SIAM Conference on Computational Science and Engineering (CSE19)

How to Assure Quality of Software Preservation Early in a Project Life Cycle and Ongoing Efforts

Sandra Gesing, Natalie Meyers, Richard Johnson, John Wang

sandra.gesing@nd.edu

https://presqt.crc.nd.edu/

MS2 Scientific Software: Practices, Concerns, and Solution Strategies

February 25, 2019

Bridging the Gap to Data and Software Sharing

Image Credit Peter Alfred Hess (CC BY 2.0)

Researchers

“the local academic community struggles to effectively manage its assets which manifested itself in a number of challenges, and as for researchers, they lacked storage capacity and data curation processes, and the institution lacked standard metadata and indexing technologies, as well as tools that would support the whole research workflow” - Digital Asset Strategy Committee, DigitalND, 2011

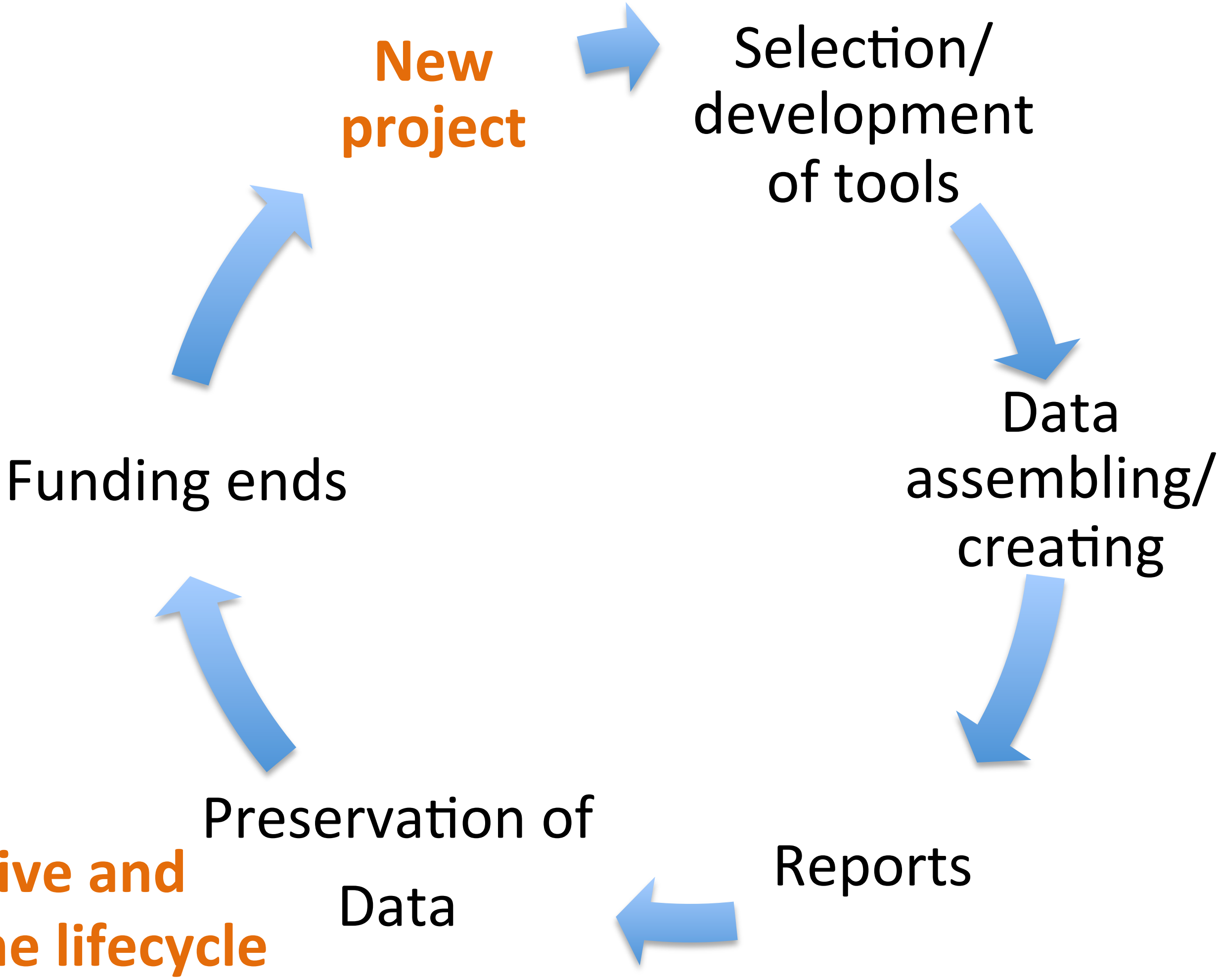
Libraries

Typically, data curation happens retroactively, and as a result data is either not captured at all or available metadata is incomplete.

Pressures from the Outside

“...digitally formatted scientific data resulting from unclassified research supported wholly or in part should be stored and publicly accessible to search, retrieve, and analyze.” - White House OSTP Public Access Memo, Feb. 2013

Current Lifecycle of Research Projects



New project

Selection/ development of tools

Data assembling/ creating

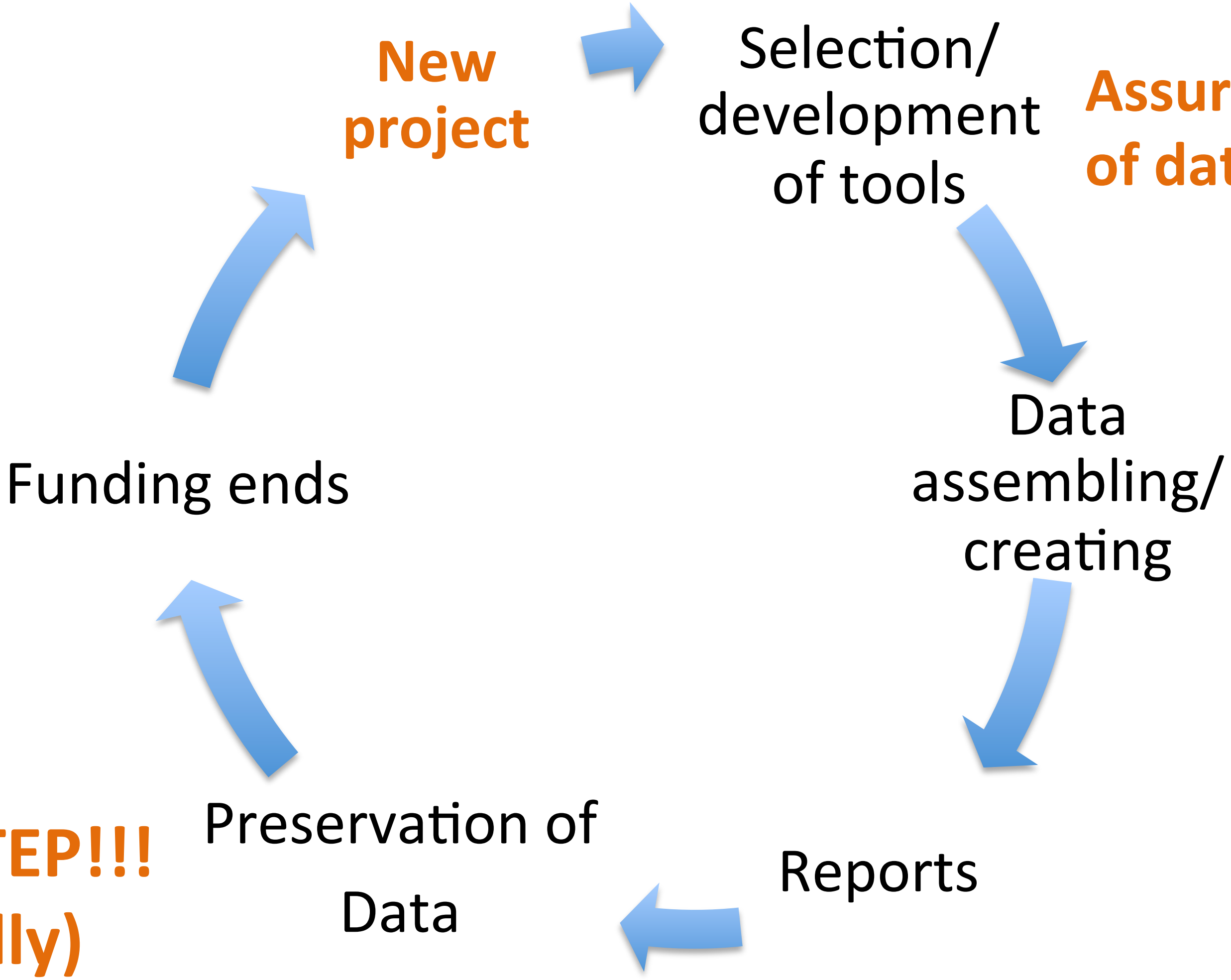
Reports

Preservation of Data

Funding ends

Work-intensive and too late in the lifecycle

Target Lifecycle of Research Projects



New project

Selection/ development of tools

Assure quality of data

Data assembling/ creating

Assure quality of data

Reports

Preservation of Data

EASY STEP!!! (ideally)

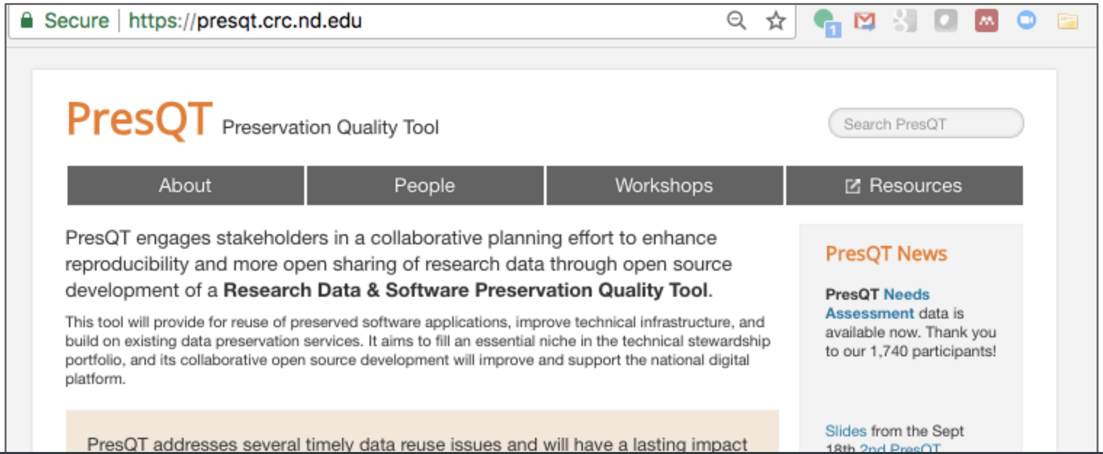
Funding ends

PresQT

A collaborative design effort to enhance reproducibility and more **open sharing** of research data through **open source development** (July 2018-June 2020) of **Tools and RESTful Services to Improve Preservation and Re-use of Research Data & Software**.

INSTITUTE *of* Museum and Library SERVICES

https://www.imls.gov/grants/awarded/lg-72-16-0122-16
https://www.imls.gov/grants/awarded/lg-70-18-0082-18



PresQT Preservation Quality Tool

PresQT engages stakeholders in a collaborative planning effort to enhance reproducibility and more open sharing of research data through open source development of a **Research Data & Software Preservation Quality Tool.**

This tool will provide for reuse of preserved software applications, improve technical infrastructure, and build on existing data preservation services. It aims to fill an essential niche in the technical stewardship portfolio, and its collaborative open source development will improve and support the national digital platform.

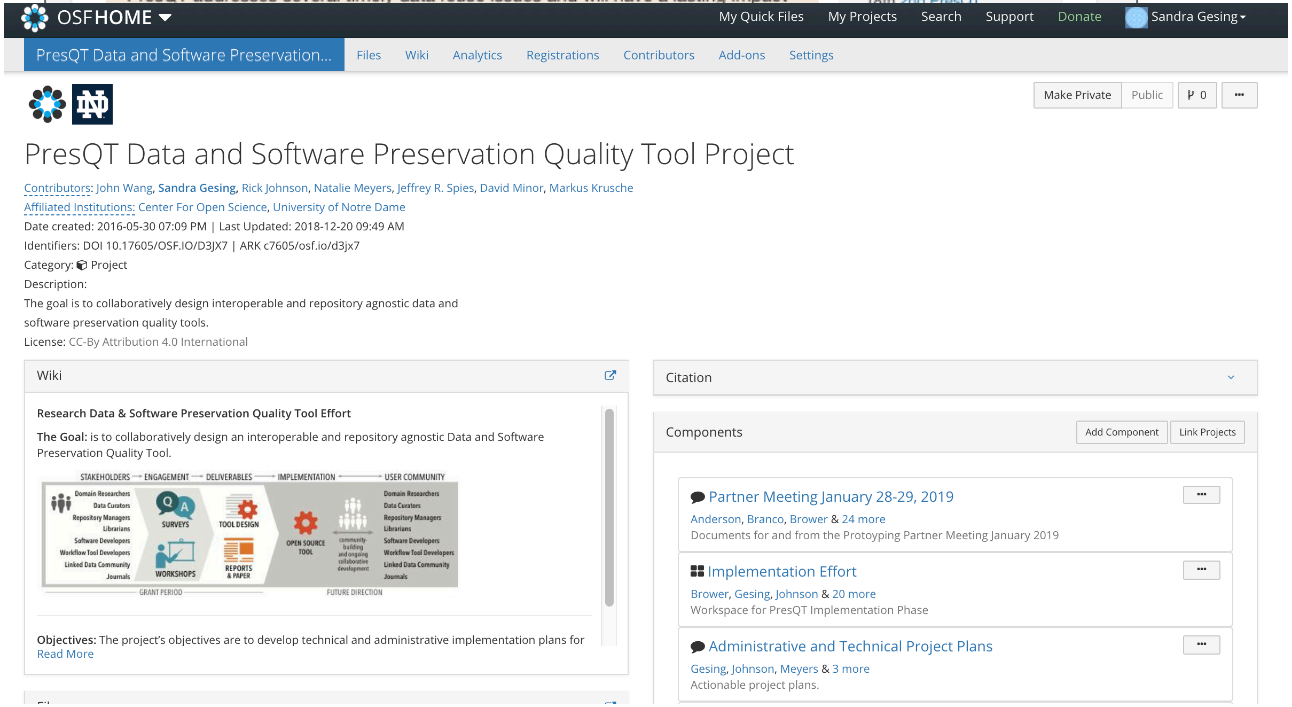
PresQT addresses several timely data reuse issues and will have a lasting impact

PresQT News

PresQT **Needs Assessment** data is available now. Thank you to our 1,740 participants!

Slides from the Sept 18th 2nd PresQT

http://presqt.crc.nd.edu/



PresQT Data and Software Preservation Quality Tool Project

Contributors: John Wang, **Sandra Gesing**, Rick Johnson, Natalie Meyers, Jeffrey R. Spies, David Minor, Markus Krusche

Affiliated Institutions: Center For Open Science, University of Notre Dame

Date created: 2016-05-30 07:09 PM | Last Updated: 2018-12-20 09:49 AM

Identifiers: DOI 10.17605/OSF.IO/D3JX7 | ARK c7605/osf.io/d3jx7

Category: Project

Description:

The goal is to collaboratively design interoperable and repository agnostic data and software preservation quality tools.

License: CC-By Attribution 4.0 International

Wiki

Research Data & Software Preservation Quality Tool Effort

The Goal: is to collaboratively design an interoperable and repository agnostic Data and Software Preservation Quality Tool.

Objectives: The project's objectives are to develop technical and administrative implementation plans for

Read More

Citation

Components

Partner Meeting January 28-29, 2019
Anderson, Branco, Brower & 24 more
Documents for and from the Protoyping Partner Meeting January 2019

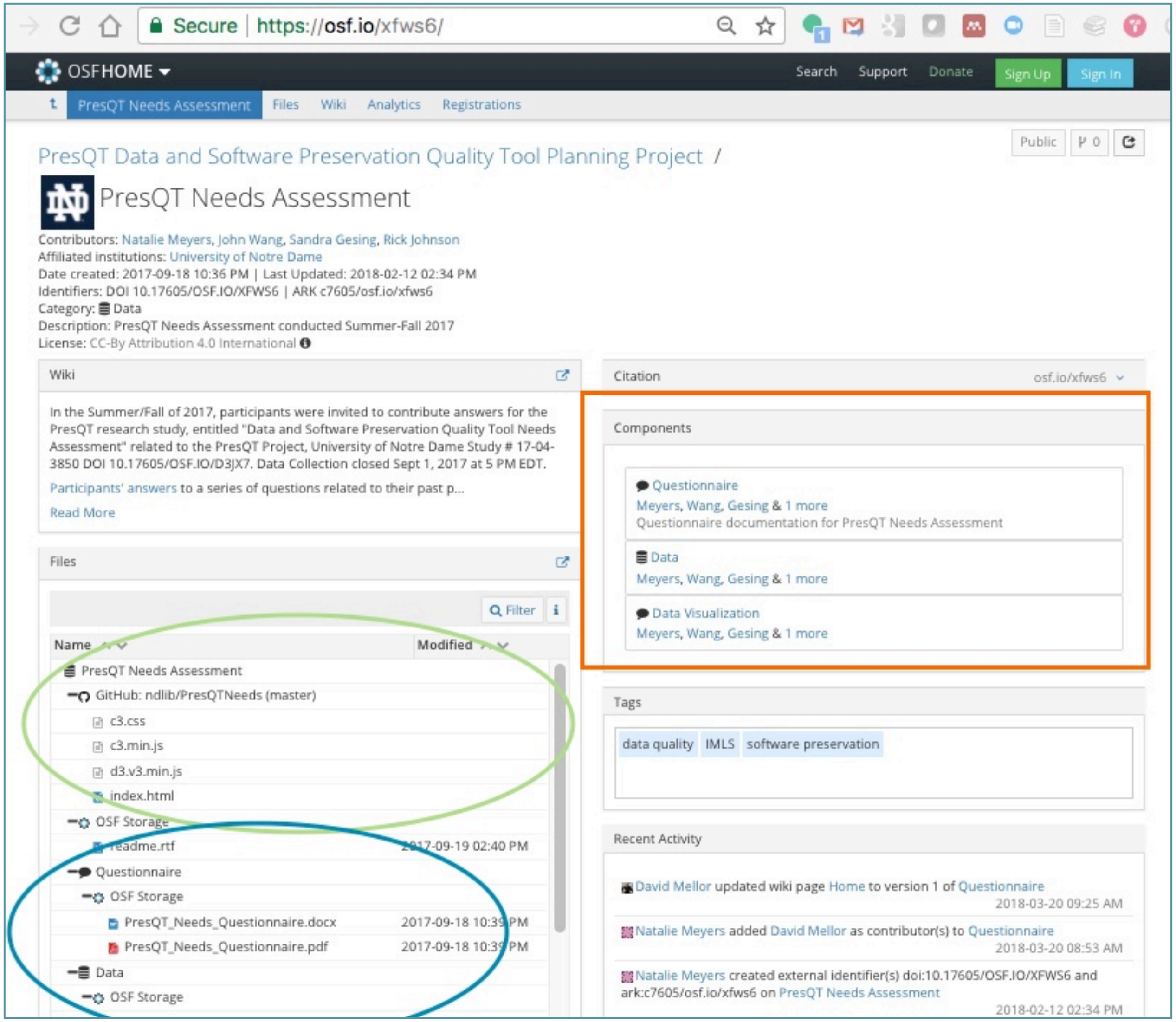
Implementation Effort
Brower, Gesing, Johnson & 20 more
Workspace for PresQT Implementation Phase

Administrative and Technical Project Plans
Gesing, Johnson, Meyers & 3 more
Actionable project plans.

https://osf.io/d3jx7/

All Resources available online

Needs Assessment Data & More



PresQT Data and Software Preservation Quality Tool Planning Project /

PresQT Needs Assessment

Contributors: Natalie Meyers, John Wang, Sandra Gesing, Rick Johnson
Affiliated institutions: University of Notre Dame
Date created: 2017-09-18 10:36 PM | Last Updated: 2018-02-12 02:34 PM
Identifiers: DOI 10.17605/OSF.IO/XFWS6 | ARK c7605/osf.io/xfws6
Category: Data
Description: PresQT Needs Assessment conducted Summer-Fall 2017
License: CC-By Attribution 4.0 International

Wiki

In the Summer/Fall of 2017, participants were invited to contribute answers for the PresQT research study, entitled "Data and Software Preservation Quality Tool Needs Assessment" related to the PresQT Project, University of Notre Dame Study # 17-04-3850 DOI 10.17605/OSF.IO/D3JX7. Data Collection closed Sept 1, 2017 at 5 PM EDT.

Participants' answers to a series of questions related to their past p...

Read More

Files

Filter

Name	Modified
PresQT Needs Assessment	
GitHub: ndlib/PresQTNeeds (master)	
c3.css	
c3.min.js	
d3.v3.min.js	
index.html	
OSF Storage	
readme.rtf	2017-09-19 02:40 PM
Questionnaire	
OSF Storage	
PresQT_Needs_Questionnaire.docx	2017-09-18 10:39 PM
PresQT_Needs_Questionnaire.pdf	2017-09-18 10:39 PM
Data	
OSF Storage	

Citation osf.io/xfws6

Components

Questionnaire
Meyers, Wang, Gesing & 1 more
Questionnaire documentation for PresQT Needs Assessment

Data
Meyers, Wang, Gesing & 1 more

Data Visualization
Meyers, Wang, Gesing & 1 more

Tags

data quality · IMLS · software preservation

Recent Activity

David Mellor updated wiki page Home to version 1 of Questionnaire
2018-03-20 09:25 AM

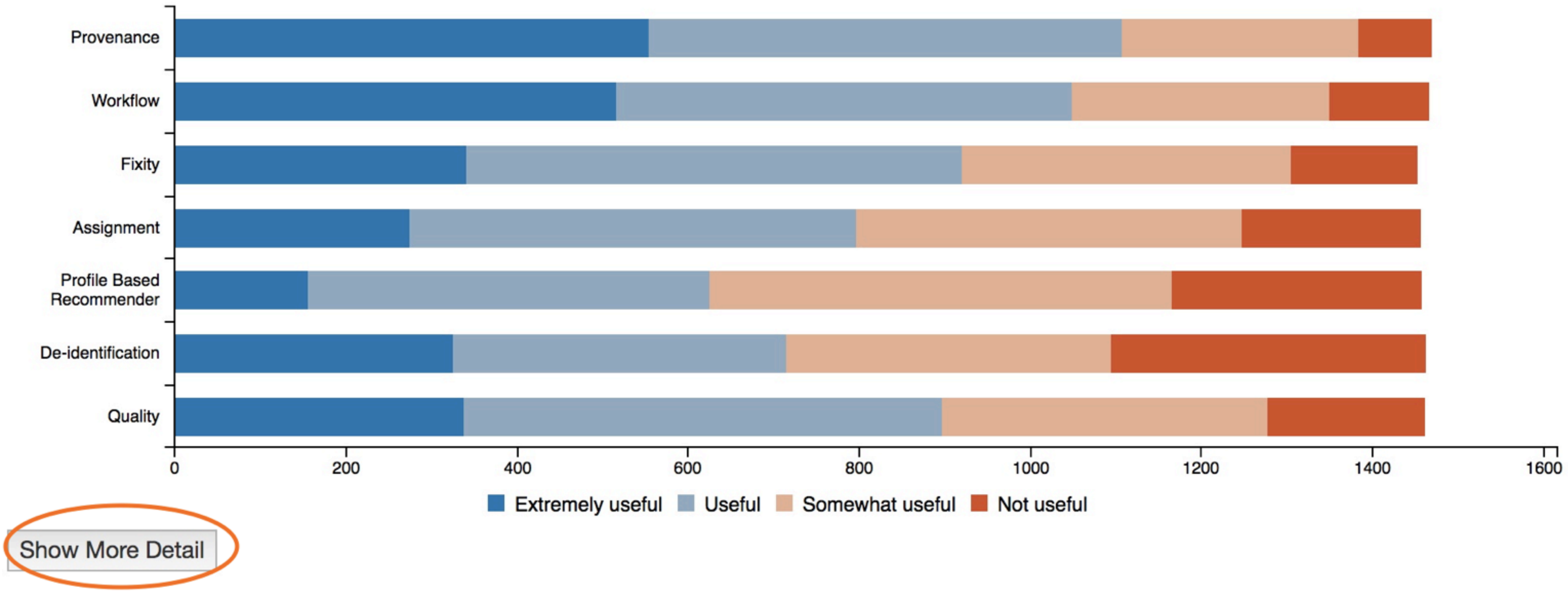
Natalie Meyers added David Mellor as contributor(s) to Questionnaire
2018-03-20 08:53 AM

Natalie Meyers created external identifier(s) doi:10.17605/OSF.IO/XFWS6 and ark:c7605/osf.io/xfws6 on PresQT Needs Assessment
2018-02-12 02:34 PM

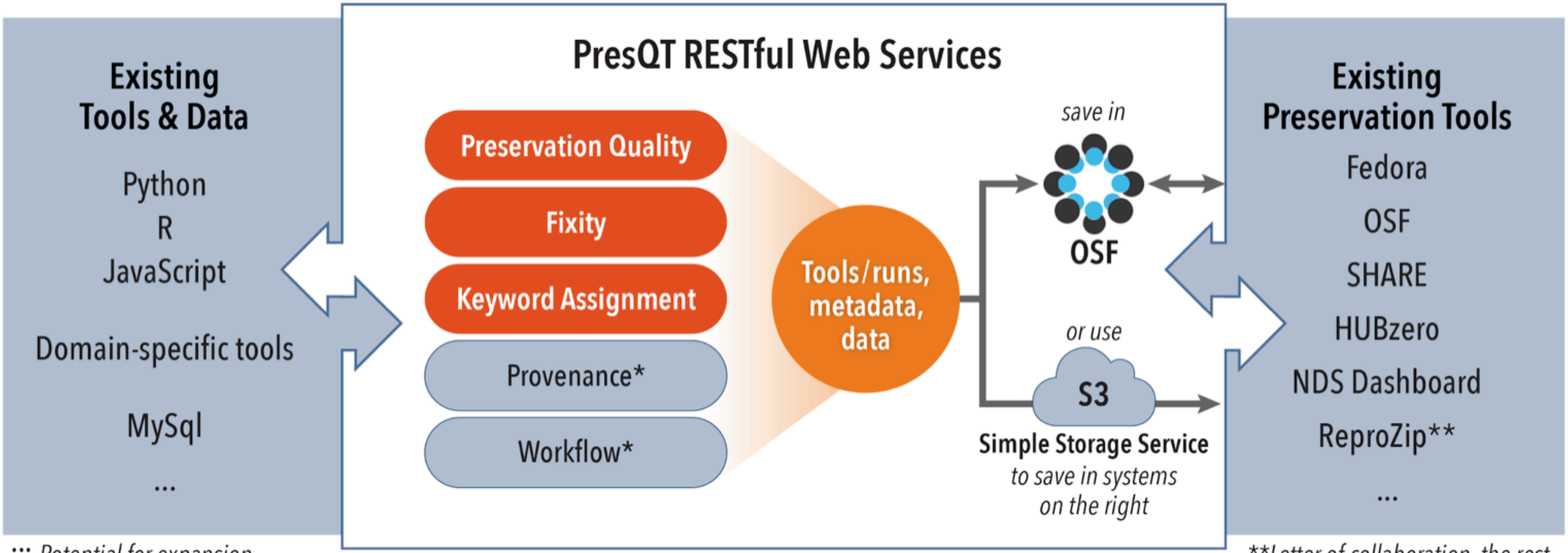
Needs Assessment Results

https://ndlib.github.io/PresQTNeeds/

Indicate whether implementation or integration of tools like those below would ease your path to publishing, sharing, curating, or reusing data or software



Repository and Tool Agnostic Solutions



... *Potential for expansion beyond initial tools/data*

**Highly ranked and already mature and available in existing systems*

***Letter of collaboration, the rest are subcontractors*

- Open design of tools and services using standards
- Integrate with workflows, tools, and virtual environments
- Priority Focus Areas

→ Available for anyone to adopt what they need and build upon it!

Open Design Goals

- The PresQT services **will not be standalone solutions** but extend the preservation tool landscape in a way that stakeholders like researchers and librarians can keep working in their chosen computational environment and receive additional features instead of having to switch to a different software. PresQT services form the connection between tools, workflows and databases to existing repositories.
- Additional preservation features in tools, workflows and repositories should be **easily integrable via standard APIs**.
- To assure a wide uptake in the community and to lower the barrier for using the novel features, we will assure a **user-centered design and collaborative development.**

Partners and Committed Collaborations

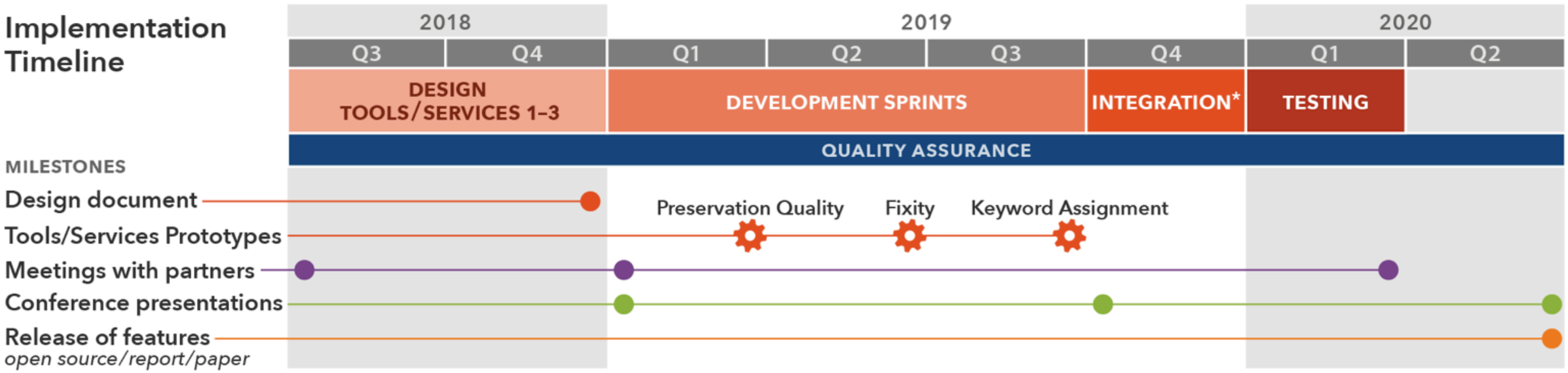
Funded sub-awardees

- Sheridan Libraries, Johns Hopkins University
- NDS
- UC San Diego Library
- HUBzero team, Purdue University
- Yale University Library (EaaSI)

Collaborators and Testing Partners

- Libraries at Amherst College, Fontbonne University, Tuskegee University, Confederation of Open Access Repositories (COAR)
- ReproZip, Jupyter, CERN, RDA groups
- Midwest Big Data Hub, Science Gateways Community Institute, URSSI, Center for Open Science, Data Curation Network, Software Preservation Network

Implementation Grant Timeline



Implementation Timeline

	2018		2019				2020	
	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2
	DESIGN TOOLS/SERVICES 1–3		DEVELOPMENT SPRINTS			INTEGRATION*	TESTING	
	QUALITY ASSURANCE							

MILESTONES

Design document

Tools/Services Prototypes: Preservation Quality, Fixity, Keyword Assignment

Meetings with partners

Conference presentations

Release of features
open source/report/paper

*Integration into OSF/Fedora/Partner products

IMLS NLG-Libraries-FY18-1: https://www.imls.gov/grants/awarded/LG-70-18-0082-18

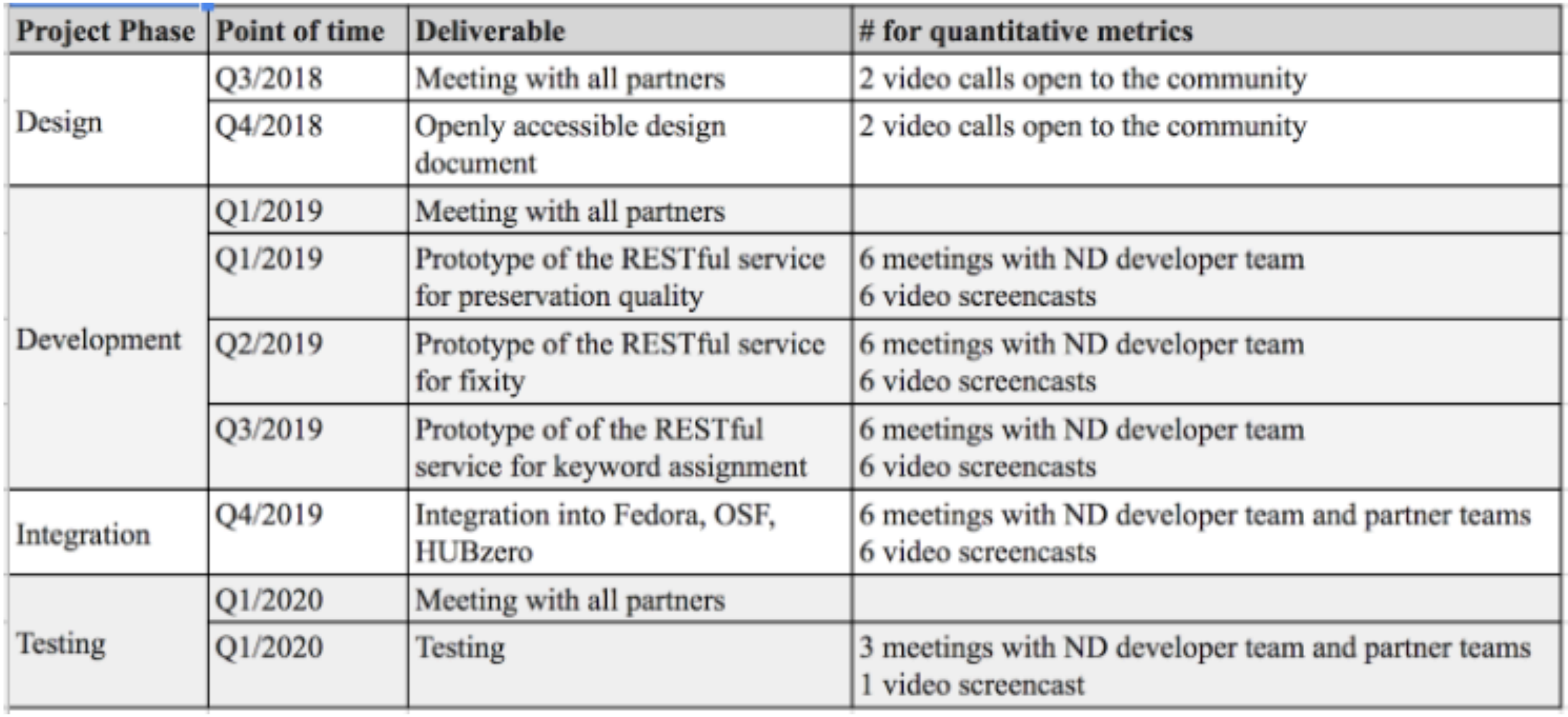
Technical Management Plan - Design

- Definition of requirements during kick-off meeting and video calls
- Definition & technical requirements from partners and stakeholders
  - Preservation quality: what does it exactly include?
  - Fixity: how to secure correct information?
  - Keyword assignment: machine-learning for previously chosen keywords?

Design Document https://osf.io/ghxec/

→

Deliverables



Project Phase	Point of time	Deliverable	# for quantitative metrics
Design	Q3/2018	Meeting with all partners	2 video calls open to the community
	Q4/2018	Openly accessible design document	2 video calls open to the community
Development	Q1/2019	Meeting with all partners	
	Q1/2019	Prototype of the RESTful service for preservation quality	6 meetings with ND developer team 6 video screencasts
	Q2/2019	Prototype of the RESTful service for fixity	6 meetings with ND developer team 6 video screencasts
	Q3/2019	Prototype of of the RESTful service for keyword assignment	6 meetings with ND developer team 6 video screencasts
Integration	Q4/2019	Integration into Fedora, OSF, HUBzero	6 meetings with ND developer team and partner teams 6 video screencasts
Testing	Q1/2020	Meeting with all partners	
	Q1/2020	Testing	3 meetings with ND developer team and partner teams 1 video screencast

Fixity Use Cases

#1 - Check file has not changed at rest over time, generate a checksum that can then be stored with the file at rest

- Then either integrated into datasource to check against checksum, or trigger check manually later?
- Have it be a web interface that is locally running javascript (open question on browser memory limits)
- For example, something like: md5sum

#2 - Verify a file has not changed after network file transfer

- Potentially a client based utility (CLI) that can generate a checksum.
- After file transfer on client supply expected checksum (transferred in separate transaction from file?), and then client generates new checksum and checks if a match

Fixity Use Cases

#3 - As aggregating files from different sources together, check to see if may already exist, and if so can combine/extract already existing metadata to add.   If so may not need to grab file again.

- May need to consider what level of granularity to validate (individual file, bag, etc.)
- Store knowledge how to canonicalize files, and have common manifests to compare

#4 - Capturing Chain of Custody and make it public and verifiable, verifying object at each change

Keyword Use Cases

#1 - Metadata Mapping Between Systems

- PresQT must support direct metadata translations between various target systems.

#2 - Keyword Suggestion and Generation

- “Expansion through ontologies…
- Employ topic modelling related techniques

Preservation Quality Use Cases

#1 - Transfer File(s) from one Repository or Workspace to another Repository

- Implementing OSF to CurateND (Notre Dame IR) first
- Will then implement with other source/target pairs with partners

#2 - Metadata Completeness Check / Score using MetaDIG

- Generate FAIR measurement to determine level of “FAIRness”

#3 - Check file if a desired file format for preservation quality, and if not recommend new format

- MetaDIG in #2 be able to meet this use case
- Deprecation warnings about particular formats

#4 - Detect if available for emulation of dependent software in emulation (e.g., EaaSi) and then connect the two together

- Check for computational reproducibility

Preservation Quality Use Case Priorities

Other Use Cases

#1: When transfer of ownership/stewardship of a resource, would like to already have metadata in place that gives provenance, usage, or generation information

- Define minimum set of metadata to capture

#2: Ask for some kind of template or blueprint for an item to preserve, examples include:

- The code itself if software (or some kind of pseudo-code of the algorithms)
- Some kind of simpler example (like stick figure drawings of choreography)
- Jisc software depositor guidelines

Community Calls

- 6 community calls recorded and shared each quarter through the end of 2019
- Next calls:
  - March 4, 2019 - 1pm EST
  - March 18, 2019 - 1pm EST
- All calls will be recorded and shared

Partners and Committed Collaborations

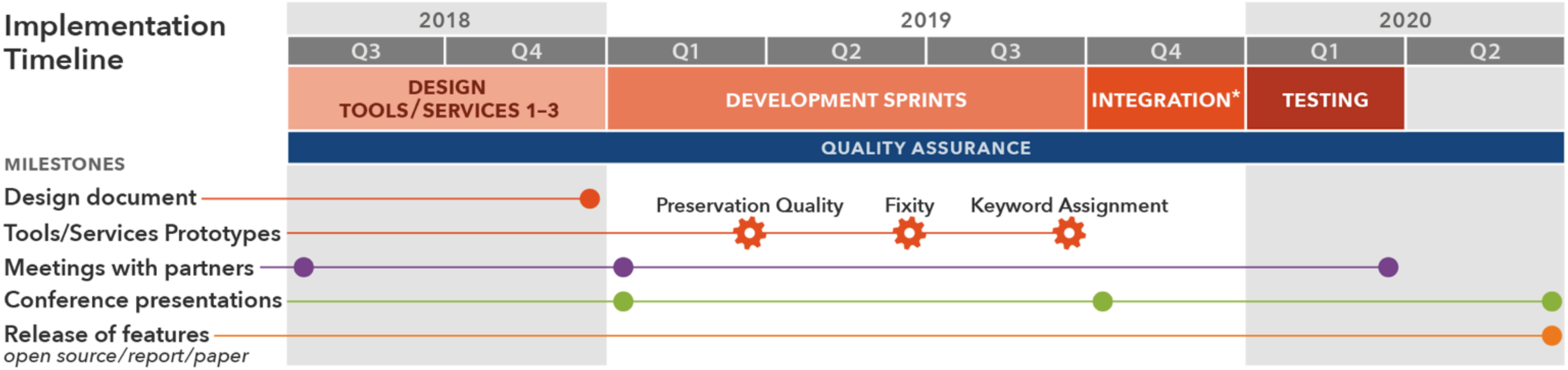
- Sheridan Libraries, John Hopkins University
- NDS
- UC San Diego Library
- HUBzero team, Purdue University
- Yale University Library

JOIN US!

- Libraries at Amherst College, Fontbonne University, Tuskegee University, Confederation of Open Access Repositories (COAR)
- ReproZip, Jupyter, CERN, RDA groups

- Midwest Big Data Hub, Science Gateways Community Institute, URSSI, Center for Open Science, Data Curation Network, Software Preservation Network

Thank you!



*Integration into OSF/Fedora/Partner products

Contact us: presqt-contact-list@nd.edu

PresQT on the web: https://presqt.crc.nd.edu/

https://osf.io/d3jx7/

Subscribe to our newsletter!