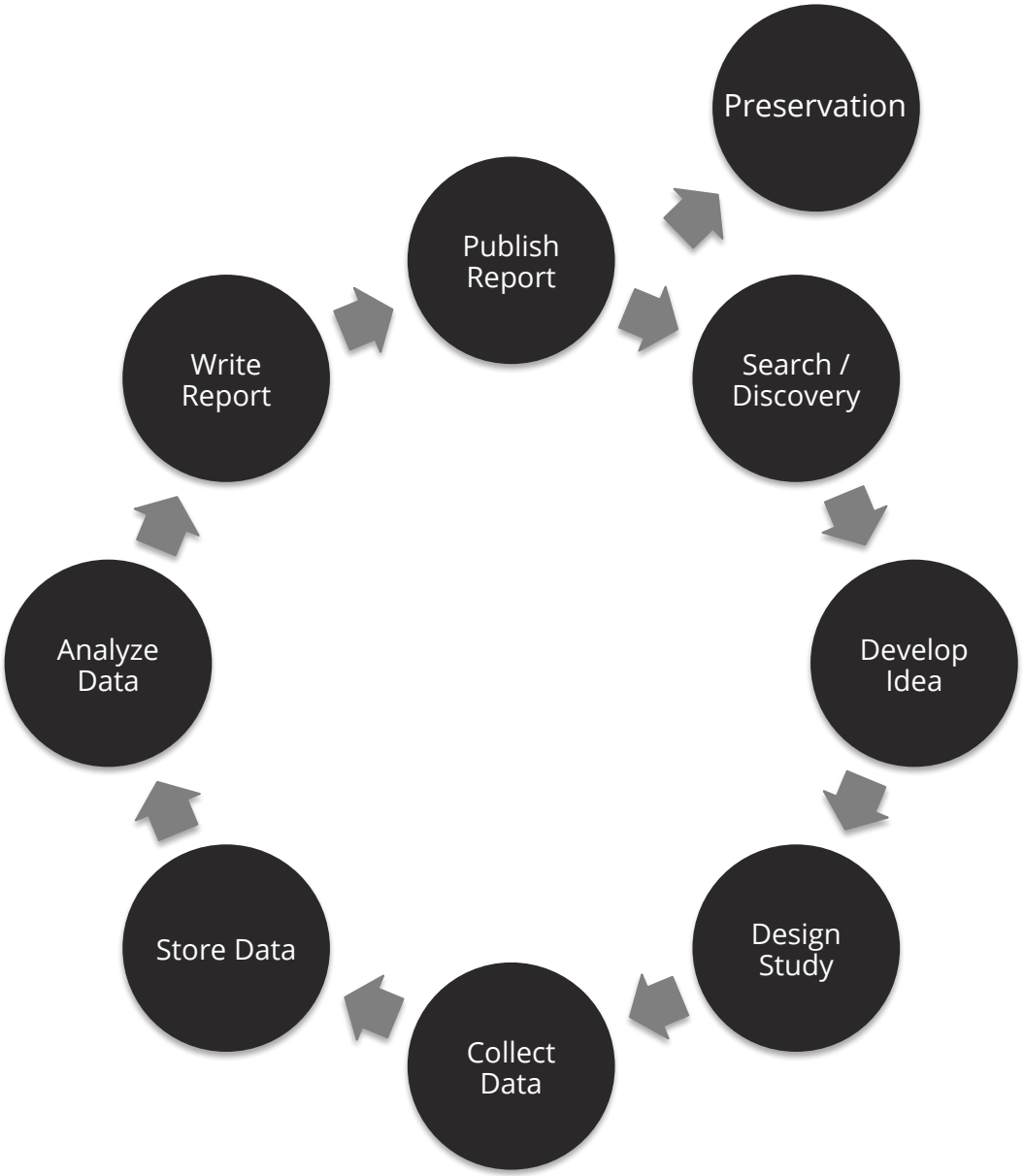
- Reproducibility
- Replicability
- Extensibility
- **Management**

in order to increase **software** . . .

- Efficiency
- Quality
- Accessibility
- Diversity

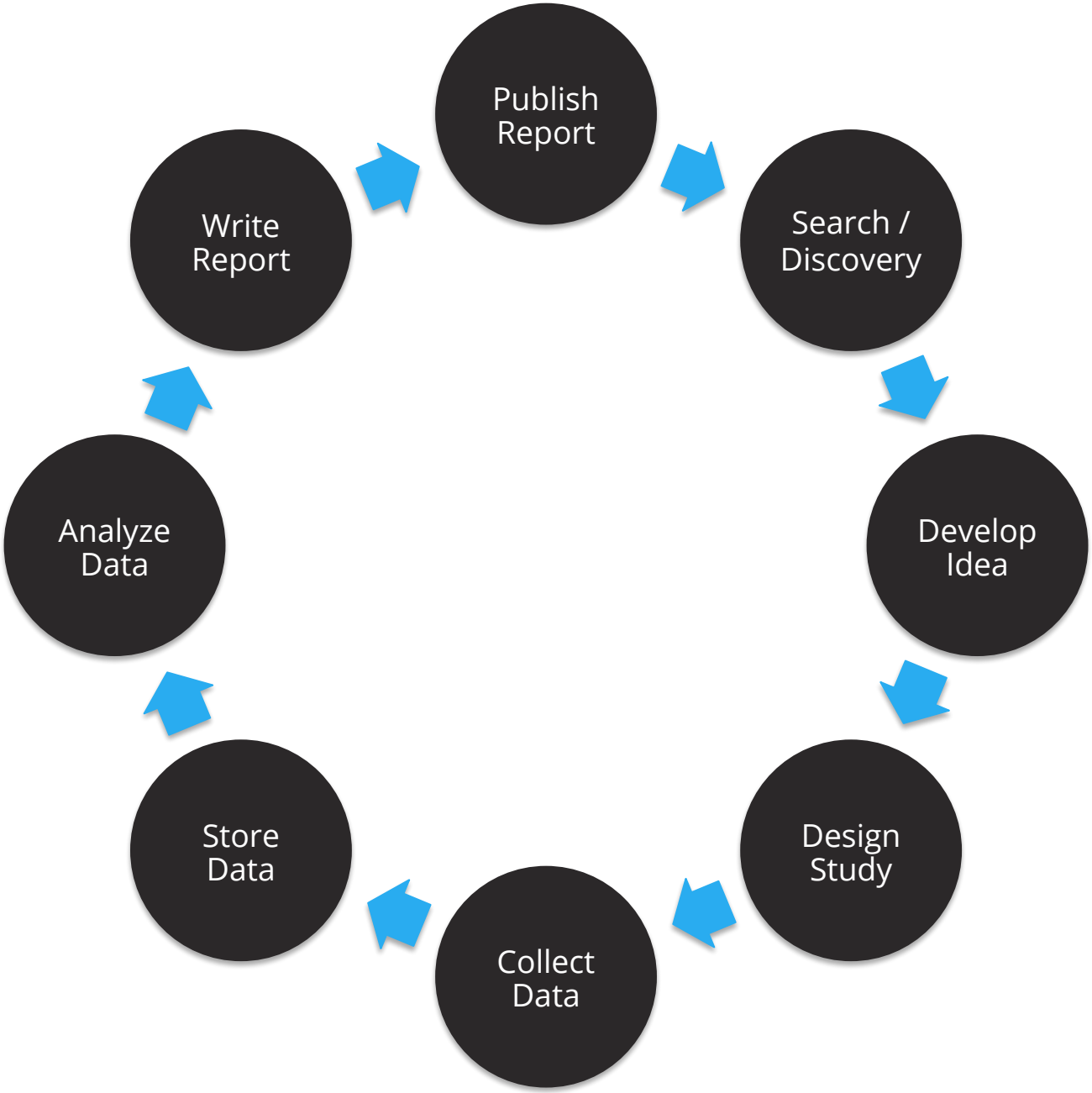
DevOps + Librarians =

Research Software
Preservation and Sharing

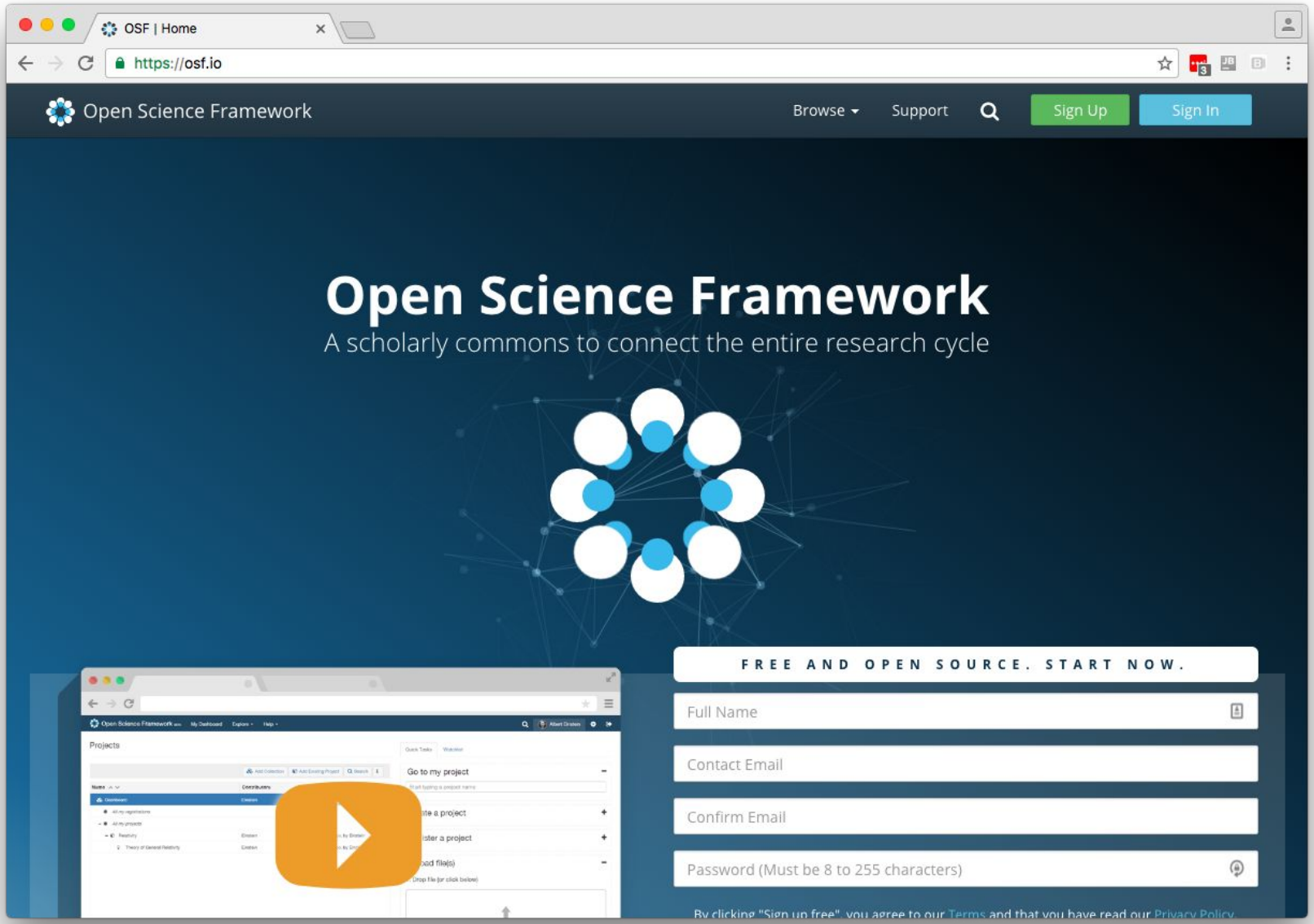

Preservation
Search / Discovery
Develop Idea
Design Study
Collect Data
Store Data
Analyze Data
Write Report
Publish Report

OSF

OSF


Publish Report
Search / Discovery
Develop Idea
Design Study
Collect Data
Store Data
Analyze Data
Write Report

http://osf.io


OSF | Home
https://osf.io
Open Science Framework
Browse
Support
Sign Up
Sign In
Open Science Framework
A scholarly commons to connect the entire research cycle
FREE AND OPEN SOURCE. START NOW.
Full Name
Contact Email
Confirm Email
Password (Must be 8 to 255 characters)

Collaboration
Documentation
Archiving

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: Gupta et al. 2010, Nature

Contributors: Tim Errington, Elizabeth Iorns, William Gunn, Fraser Elisabeth Tan, Sarah Statt, Joelle Lomax, Nicole Perfito

Date Created: 2013-10-22 02:04 PM | Last Updated: 2015-01-20 06:16 PM

Category: Project

Wiki

This project contains all information pertaining to the replication of key experiments from this paper. It includes the detailed protocols, including reagents and author clarifications. We also include any comments from other contributors, researchers from the Science Exchange network, and further information with the original authors that we have learned since the beginning of the project. When experimental studies begin all data collected will also be deposited here, including data analysis...

Files

Search

Name

Project: Study 3: Gupta et al. 2010, Nature

Citation

osf.io/4bokd

Components

Coded Paper

Errington, Iorns, Gunn & 2 more

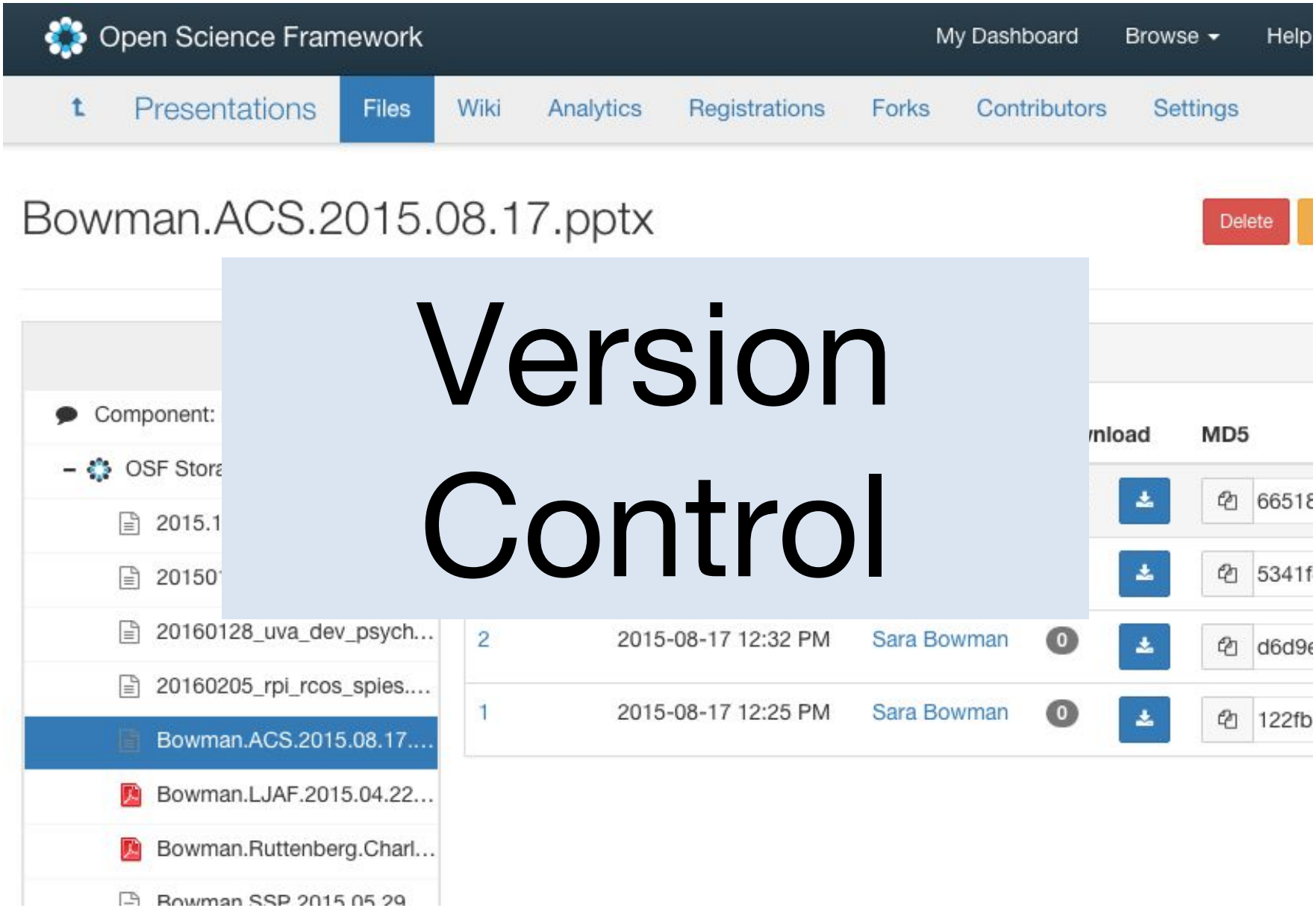
8 contributions

Recent Activity

All times displayed at -0700 UTC offset.

2015-01-20 06:16 PM Tim Errington added Nicole Perfito as contributor(s) to

Version Control



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...
- Bowman.SSP.2015.05.29

Revisions

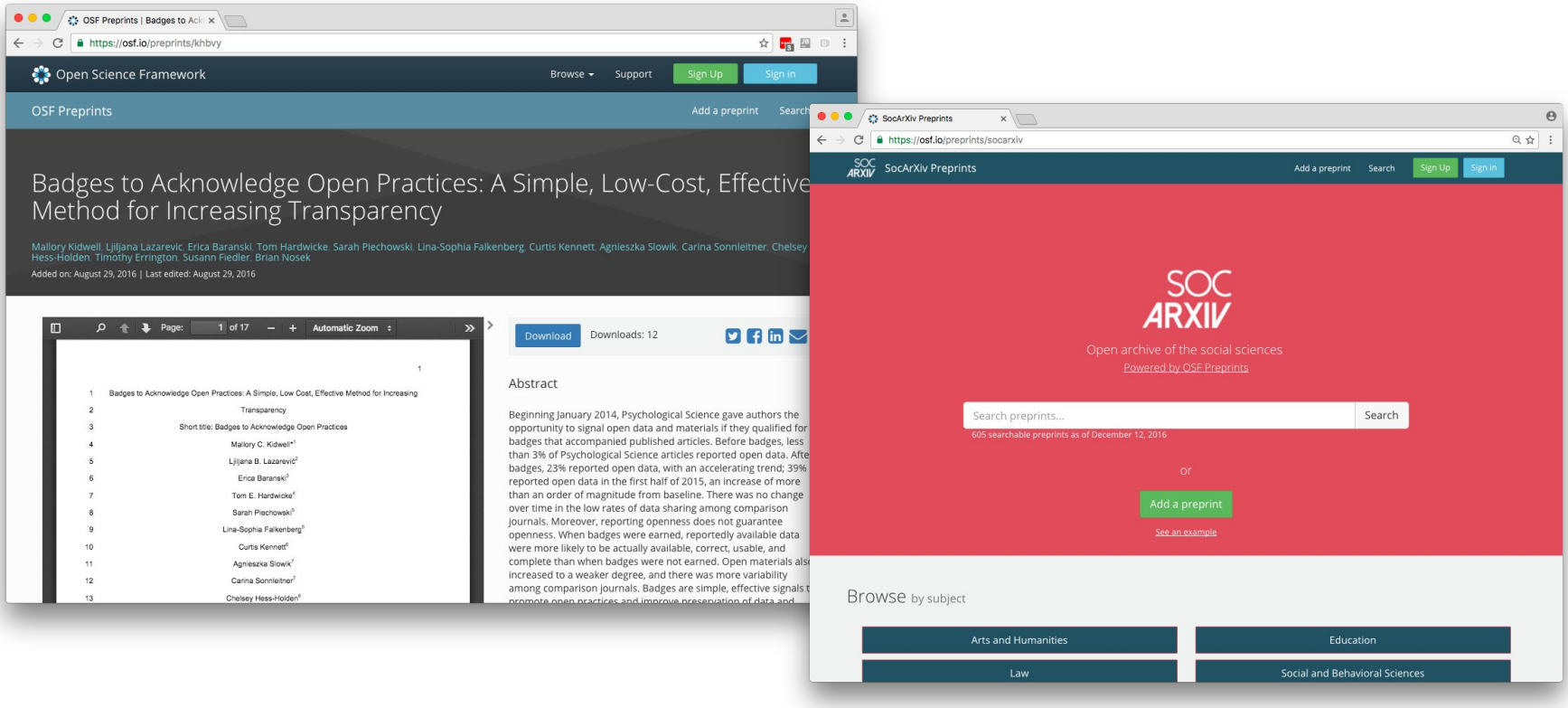
Version ID	Date	User	Download	MD5
4	2015-08-17 01:05 PM	Sara Bowman	14	66518
3	2015-08-17 12:49 PM	Sara Bowman	0	5341f
2	2015-08-17 12:32 PM	Sara Bowman	0	d6d9e
1	2015-08-17 12:25 PM	Sara Bowman	0	122fb

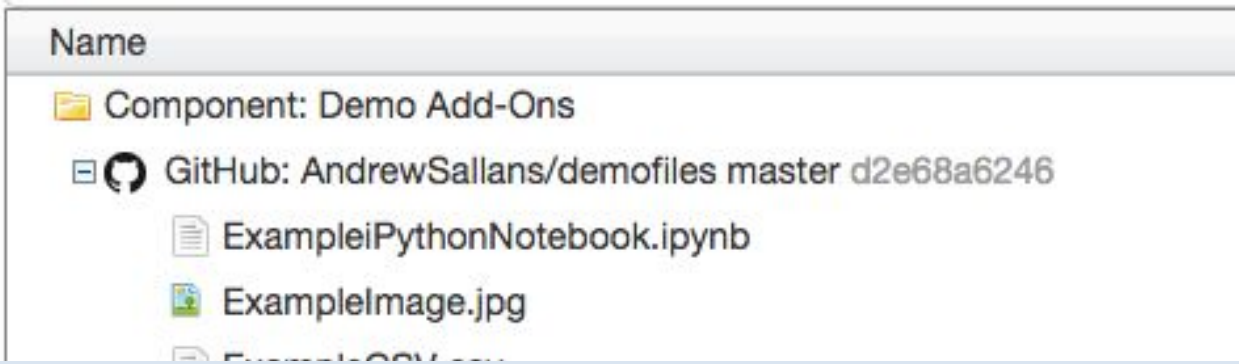
Other Features

- Granular privacy/sharing
- Granular permissions
- Analytics dashboards
- Persistent, citable identifiers
- Persistent content
- Project snapshotting (i.e., registration)
- Licensing
- Forking

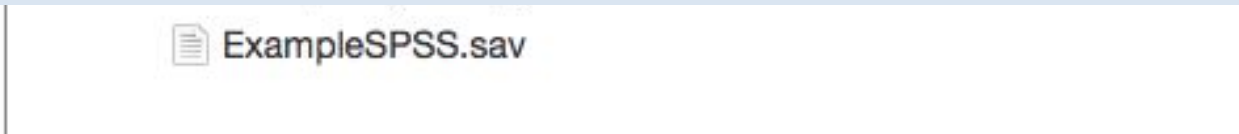
And we build other services with the OSF API (+ SHARE) like

http://osf.io/preprints





Connects Services Researchers Use



Name

- Component: Demo Add-Ons
  - GitHub: AndrewSallans/demofiles master d2e68a6246
    - ExampleiPythonNotebook.ipynb
    - ExampleImage.jpg
    - ExampleCSV.csv
    - ExampleImage.png
    - ExampleSPSS.sav
    - ExamplePython.py
    - ExampleSpreadsheet.xlsx
    - ExamplePDF.pdf
    - ExampleR.r
    - ExampleWordDocument.docx
  - Amazon Simple Storage Service: osfdemofiles
  - FigShare: demofiles:892
  - Dropbox: /demofiles
    - ExampleImage.jpg
    - ExampleImage.png
    - ExamplePDF.pdf
    - ExamplePython.py
    - ExampleSPSS.sav

Open Science Framework

My Dashboard Browse Help

Lepadogaster lepadogaster | Files | Wiki | Analytics | Registrations | Forks

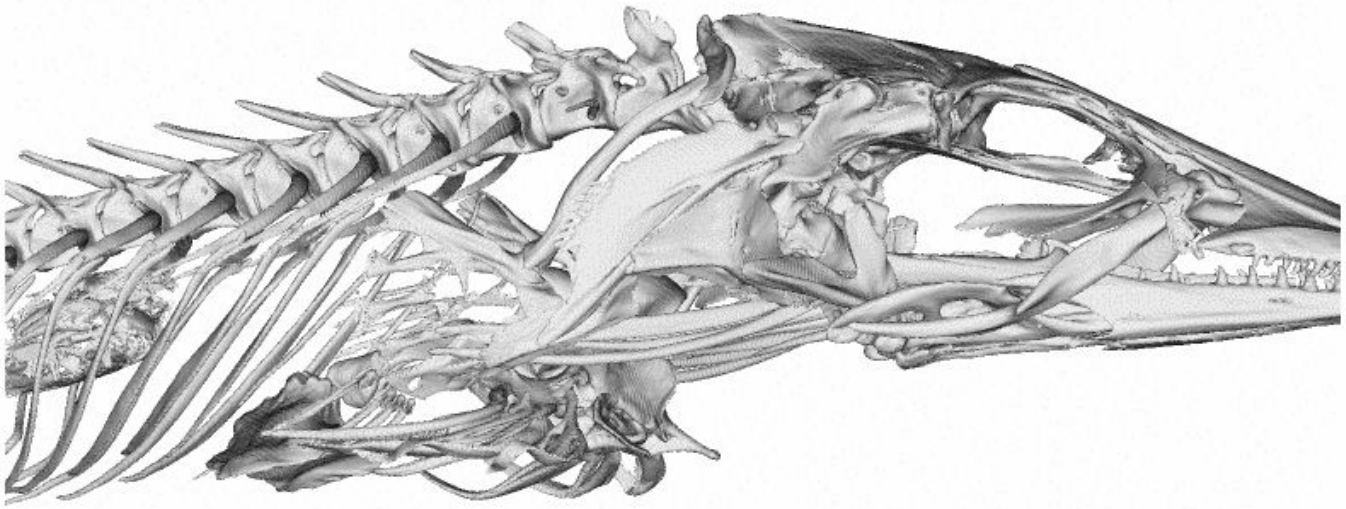
Lepadogaster.stl

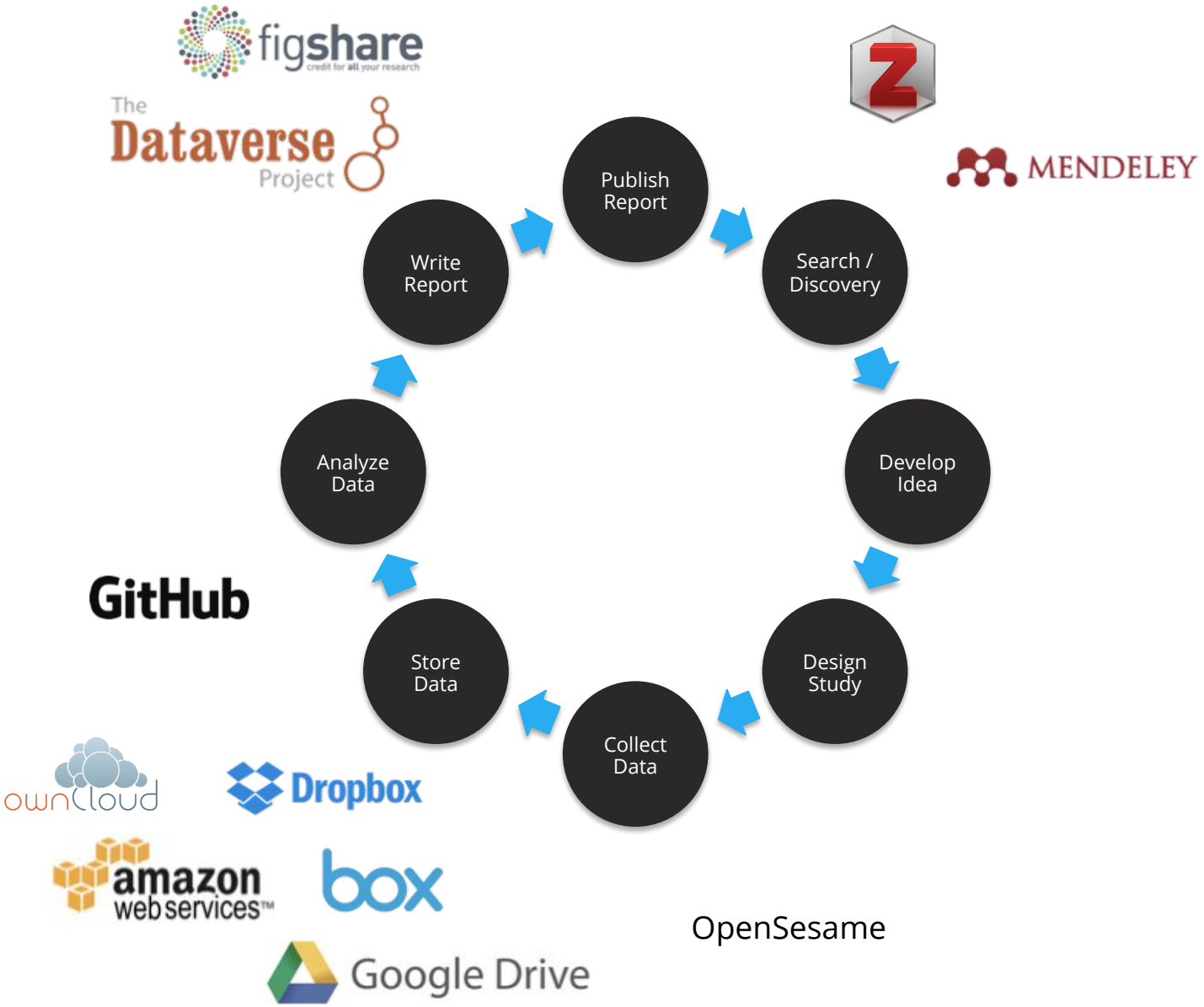
Share Download View Revisions

Search

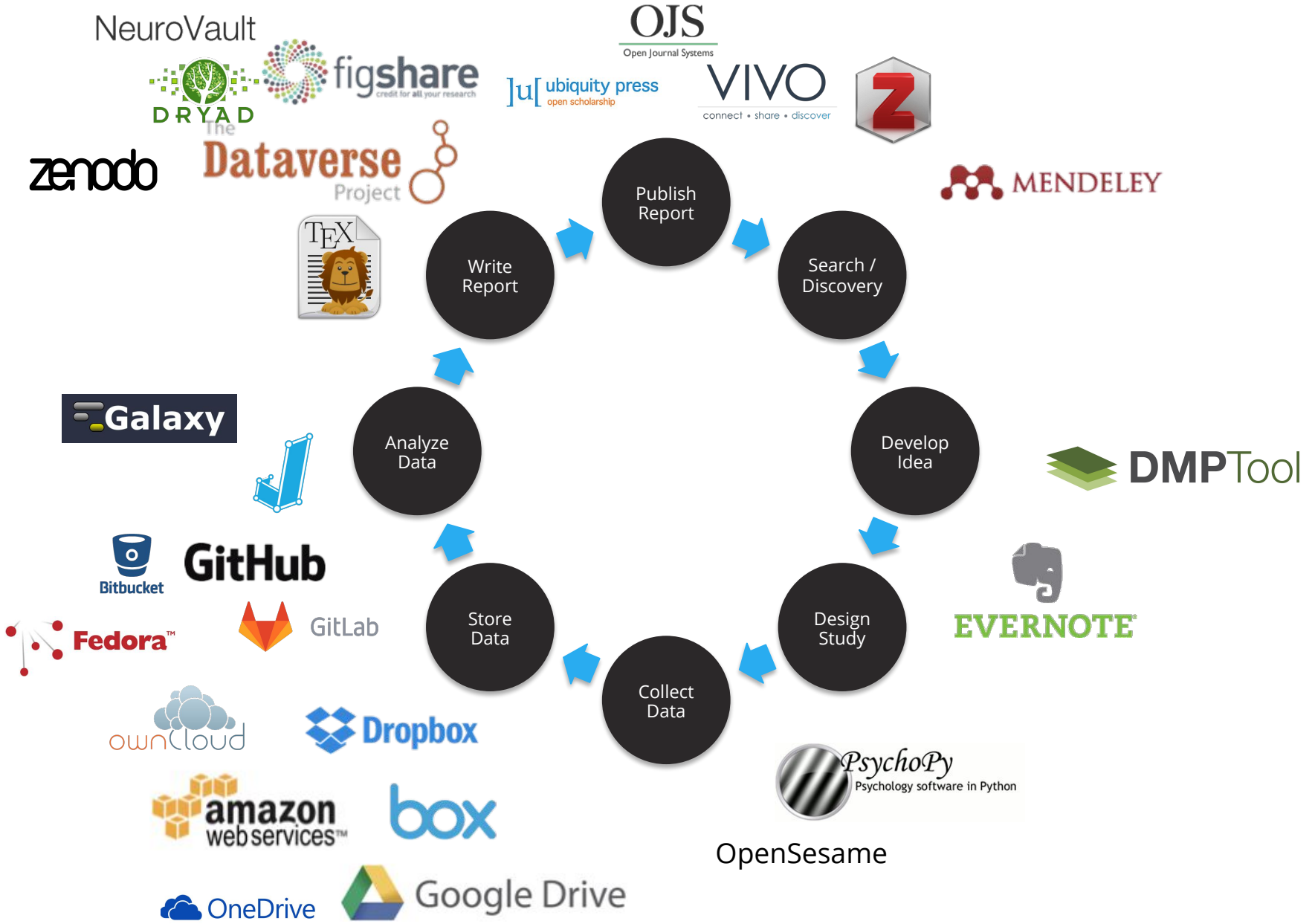
Component: Lepadogaster lep…

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




figshare
credit for all your research
The Dataverse Project
MENDELEY
Publish Report
Search / Discovery
Develop Idea
Design Study
Collect Data
Store Data
Analyze Data
Write Report
GitHub
owncloud
Dropbox
amazon web services
box
Google Drive
OpenSesame

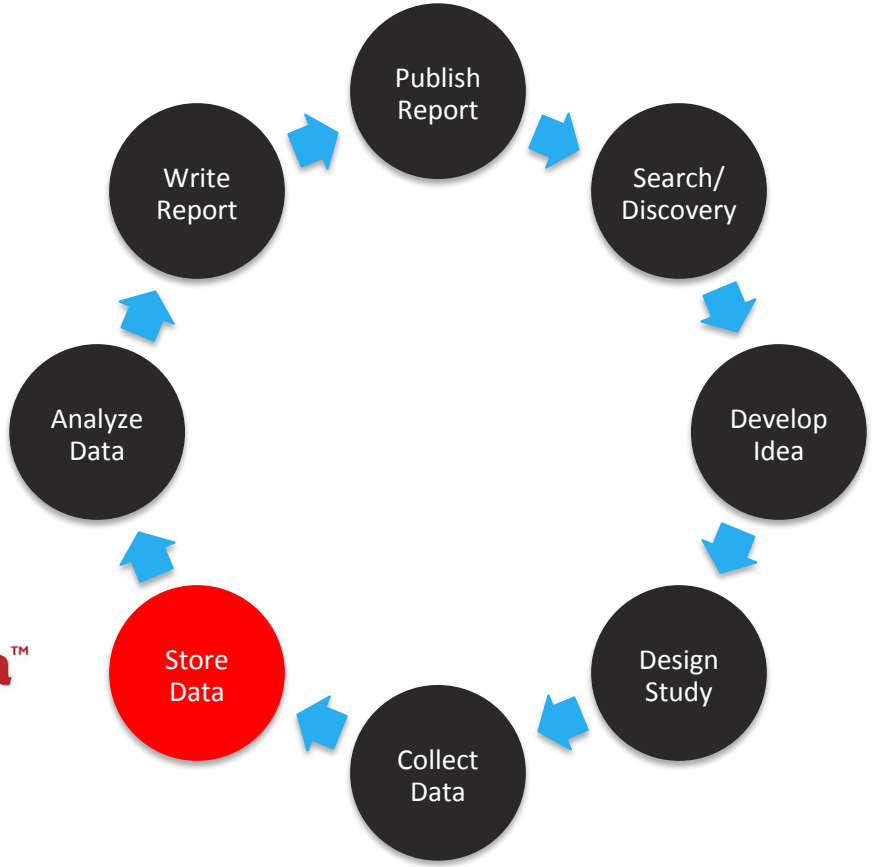
Now


NeuroVault
figshare
credit for all your research
DRYAD
OJS
Open Journal Systems
ubiquity press
open scholarship
VIVO
connect • share • discover
zenodo
The Dataverse Project
MENDELEY
Publish Report
Search / Discovery
Develop Idea
Design Study
Collect Data
Store Data
Analyze Data
Write Report
Galaxy
DMPTool
GitHub
Bitbucket
EVERNOTE
Fedora
GitLab
ownCloud
Dropbox
PsychoPy
Psychology software in Python
amazon web services
box
OpenSesame
OneDrive
Google Drive

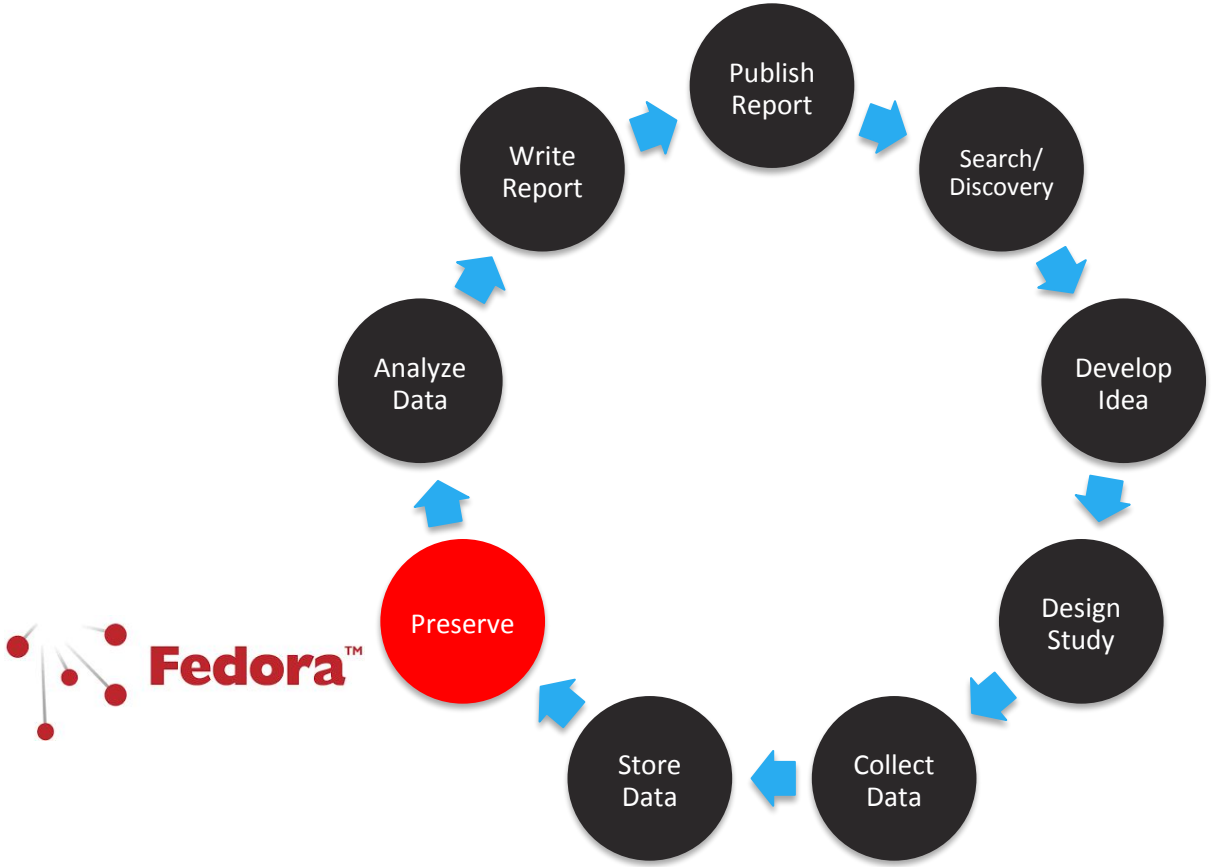
Soon


Publish Report
Search/ Discovery
Develop Idea
Design Study
Collect Data
Store Data
Analyze Data
Write Report

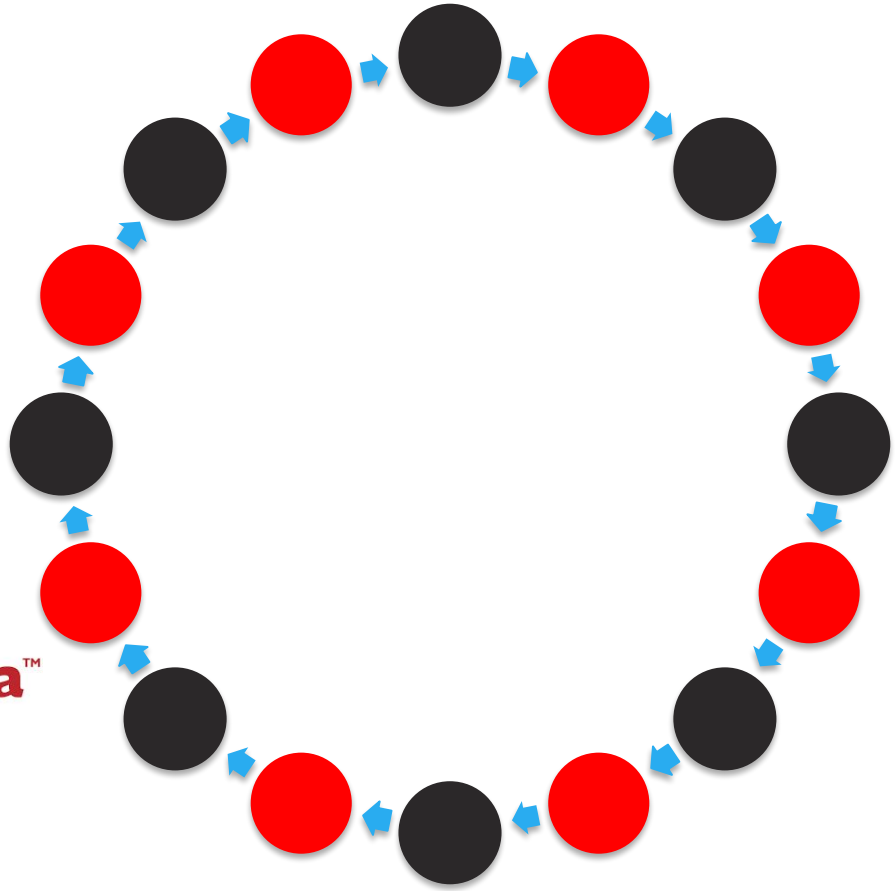
Fedora™


Publish Report
Search/ Discovery
Develop Idea
Design Study
Collect Data
Store Data
Analyze Data
Write Report

Fedora™

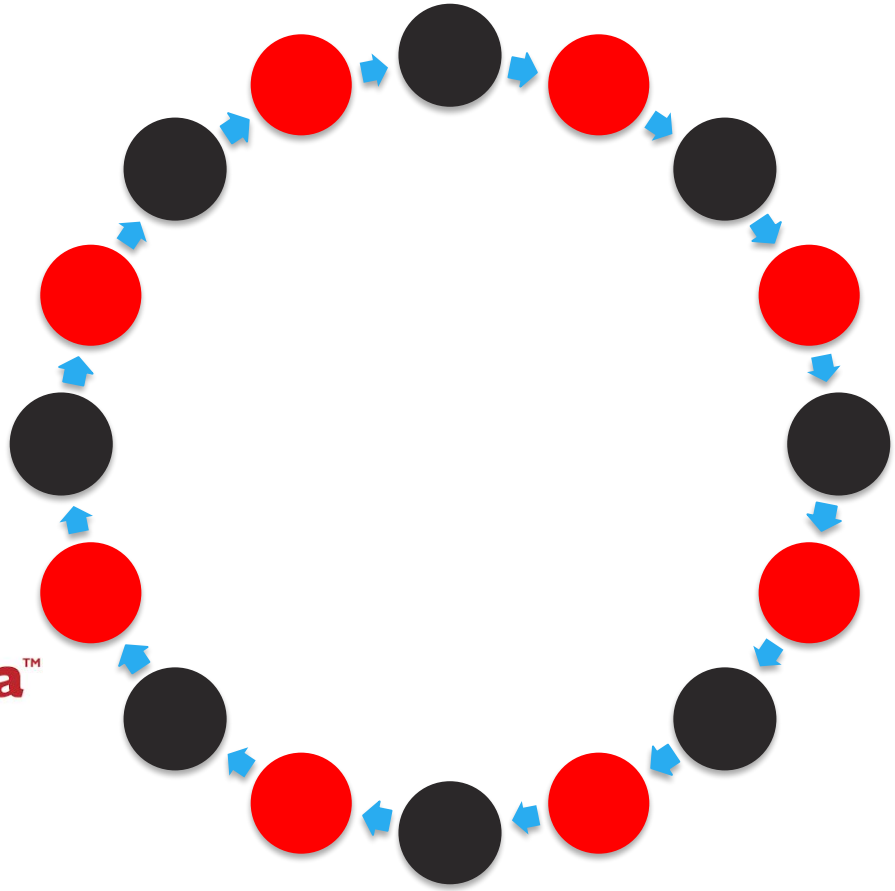


But computing is pervasive....



Fedora™

ReproZip +



jeff@cos.io

@jeffspies

Rick Johnson

- Co-Program Director of Digital Initiatives and Scholarship, Hesburgh Libraries, University of Notre Dame
- Visiting Program Officer (VPO) for SHARE, Association of Research Libraries

rick.johnson@nd.edu

@rick_nd

Notre Dame Perspectives

- Computational Analysis Reproducibility is complex
- Need to align preservation with existing workflows
- Simple, user-driven archival process
- Data flow between computational and preservation environments
- Reuse preserved data and software in other projects
- Cannot be solved just with technology

Notre Dame Data & Software Capture and Reproducibility Efforts & Partners

DASPOS Data and Software Preservation for Open Science

http://daspos.org

CurateND

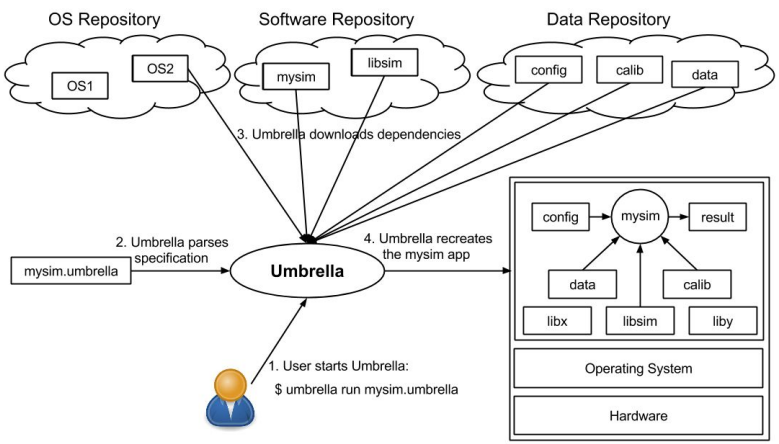
http://curate.nd.edu

COS CENTER FOR OPEN SCIENCE

http://cos.io

JOHNS HOPKINS SHERIDAN LIBRARIES

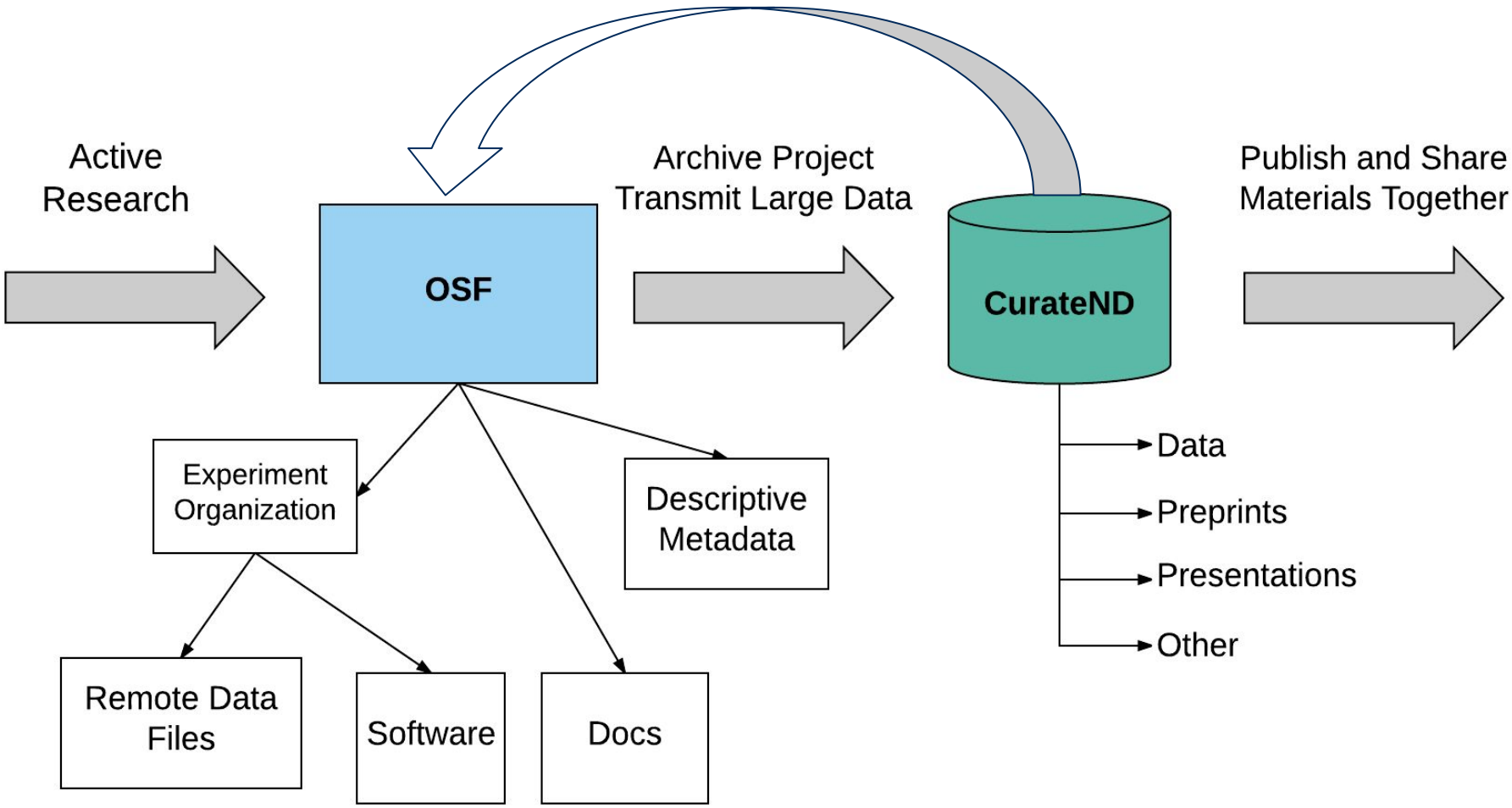
http://www.library.jhu.edu



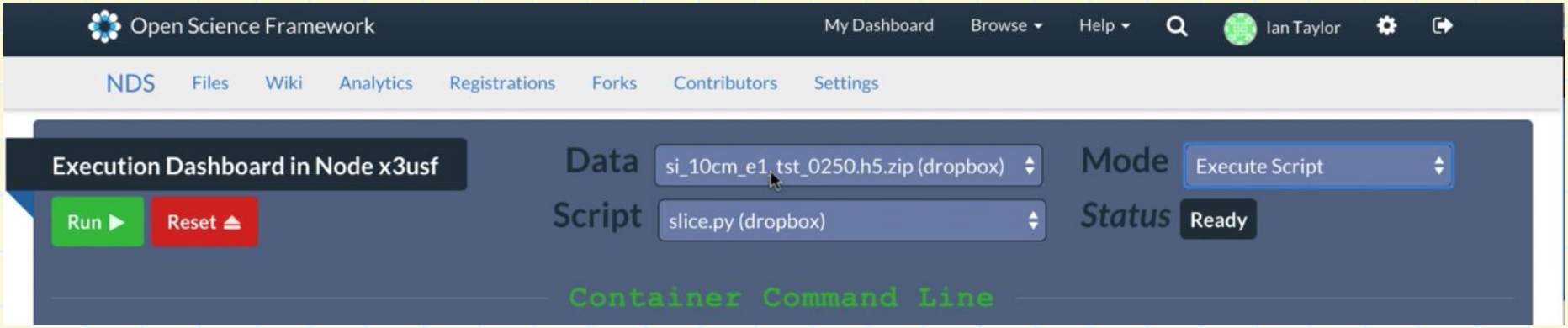
Umbrella, http://ccl.cse.nd.edu/

NATIONAL DATA SERVICE PARTICIPANT

http://www.nationaldataservice.org

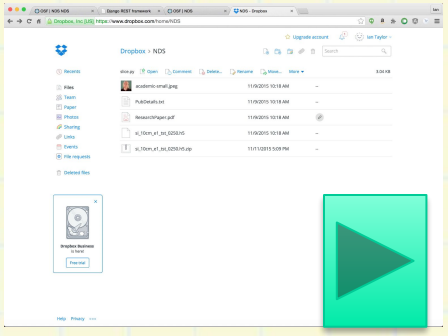

Active
Research
OSF
Archive Project
Transmit Large Data
CurateND
Publish and Share
Materials Together
Experiment
Organization
Descriptive
Metadata
Remote Data
Files
Software
Docs
Data
Preprints
Presentations
Other

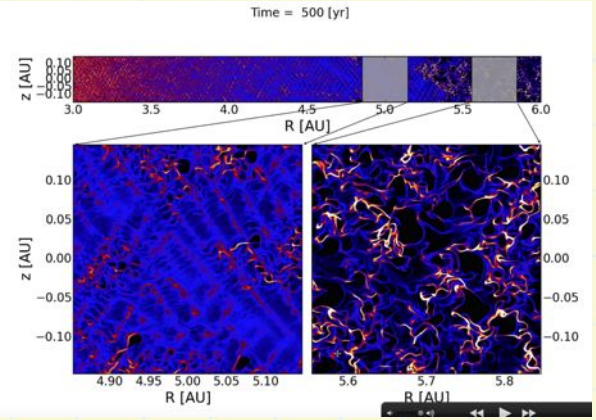
NDS OSF Dashboard integration



Contact:
Ian Taylor

ian.j.taylor@gmail.com

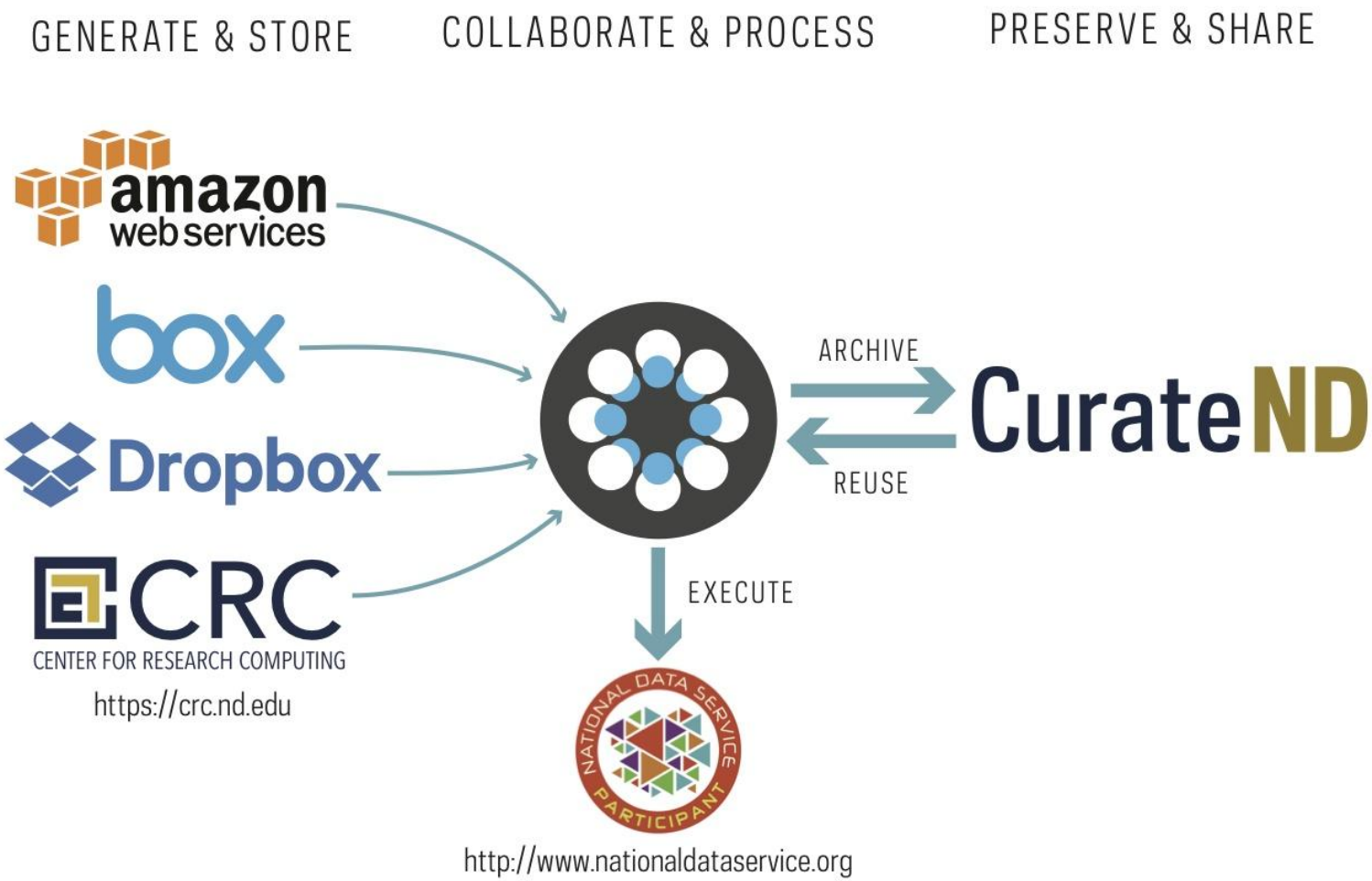




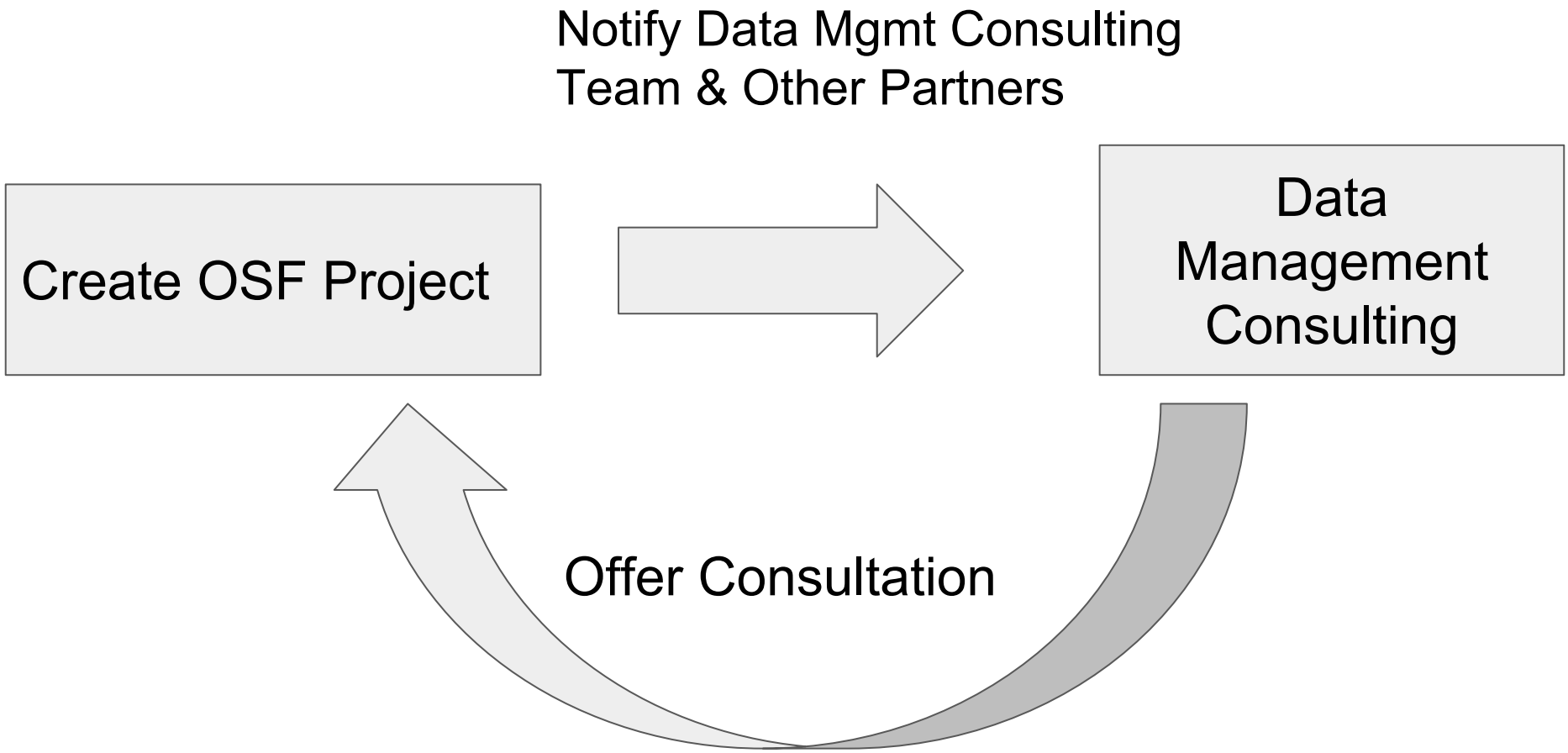
bitbucket.org/nds-org/nds-dashboard

http://www.nationaldataservice.org/
http://ndspilot.com

Computational and Preservation Environment Data Flow



Not Just Technology

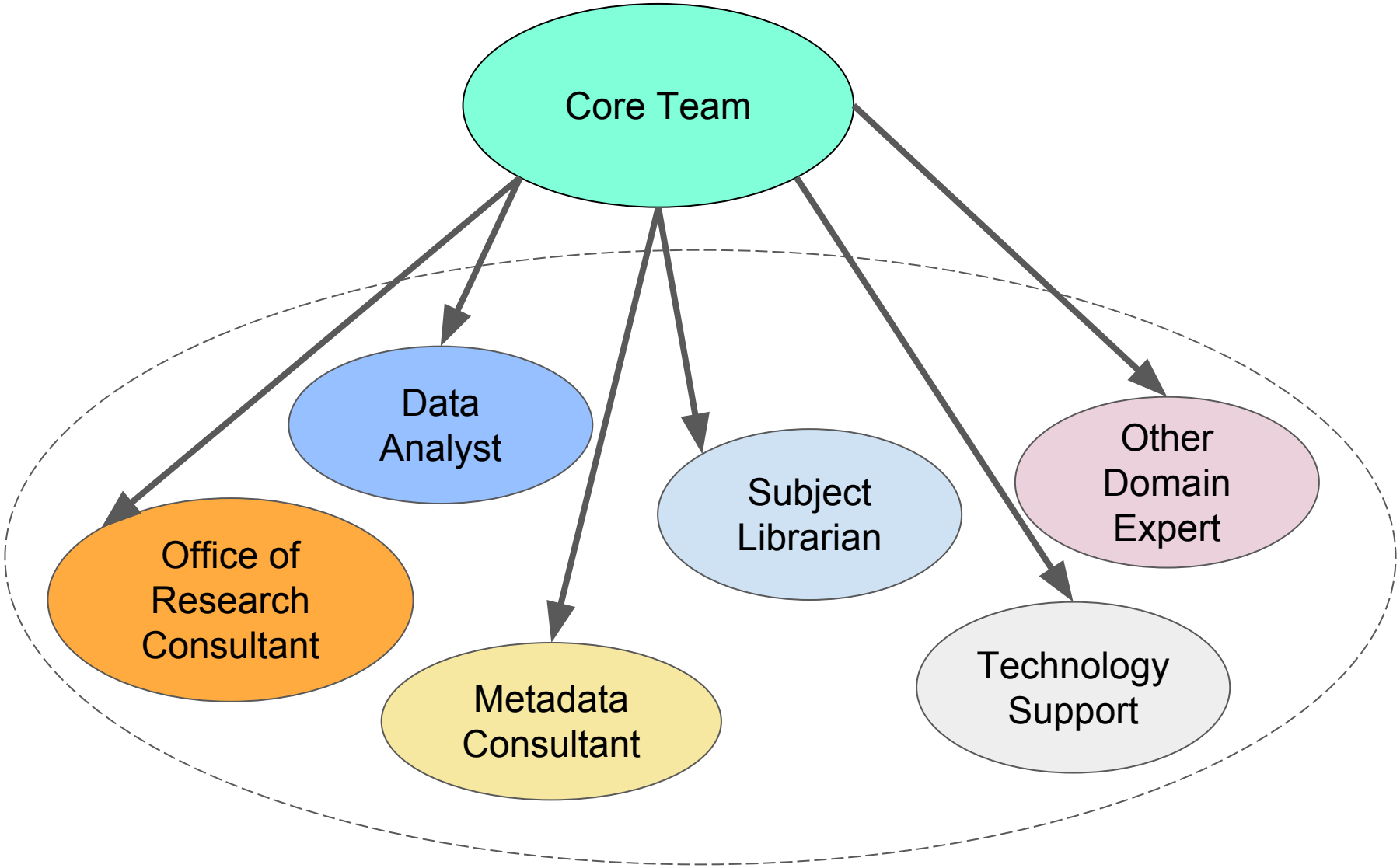


Data Management Consulting Team - Key Support Roles

UNIVERSITY OF NOTRE DAME

Role	Function
Consultant	Provides relevant knowledge, experience, and task execution
Data Analyst	Technical coordinator for data products and other computational needs associated with the project
Developer	Technical coordinator/programmer for data products and other high level computational needs
Domain Expert*	Provides project-specific contextual and cultural expertise from the perspective of the related domain
Metadata Consultant	Oversees project metadata components
Policy Expert*	Ensures project is in compliance with applicable standards and regulations
Request Lead	Key point of contact/responsibility for project

ND Virtual Data Management Consulting Team



Supporting software access and preservation

A library RDM services perspective

Fernando Rios - rios@jhu.edu
CLIR Fellow - JHU Data Management Services

Data

Software

Consulting + training
+ archiving/sharing

JOHNS HOPKINS UNIVERSITY
Data Management Services

Funder requirements for open access	Preserving the scholarly record
Publisher requirements for open access	Reproducing results
Need of specific research communities	Research efficiency via reuse

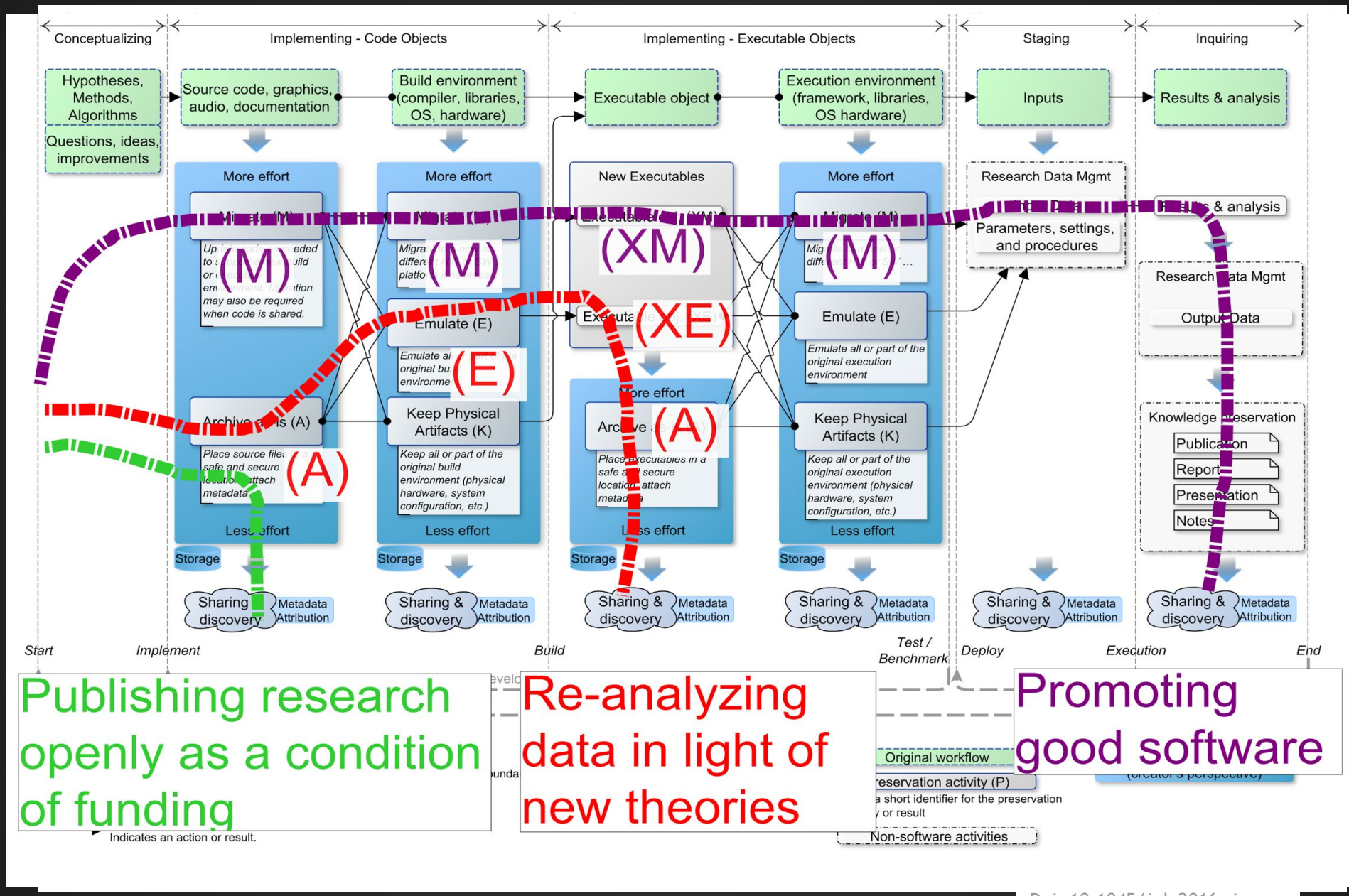
Starting from scratch

Problem: what do we even need to think about?

Scan the literature, work by different communities, current preservation offerings, technology, state of policy etc.

- Software not considered 1st class research output
- Availability of software in the literature is low
  - Even in computationally-oriented fields
- The biggest challenges are more cultural and legal in nature. Technical solutions are still not trivial (eg., use of virtualization/containers)

Identifying what we could support and how -- a plan



Doi: 10.1045/july2016-rios

Plan

Initially, support outputs tied to publication, end-of-project

- Archiving only
- What software-specific metadata to include?
- Make use of existing data infrastructure

Extend internal expertise

Archiving

Metadata: minimal/useful metadata?

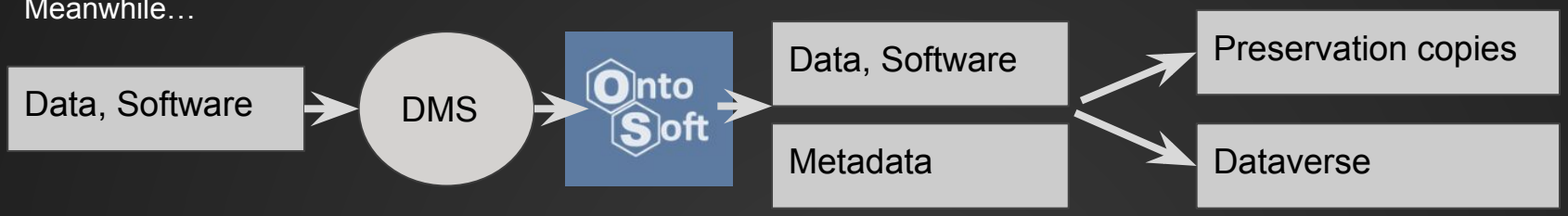
- What does it do, does it do what I want, how do I use it?

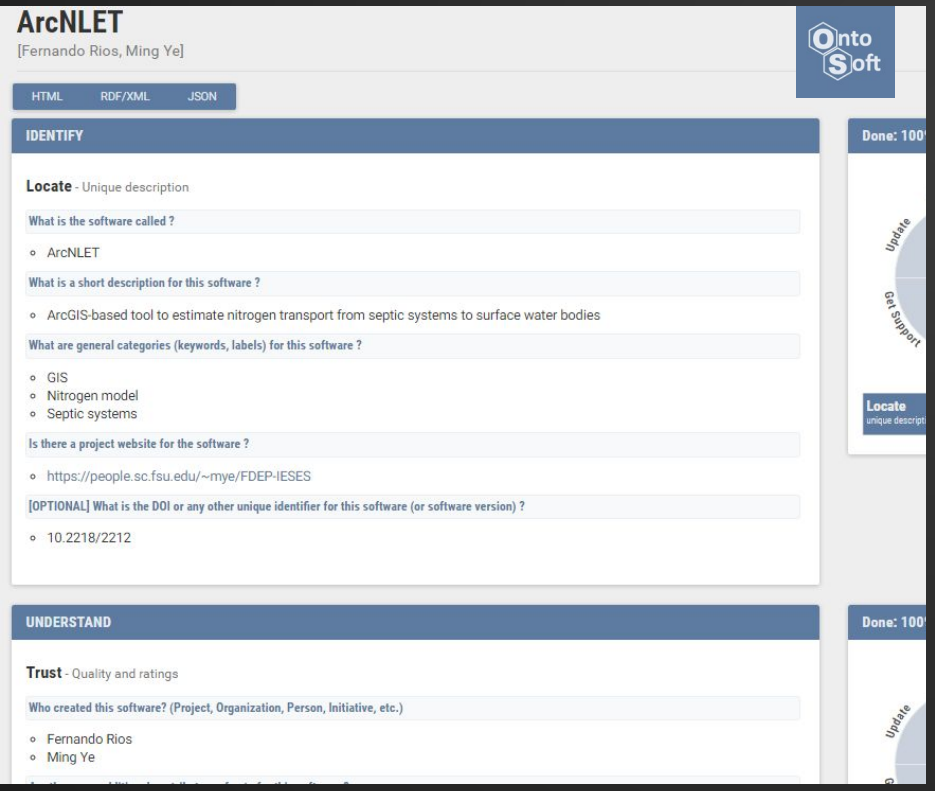
Looked at ~10 schemas

SoftwareTitle	Required (1) SW must be named.
SoftwareIdentifier	Required (1)
SoftwareContributor	Required (1+)
AuthorId	Optional (1+) ORCID etc
Version Captures version	Required (1)
ObjectType	Required (1+) Is code or exe, language used
Description	Required (1) Human-readable synopsis
RelatedLink *need ability to add later (eg if paper isn't out)	Optional (0+) e.g., website, docs, other packages in DV. Don't use for repo
CodeRepositoryLink	Optional (1) e.g., GitHub, SourceForge etc.
License	Required (1)
datePublished	Required (1)
Language	Optional (1+) Python, C++, etc.
Dependency Capures variant	Encouraged (0+) Architecture (e.g., x86), environment (CentOS 5.5 + Python 2.7 + NumPy 1.8)
InteractionMethod, Function, I/O	Encouraged (0+ of each)

Archiving

Meanwhile…





Metadata for software is deceptively easy

- Defining who should count as an "author" or contributor is not trivial.
- Granularity: at what level should it be described?

Software is highly dynamic

Workflow integration is important

- "End of project" is only one of several points of preservation

Consulting Expertise

Publication support
SW Citation Principles →

Smith AM, Katz DS, Niemeyer KE, FORCE11 Software Citation Working Group. (2016) Software citation principles. *PeerJ Computer Science* 2:e86 https://doi.org/10.7717/peerj-cs.86

Funder guidelines/DMPs, publishers

Understanding best practices for reproducible computational work

How to communicate about software preservation

Summary

Extending RDM services to (basic) software services is doable w/o huge changes to existing data practices

Archiving is 1st step

- full workflow capture and replay would be a dream goal
- Service models?

Explore "curation-ready" software

- Involvement in active research phase