PresQT
Assessing Researcher and Library Needs for Research Data & Software Preservation Quality Tools

CNI Spring 2018 Membership Meeting
April 12-13, 2018 | San Diego, CA
Richard Johnson, Natalie Meyers, John Wang
presqt-contact-list@nd.edu

https://presqt.crc.nd.edu/
https://osf.io/d3jx7/
PresQT

A planning grant funded effort to address needs for preserving data and software. The goal is to collaboratively design interoperable and repository agnostic Data and Software Preservation Quality Tools.

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

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

Preservation of Data

Reports

Funding ends

EASY STEP!!! (ideally)
Research Data Management Archipelago
PresQT Stakeholder Engagement

Identifying Data and Software Preservation Tool Gaps
Two Workshops & the Needs Assessment 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 Proposal Submitted
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/
Related Surveys & Case Studies

• Several documented in: *Additional Surveys of Interest to the MPS Community*
• *Science Gateways Today and Tomorrow* (2015)
• American Physical Society Open Data (March 2016)
  
  See: *Final Report: Workshops to Gauge the Impact of Requirements for Public Access to Data Produced by NSF-funded Research in Mathematics and the Physical Sciences*

  – Springer Nature in association with Figshare and Digital Science

• *Open Data The Researcher Perspective* (April 2017)
  – Elsevier and the Centre for Science and Technology Studies (CWTS)

• *Realities of Research Data Management* (2017, 2018)
  – OCLC Research Four Part Series
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)
Researcher Behavior

How familiar are you with tools used to share, publish, cite and preserve data or software?

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:

<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 Recommmender: 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/
Do you have a data or software preservation quality tool need this project could help you develop?

*If so, please describe*

**Sample of responses:**

- Wow, that's a really interesting question. My primary focus these days seems to be Visual Analytics. I would be very interested in a system for preserving and annotating visualizations. One of the most critical failures I see in final analyses of data, is that dozens of visualizations of the data may be generated, and these are so poorly annotated and cataloged that it becomes almost impossible to reproduce an identical visualization after even a few days of mental bit-rot. The result is an ever-growing stack of randomly stored visual analyses that are essentially useless, because it's impossible to completely understand their content. If you're actually interested in collaborating, this is a sufficiently interesting project that it might be worth talking to the NSF about specifically funding it.

- Yes, I am part of a team working on the development of software for biodiversity specialists. We are trying to envision all those issues in our product.

- Yes, data about pharmaceutical quality and about lead assay results

- Yes, metadata and persistent identifiers for both individual data and data bundles

https://ndlib.github.io/PresQTNeeds/
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.
In the past, how often have you made your research data free to access, reuse, repurpose, and redistribute? (res_open_data)

Is any of your data or code published or shared now on a repository or website? (res_pubshare)

Do you anticipate publishing or sharing your own data or code over the next five years? (res_pubshare_5)
In the past three years, have you or your research group made publicly accessible the following items through your or your institution's website or a third-party repository? (res_open_data3)
Do you need better tools to share, re-use, cite, publish or preserve your own or others' Data and/or Software? (res_tools_share)

https://ndlib.github.io/PresQTNeeds/
When asked to submit keywords to describe your own or others' research is it usually: (res_keywords)

When asked to submit keywords to describe your own or others' research how accurately do the terms available usually describe your work? (res_keywords_avail)
Do you actively create and manage metadata to make sharing, finding, or documenting provenance of your own or others' data or code easier? (res_metadata)

https://ndlib.github.io/PresQTNeeds/
Does your employer require you to make any of your publications or data openly available? (res_employ_open)

Do any of the organizations who fund your work require you to make any of your publications or data openly available? (res_fund_open)

https://ndlib.github.io/PresQTNeeds/
In your estimation, which of the following currently have the infrastructure required to provide long-term public access to your research data?

(res_pub_infrastruct)

To compare to CNI perspectives on IR landscape see:

- Institutional Repository Strategies: What We Learned at the Executive Roundtables (March '17)

https://ndlib.github.io/PresQTNeeds/
Where’d my Data go?

The current editors of Biostatistics indicated that when the publisher (Oxford) switched to a new publishing platform in January 2017, some of the supplemental material was lost in the transfer (personal communication, J Leek, 8 November 2017). As such, we conducted a sensitivity analysis assuming these broken links worked before Oxford changed publishing platforms.

The data analysis was made using the statistical software R (version 3.2.3).

Results

Broken links

We often encountered issues with broken hyperlinks at both journals. Forty-nine out of 76 (64%) articles that provided links to data and code at Biostatistics had broken links and at Statistics in Medicine, 21 out of 53 (40%) articles that provided links to data and code had broken links. We examine the impact of these broken links in sensitivity analyses.
How important to your work are Web-based applications that provide access to the following specialized resources? (res_tools)

https://ndlib.github.io/PresQTNeeds/
For others to reproduce your results would they need your software, code or scripts? (res_reproduce)

Do you create and/or use software, code, or scripts in your research? (res_code)

https://ndlib.github.io/PresQTNeeds/
Have you ever authored software, code or scripts to analyze or produce your data? (res_author_sw)

Have you ever hired or supervised someone to author software code or scripts to analyze or produce your data? (res_hire_swdev)

https://ndlib.github.io/PresQTNeeds/
Do you use or revise commercial and/or freeware software, scripts, or tools to analyze or produce your data more often than writing new code or hiring someone to write custom software to analyze or produce data for your project(s)? (res_external_sw)
Developer Questions

Do you collaboratively develop and/or publicly share code in an version control repository like GitHub or bitbucket? (dev_code_verCon)

Have you containerized any code to make sharing or distribution easier? (dev_code_container)

https://ndlib.github.io/PresQTNeeds/
For Software, Scripts or Code] Do you develop, administer, maintain or support any tools used to:

- **SHARE**
- **PUBLISH**
- **CITE**
- **PRESERVE**
- **OTHER**

How long do you expect people to use the software you develop, administer, maintain or support?
Needs Assessment Results

For the typical software you develop, administer, maintain or support, how many users are there:

(dev_user_count)

How do you typically license the software you develop? Select all that apply:
(dev_license)

https://ndlib.github.io/PresQTNeeds/
Visit the 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)
Priority Focus Areas for Tools that address needs

PresQT Research Workflows
- CERN
- NDS
- Local
- ICSG

Computational Environments

External Services
- ReproZip
- Fedora
- DCN
- DLI
- OSF
- R
- Matlab

PresQT Services
- Fixity
- Provenance
- Workflow Capture
- Virtualization
- Metadata Tagging

Target Lifecycle of Research Projects
- New project → Selection/development of tools → Assure quality of data
- Data assembling/creating → Assure quality of data
- Funding ends → Preservation of Data → Reports

Preservation and Sharing
- Data Registries
- Code Sharing
- Institutional Repositories
- Vended Services
- Community Repositories
- Publishers

Active Workstation(s)
Implementation Proposal Timeline

IMLS NLG-Libraries-FY18-1: Cycle One
https://www.imls.gov/nofo/national-leadership-grants-libraries-fy18-notice-funding-opportunity
Repository Agnostic Solutions

• Open design of tools and services using standards
• Integrate with workflows, tools, and virtual environments

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

presqt-contact-list@nd.edu

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

These Slides: https://osf.io/d72cm/