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/
Bridging the Gap to Data Sharing

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

- **Selection/development of tools**
- **Data assembling/creating**
- **Preservation of Data**
- **Funding ends**
- **New project**
- **Reports**

*Work-intensive and too late in the lifecycle*
Target Lifecycle of Research Projects

- Funding ends
- Preservation of Data
- Reports
- Selection/development of tools
- Data assembling/creating
- Assure quality of data
- New project
- Assure quality of data

EASY STEP!!! (ideally)
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.

https://www.imls.gov/grants/awarded/lg-72-16-0122-16
Two Workshops & the Needs Assessment answered by 1740 stakeholders have been completed.

All Resources avail online

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

https://osf.io/d3jx7/

Project Partner

https://cos.io/
Needs Assessment Results

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

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

<table>
<thead>
<tr>
<th>Tool Description</th>
<th>Extremely useful</th>
<th>Useful</th>
<th>Somewhat useful</th>
<th>Not useful</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provenance: Tools that show who did what when, or what changed when</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Workflow: Tools that let you preserve your own or reuse others' workflows</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Fixity: Tools that help users or data curators identify whether a digital file is</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>fixed, or unchanged.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keyword Assignment: Tools that automate or nudge for better or easier tagging</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Profile Based Recommendor: Tool that helps users identify digital resources of</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>interest based on their profile</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>De-identification: Tools that make it easier to de-identify or anonymise data</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>so you can share it</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality: Tools that provide an assessment of a digital object's metadata</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>completeness or preservation quality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Indicate whether implementation or integration of tools like those below would ease your path to publishing, sharing, curating, or reusing data or software

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.
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, URSSSI, 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/