Tools and Services to Improve Preservation and Re-use of Research Data & Software

Brief Project Description: We propose to execute the administrative and technical project plans produced by the PresQT planning effort funded by IMLS award LG-72-16-0122-16. Coordinators at the Hesburgh Libraries and the Center for Research Computing (CRC) at the University of Notre Dame (ND), with dedicated partner participation from the Center for Open Science (COS), seek $570,544 to support community open source development and interoperability testing with stakeholders of the top 3-5 services and features identified as priorities during the PresQT planning effort. Project participants from Johns Hopkins University Sheridan Libraries, NYU Libraries, Purdue University Libraries, and UC San Diego Library along with collaborators from Software Preservation Network (SPN), Data Curation Network (DCN), National Data Service (NDS), Science Gateways Community Institute (SGCI), and Research Data Alliance (RDA) will participate as collaborative developers on interoperability testing, and/or as usability stakeholders under subaward, cost share and/or through participant support. The tools and RESTful services developed with the requested funding will enable improved reuse of preserved data and software in repository systems (e.g., Fedora), make such research data more discoverable through community aggregators (e.g., SHARE), and more interoperable with the Open Science Framework (OSF) and other Science gateway tools like HUBzero. Cost-sharing resources ($570,555) will be contributed, including: project management, other personnel time, and fringe benefits. The project furthers the IMLS agency level goal of *Creating a Nation of Learners* by improving library technology that facilitates discovery and reuse of data and software knowledge assets.

Impact on Research and Reproducibility - Addressing Fieldwide Need: Researchers and their parent institutions often respond reluctantly and retroactively to funder and publisher mandates for data and software sharing. Our project bridges gaps between existing digital library infrastructure, repositories, and software reuse. Interoperability with existing tools and platforms improves the quality of preserved scientific digital content making it more reusable and reproducible, aligning well with IMLS' goal to promote the use of technology to facilitate discovery of knowledge. The tools and RESTful services proposed fill identified areas of need in the technical stewardship portfolio. Stakeholder input via two workshops, and through the widely circulated *PresQT Needs Assessment*[1] inform our timing and approach. Over 1,700 researchers, tool developers and platform providers responded. Input from PresQT Workshop One[2] held in May 2017, and our preliminary analysis of the needs assessment survey data show a preliminary priority order of interest in: Provenance tools, Workflow Preservation & Re-use, Fixity tools, Metadata Completeness & Preservation Quality Assessment tools, Keyword Assignment, Profile-based Recommenders, and Data De-Identification[3] tools. The second PresQT Workshop[4] Sept 18, 2017[5] concludes a final round of scheduled community input. After the September workshop the list of tools will be further narrowed to prioritize development of the top 3-5 community identified features/services. We will openly and publicly share the priorities, along with our draft technical project plan well before the January 2018 full-proposal deadline. Academic library expertise from private and public universities on both coasts and in the midwest will combine to improve functionality that will benefit providers and users of existing valued formats, tools and services like BagIt, ReproZip, Fedora, HUBzero, OSF, and SHARE. Collaboration with DCN, RDA, SGCI, and SPN will focus on interoperability and usability testing. Interoperability testing with data from metadata aggregation platform providers like SHARE and Scholix implementers (e.g., Crossref[6], Data-Literature Interlinking Service[7]) will advance reproducible science and ease data re-use. The project will improve and support the national digital

[1] https://presqt.crc.nd.edu/index.php/about/survey

[2] https://presqt.crc.nd.edu/index.php/workshops/workshop-one

[3] Workshop input thus far suggests development of de-identification tools is out-of scope for this effort

[4] https://presqt.crc.nd.edu/index.php/workshops/workshop-two

[5] subsequent to this proposal deadline of Sept 1, 2017

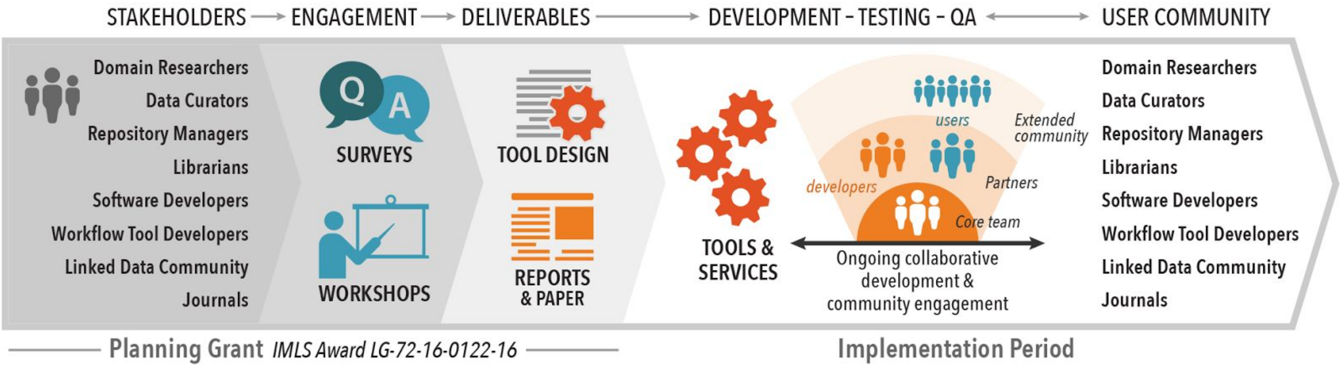
[6] https://www.crossref.org/

[7] https://dliservice.research-infrastructures.eu

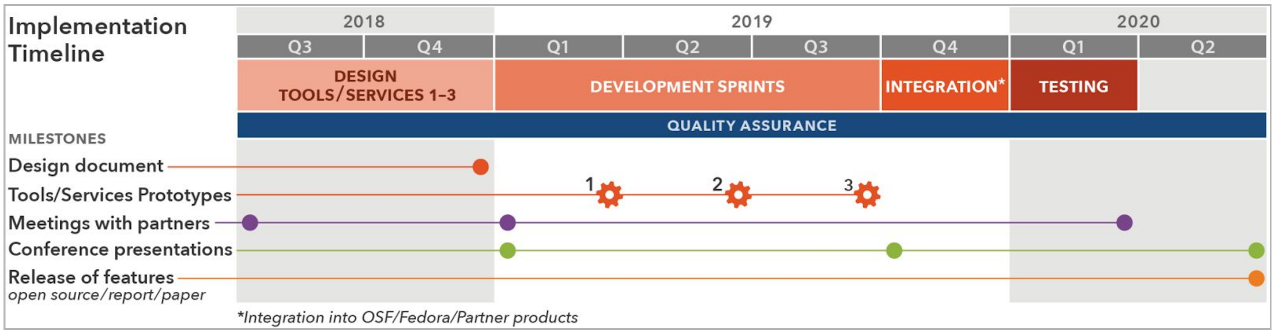
platform enabled through collaborations with Midwest Big Data Hub, NCSA, and NDS, and improve interoperability with international platforms like those at CERN.

Leadership: Zheng (John) Wang, AUL, Richard Johnson, Co-Program Director, Digital Initiatives and Scholarship, and Natalie Meyers, E-Research Librarian all from the Hesburgh Libraries and Sandra Gesing, Computational Scientist (CRC) will serve as Co-PIs to lead the activities of dedicated project personnel, collaborators and stakeholders. Gesing will be responsible for supervising technical implementation. Dedicated partner participation from COS will focus on interoperability and reproducibility.

Work Plan: The proposed work takes up where the previous planning grant period ends. In the implementation phase as shown below our workplan shifts to focus on collaborative development, and then to interoperability testing and usability analysis with stakeholders.



Projected performance goals and outcomes: Implementation of the detailed technical and administrative project plans created in the planning phase above are the performance goals of the currently proposed project. The tools and RESTful services will be thoroughly designed, implemented in development sprints and carefully tested with collaborators to assure their extendibility and interoperability (see below).



Budget Summary: We request $570,544 to support software development and interoperability testing conducted by the lead organization and named project partner(s). Stakeholder engagement is a cornerstone of the project's success thus far. Requested meeting, travel and participant support funds will be used judiciously to provide modest travel support and meeting opportunities to facilitate continued collaborator/stakeholder engagement ensuring the valuable input of our named partners (Johns Hopkins University Sheridan Libraries, NYU Libraries, Purdue University Libraries, and UC San Diego Library) and organizational participants (CERN, DCC, DCN, Midwest Big Data Hub, NCSA, NDS, RDA, SGCI, SHARE, and SPN).

Requested IMLS Funding	**YR 1**	**YR 2**	**570,544**
Labor (Software, Lead & QA)	168,574	240,460	409,034
Participant Travel	9,000	9,000	18,000
Meeting Expenses	8,000	8,000	16,000
Indirect Costs 30%	52,972	74,538	127,510
Cost Share			***570,555***
ND Research Cost Share Commitment	71,610	74,912	146,522
NDS Integration support		8,724	8,724
ND Labor Cost Share	151,227	132,508	283,735
ND Indirect Costs 30%	66,731	64,843	131,574