MIDWEST BIG DATA HUB

The National DATA SERVICE

PresQT: Research Data & Software Preservation Quality Tools Across the Data Ecosystem

Data Science Tools & Methods Workshop, Rosemont, IL, July 11-12, 2018

Sandra Gesing, Richard Johnson, Natalie Meyers, John Wang

presqt-contact-list@nd.edu

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

https://osf.io/d3jx7/

UNIVERSITY OF NOTRE DAME Hesburgh Libraries

CRC CENTER FOR RESEARCH COMPUTING

Bridging the Gap to Data 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

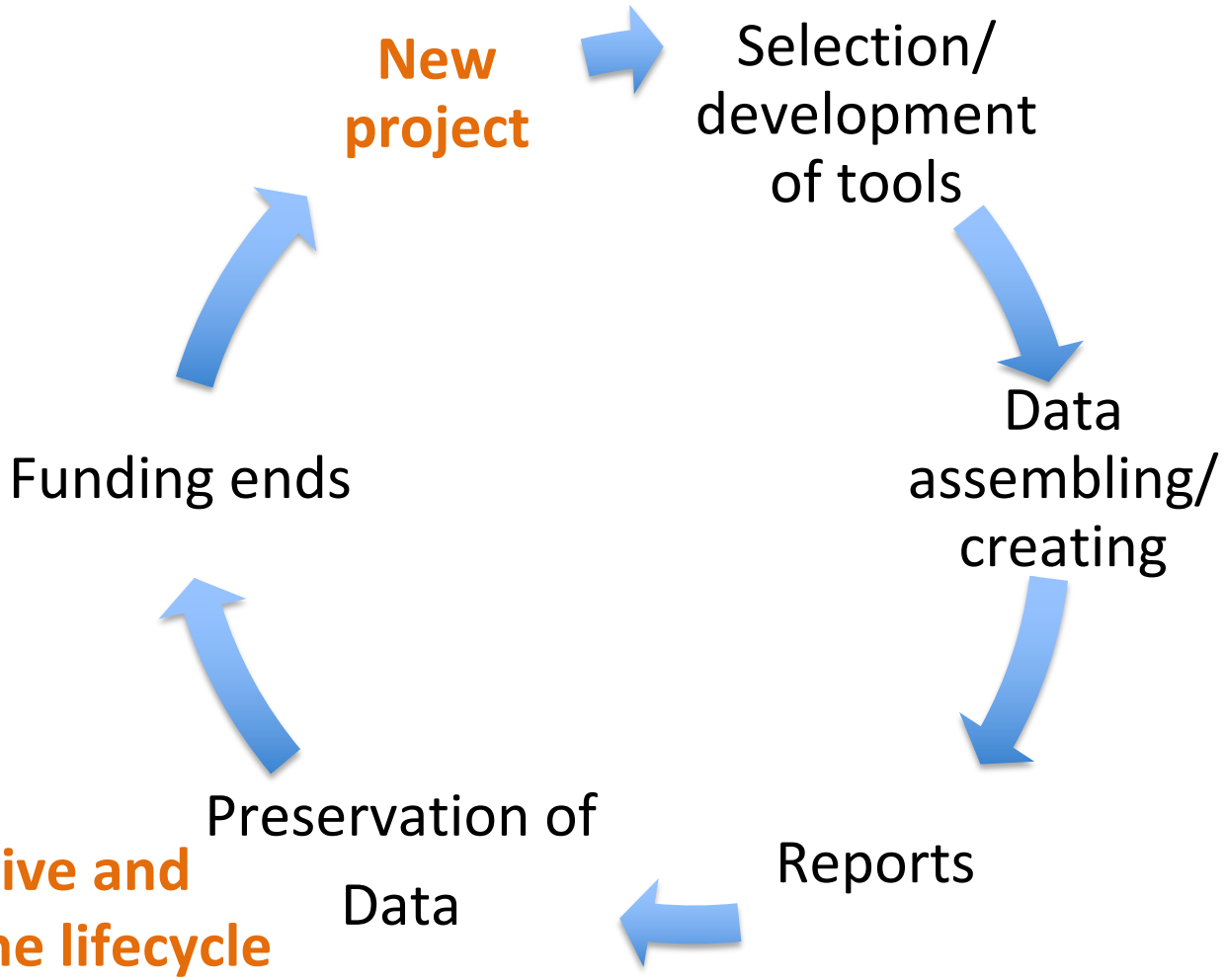
Libraries

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

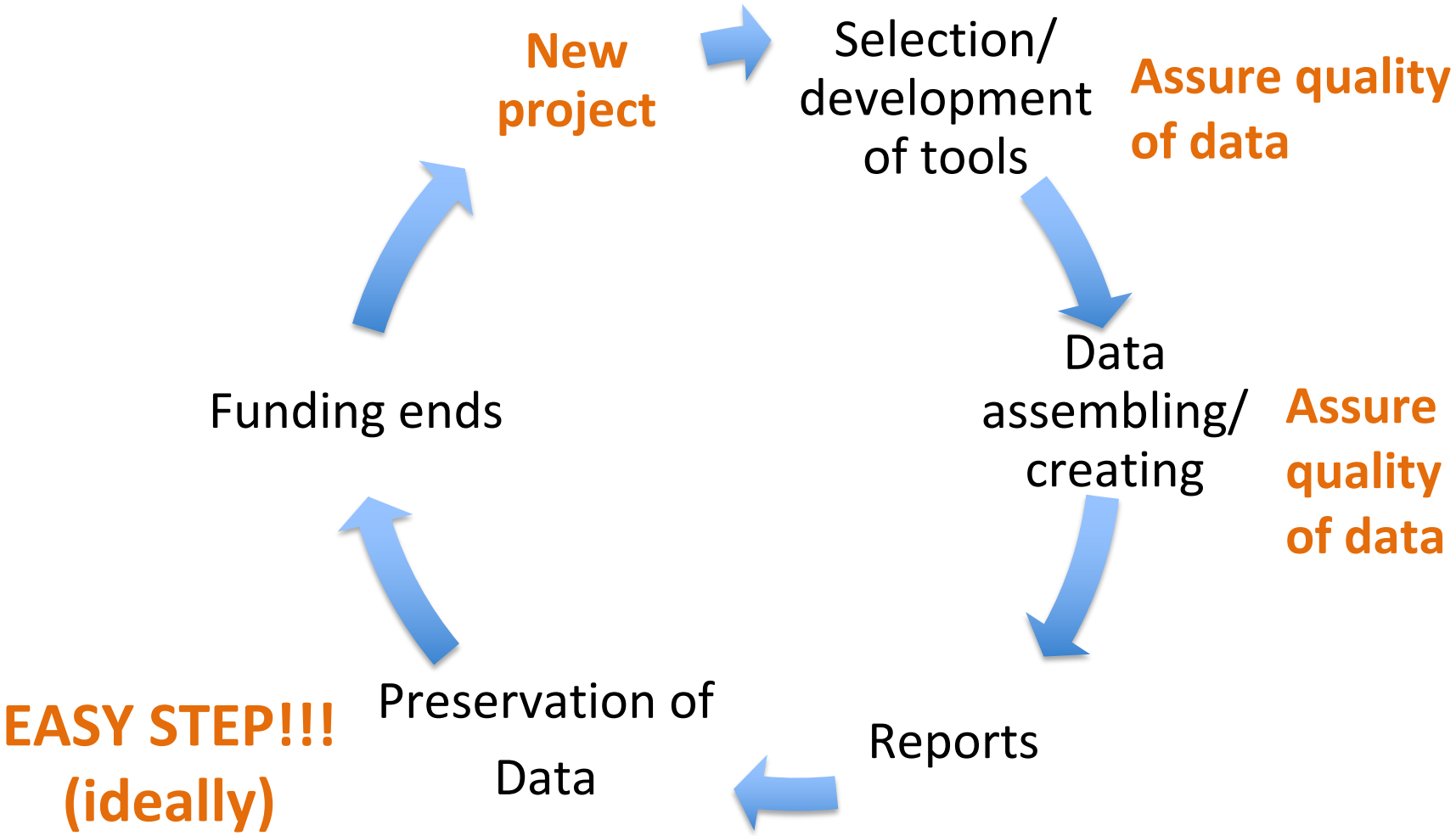
"...digitally formatted scientific data resulting from unclassified research supported wholly or in part should be stored and publicly accessible to search, retrieve, and analyze." - White House OSTP Public Access Memo, Feb. 2013

Current Lifecycle of Research Projects



Work-intensive and too late in the lifecycle

Target Lifecycle of Research Projects



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

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

PresQT Preservation Quality Tool

Search PresQT

About | People | Workshops | Resources

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 on the field by affording researchers and data curators methods to:

- Better represent digital workflow methodologies
- Improve data and software provenance
- Automatically enhance metadata
- Perform schema validation
- Improve file format recognition, interoperability and data integrity
- Facilitate scientific reproducibility

The project will design a **Research Data & Software Preservation Quality Tool**, which supports interoperability with existing platforms and solutions and improves the quality of preserved scientific digital content making it more reusable and reproducible, aligning well with the Institute of Museum and Library Services' (IMLS) goal to promote the use of technology to facilitate discovery of knowledge.

PresQT News

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

Slides from the Sept 18th 2nd PresQT workshop collocated with the RDA Plenary in Montreal are available online

PresQT project resources now available on the OSF.

View all news briefs

Open Science Framework

Everything produced for the PresQT project is shared on the Open Science Framework (OSF).

Explore Project Resources on OSF

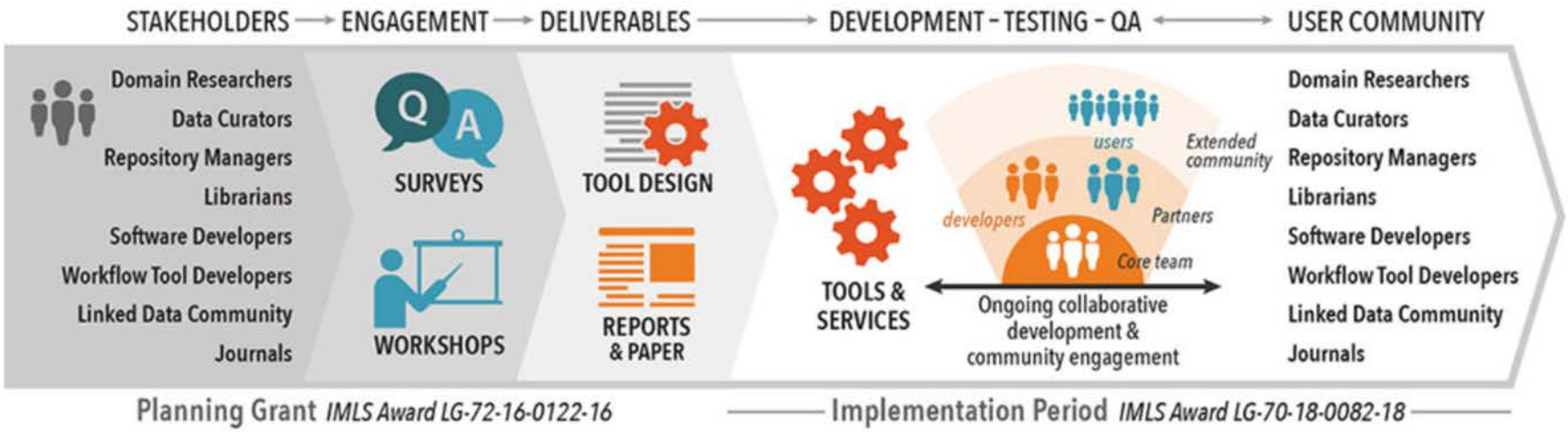
DOI: 10.17605/OSF.IO/D3JX7

Two Workshops & the Needs Assessment answered by 1740 stakeholders have been completed.

All Resources avail online

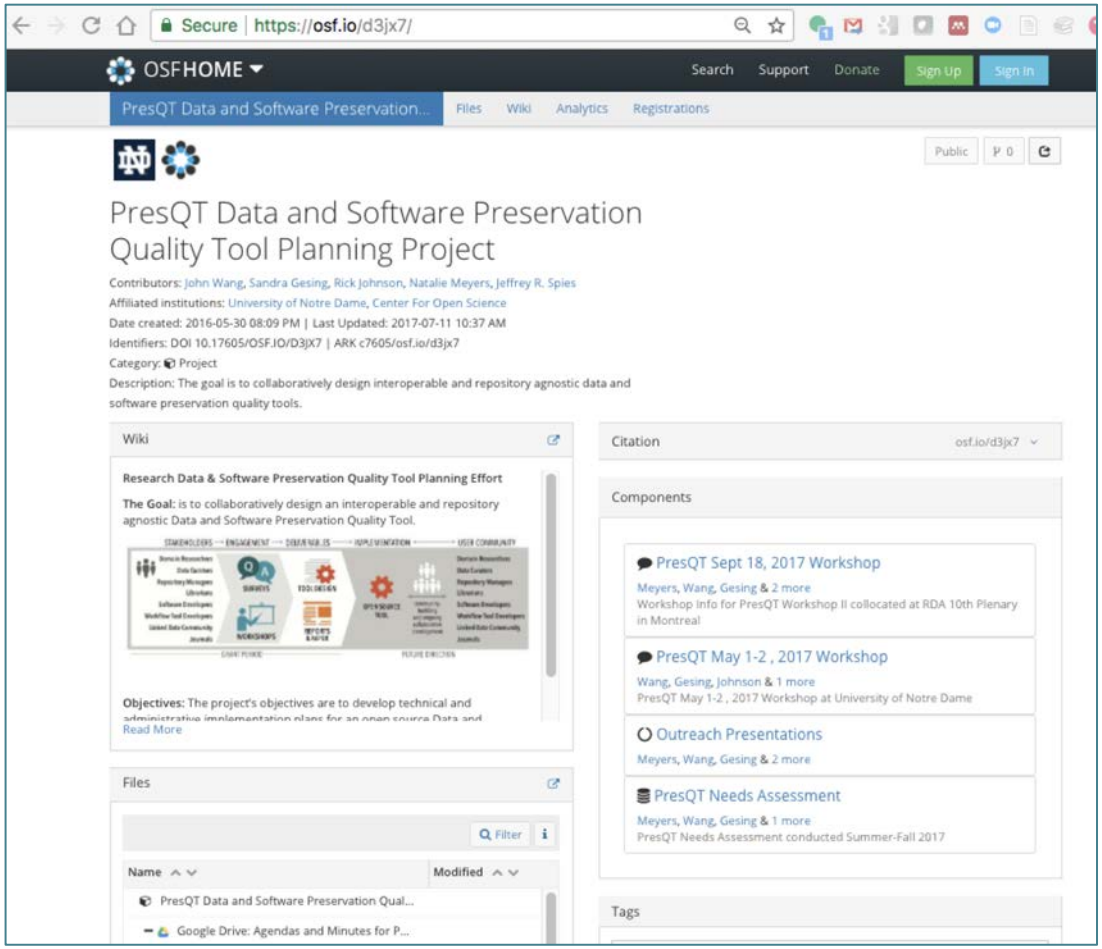
http://presqt.crc.nd.edu/

Collaborative Effort



Where we are now – Toward end of planning grant, the two workshops are concluded, Survey Results Shared, Implementation grant has started, Kick-off meeting in September

PresQT OSF Project



An open project with all stakeholder input, workshop materials, and meeting info shared on Open Science Framework.

Project Partner

CENTER FOR OPEN SCIENCE

https://cos.io/

https://osf.io/d3jx7/

Needs Assessment Results

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

Secure | https://ndlib.github.io/PresQTNeeds/test.html

PresQT Needs Assessment Results

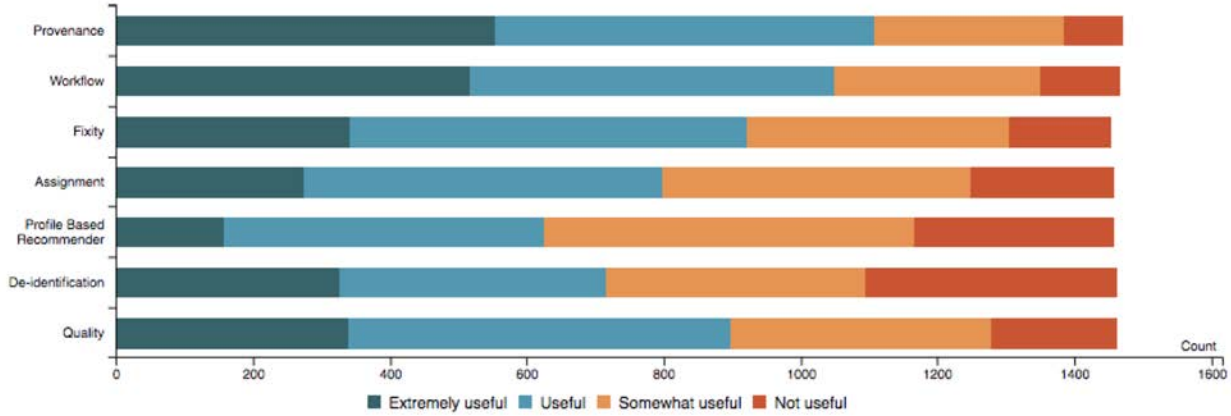
In the Summer/Fall of 2017 Participants were invited to contribute answers for the PresQT research study, entitled "Data and Software Preservation Quality Tool Needs Assessment" related to the PresQT Project, University of Notre Dame Study # 17-04-3850 DOI 10.17605/OSF.IO/D3JX7. Data Collection closed Sept 1, 2017 at 5 PM EDT. Participants' answers to a series of questions related to their past practice, and anticipated future needs as researchers and/or software developers contribute to a better understanding of what tools and/or tool suites would be of benefit those preserving and/or sharing data and software.

The Needs Assessment questionnaire and response data are available on the project page.

- Questionnaire (PDF)
- Data

Tools/Usefulness/Sort

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

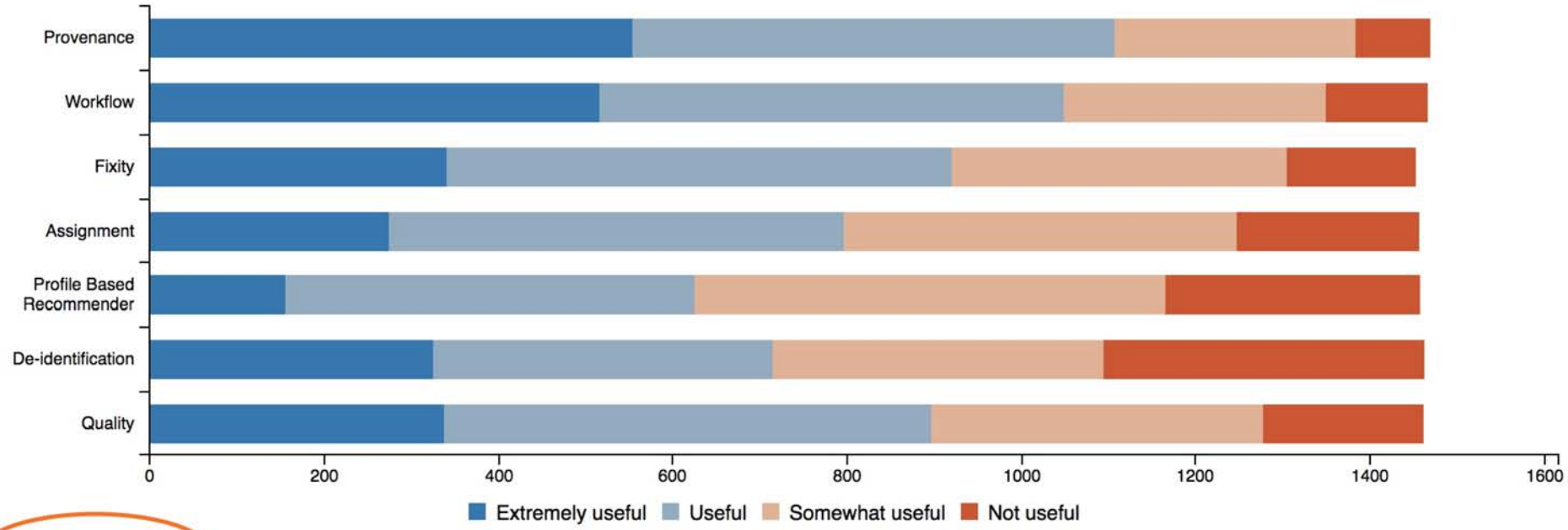


ToolsUsefulnessSort

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

	Extremely useful	Useful	Somewhat useful	Not useful
Provenance: Tools that show who did what when, or what changed when	O	O	O	O
Workflow : Tools that let you preserve your own or reuse others' workflows	O	O	O	O
Fixity: Tools that help users or data curators identify whether a digital file is fixed, or unchanged.	O	O	O	O
Keyword Assignment: Tools that automate or nudge for better or easier tagging	O	O	O	O
Profile Based Recommender: Tool that helps users identify digital resources of interest based on their profile	O	O	O	O
De-identification: Tools that make it easier to de-identify or anonymise data so you can share it	O	O	O	O
Quality:Tools that provide an assessment of a digital object's metadata completeness or preservation quality	O	O	O	O

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



Show More Detail

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

(tools_use_matrix)

Is there a tool gap in your digital ecosystem or workflow?

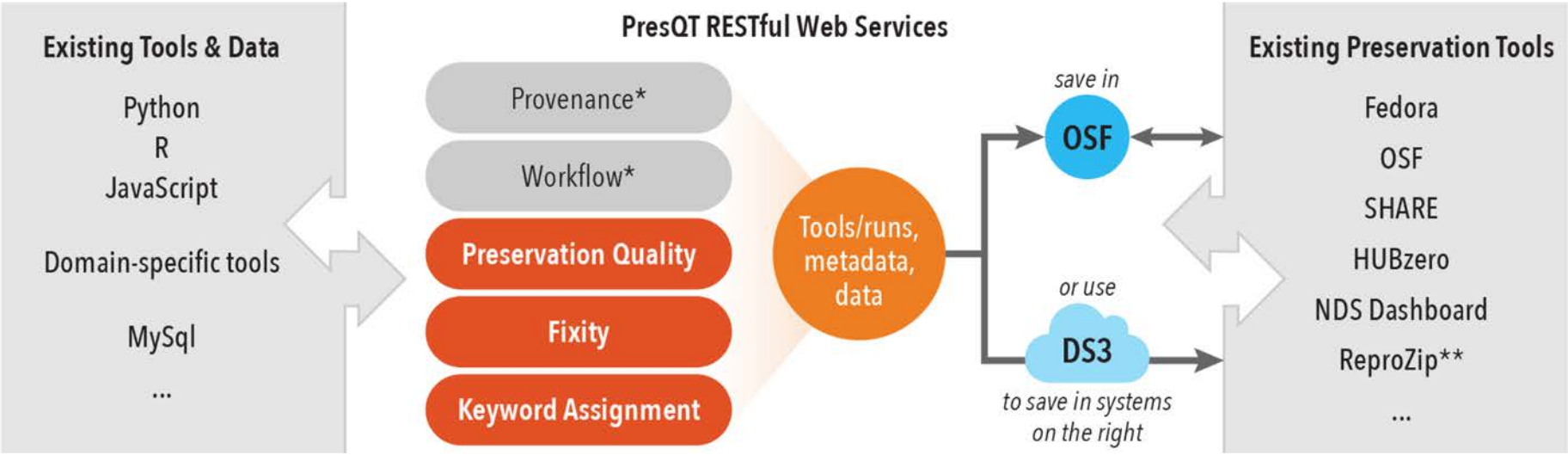
If so, please describe

We implement various ad hoc strategies for tracking workflows and provenance
We need a way to database mass spec data in a searchable way. We also need a simple way to combine mass spec sample groups for simple comparisons.
We need better facilities for making software tools work together
We need to backup multiple versions of large datasets. Also it would be useful to identify the projects and workflow associated with the datasets.
We really need to move to an electronic notebook system; however, I have many concerns that need to be addressed before doing this.
We run huge numbers of experiments, some conceived of, started and ended within minutes. It would be great to somehow archive all of this (software and data), but if it slows down our workflow then that would be problematic.

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

(tools_gap)

Repository and Tool Agnostic Solutions



- 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!

Partners and Committed Collaborations

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

- 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

Partners and Committed Collaborations

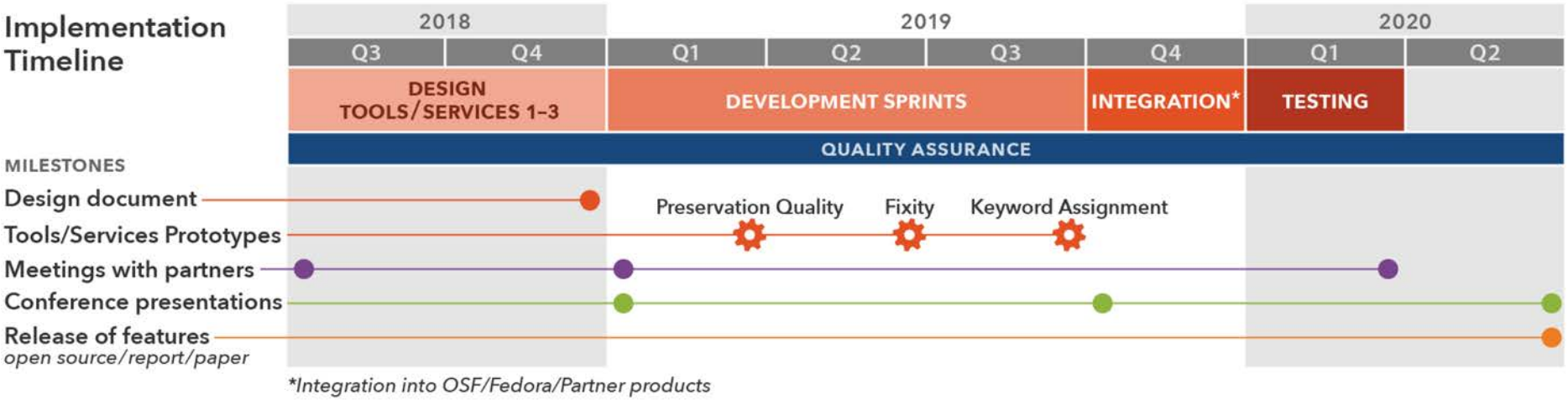
- Sheridan Libraries, John Hopkins University
- NDS
- UC San Diego Library
- HUBzero team, Purdue University
- Yale University Library

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!



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

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

Subscribe to our newsletter!

These slides are available at: https://osf.io/32z7e/