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

Brief Project Description: 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. The project executes 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 partner participation from academic libraries at small, medium and large institutions along with collaborating stakeholder organizations will develop open source data and software preservation tools and do interoperability testing with stakeholders focusing on the top three services and features identified as priorities during the PresQT planning effort. The project will develop tools and RESTful services, which address the identified needs in the PresQT planning grant and improve reuse of preserved data and software in institutional repository systems. The funding will make such research data more interoperable with repositories and Emulation as a Service (EaaS) environments following a standardized web-based approach. Project participants from Johns Hopkins University Sheridan Libraries, UC San Diego Library, Yale University Libraries, Purdue University's HUBzero Project, and National Data Service (NDS) will participate as collaborative developers on interoperability testing, and/or as usability stakeholders under subaward. Additional organizational partners include Amherst College, CalPoly's Project Jupyter, Center for Open Science, CERN, Confederation of Open Access Repositories (COAR), Fontbonne University, NYU Libraries/ReproZip Project, Exposing Data Management Plans and Metadata Interest Groups at the Research Data Alliance(RDA), Science Gateways Community Institute (SGCI), Tuskegee University, and West Big Data Innovation Hub. These organizations will contribute voluntarily as advisors, to the design, and to integration testing.

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

Statement of National Need

As computational resources become increasingly available and powerful, today's researchers can now explore and analyze hypotheses before touching a physical lab. The interactions of environmental conditions with biological organisms can be simulated to forecast possible events. Chemical compounds can be virtually created, their introduction into new environments can be simulated and evaluated for effectiveness or potentially negative outcomes. Such simulations can forecast the spread of vector borne diseases, or the effectiveness of using waste eating bacteria in the event of an oil spill. Many government agencies and funders have mandated data sharing for their funded projects and scientific progress based on the promise that life-saving scientific progress could advance more quickly with open data and interoperable systems. Yet even as scientific research is increasingly born digital, preserving and openly sharing such research is a challenging mandate because of the myriad of ways researchers can use computational resources to simulate and analyze dizzying arrays of possible scenarios in each experiment or study.

Reproducing such experiments, or even accessing their associated datasets, often depends on availability of compatible software, workflow tools and computational environments. For even the most cooperative researchers these dependencies present a barrier to sharing and data re-use, as preserving such environments and their output data for re-use can be extremely labor intensive. Funder and publisher mandates for data and software sharing can overwhelm scientific domain experts that are often not specialists in data and software curation. Curators engaged near the end of the research life cycle often receive incomplete metadata, at-risk formats, and a paucity of data documentation. Furthermore, when scientists reach out too late for preservation assistance, important intermediate data is sometimes no longer available for documentation or preservation. This subsequently impedes comprehensive data sharing. Additionally, preservation tasks are typically more labor intensive when not integrated at earlier stages of the project life cycle. Reuse and reproducibility are jeopardized in such cases. Tools that make shared scientific results more reproducible need to better handle complex data, workflows and software so that data becomes more readily re-usable. In *Self-Correction in Science at Work*, the authors emphasize that "Leaders in the research community are responsible for ensuring that management systems keep pace with revolutions in research capacity and methods."[1] As data sharing policies at funders and agencies mature, so too should preservation systems and techniques likewise evolve.

These reproducibility challenges prompted and were confirmed by our recent planning grant effort, Research Data and Software Preservation Quality Tool (PresQT) funded by IMLS award LG-72-16-0122-16. We surveyed a wide swath of the community and assembled several library and other domain experts, who are actively working to address these challenges, in two community workshops to determine highest priority needs that could be addressed now. Over 1,700 researchers, tool developers and platform providers responded to the widely circulated *PresQT Needs Assessment*[2]. Further input from PresQT Workshop One held in May 2017[3], and the second PresQT Workshop Sept 18, 2017[4] concluded a final round of scheduled community input. Our analysis of the needs assessment survey data show a priority order of interest in better: Provenance tools, Workflow Preservation & Re-use, Fixity tools, Metadata Completeness & Preservation Quality Assessment tools, Keyword Assignment, Profile-based Recommenders, and Data De-Identification tools.

Beginning to address these needs may seem daunting. However, the research and library communities are making progress as demonstrated through emerging tools and approaches shared in our community workshops.

---

[1] Bruce Alberts et al. 2015. Self-correction in science at work. *Science* Jun 2015: V 348, Issue 6242 DOI:http://dx.doi.org/10.1126/science.aab3847

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

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

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

The greatest barrier to success in libraries meeting researcher needs for data sharing may not be tool availability but the disconnected nature of our library, computational science, and research communities and asynchronous tool adoption timelines across institutions. Tighter connections and integrations between available tools and communities can bridge the gap between tool availability, researcher need, and data re-usability. Therefore, the PresQT project plan outlines a strategy to connect both existing and new services together as a foundational architecture that can be extended in the future beyond our grant period by project partners and other community members.

Related Work: In *Data and Software Preservation for Open Science* (DASPOS)[5], an NSF funded collaboration of High Energy Physicists, Computer Scientists and data librarians examined the key technical problems that must be solved to provide appropriate data, software and algorithmic preservation of the massive data sets accumulated by High Energy Physics (HEP) experiments. Many of these data are unique and represent an irreplaceable resource for potential future studies. All of our core team participated in DASPOS as either senior personnel or workshop participants, and while the archiving of HEP data requires some HEP-specific technical solutions, because DASPOS was funded to look at solutions that would be useful across many different disciplines, that work, especially related to software containerization approaches for preservation influenced the genesis of the PresQT planning project. Approaches to software preservation first explored in DASPOS were further explained, shared and demonstrated by physicists, computational scientists and data curators during the subsequent PresQT workshops.

The Software Preservation Network (SPN)[6] likewise coordinates software preservation efforts to ensure the long-term access to software. They connect and engage the legal, public policy, social science, natural science, information and communication technology, and cultural heritage preservation communities that create and use software. Participants in SPN and Data Curation Network (DCN) through PresQT workshop participation have further informed the PresQT approach to preservation quality tools. The PresQT plan here takes into account diverse repositories' needs and usefulness of preservation tools such as BagIt[7], Fedora[8], ReproZip[9], Jupyter[10], Open Science Framework (OSF)[11], and SHARE[12] as well as dashboards developed by the project WholeTale[13]. The approach of PresQT is to be repository and technology agnostic with a goal of widest interoperability and because scientists use diverse tools for their research we have also partnered with scientific software, workflow tool and science gateway providers such as HUBzero[14] to understand the available preservation features as well as the gaps in such solutions. Propopents, developers and experts using all the above have contributed to the PresQT workshops and our resultant project plan.

Through the above, as well as with regional, national, and international communities like the National Data Service (NDS)[15], Science Gateways Community Institute (SGCI)[16], West Big Data Innovation Hub[17] and Research Data Alliance (RDA)[18] PresQT engages with related research, diverse organization types, and community stakeholders.

---

[5] http://daspos.org/

[6] http://www.softwarepreservationnetwork.org/

[7] http://dataconservancy.github.io/dc-packaging-spec/

[8] http://www.duraspace.org/organization/fedora-commons

[9] https://www.reprozip.org/

[10] http://jupyter.org/

[11] https://osf.io/

[12] http://www.share-research.org/

[13] http://wholetale.org/

[14] https://www.rcac.purdue.edu/services/hubzero/

[15] http://www.nationaldataservice.org/

[16] https://sciencegateways.org/

[17] http://westbigdatahub.org/

[18] https://www.rd-alliance.org/

Project Design

Goals, projected outcomes, and assumptions: The overall goals of the project target libraries and the research community with consideration of stakeholders such as publishers and data curators: 1) One goal is the wide integration of standardized services for assuring preservation quality in existing preservation ecosystems or lower the hurdle to build an intuitive preservation ecosystem where it is not available yet; 2) A second goal is to increase the use of preservation tools during the life cycle of research projects. 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. The outcome will be RESTful web services, which provide interoperability between computer systems on the Internet and elicit a response that may be in XML, HTML, JSON or some other defined format. Additionally, these web services will be integrated in existing preservation systems such as Fedora, OSF and SHARE and researchers' science gateways, e.g., HUBzero. Thus, the whole life cycle from the researchers' use of data and software for analysing data and creating results can be more seamlessly preserved without the need for the researcher to so often switch away from their computational working environments. The proposed work takes up where the previous planning grant period ends. In the implementation phase as shown below our work plan shifts to focus on collaborative development, and then to interoperability testing and usability analysis with stakeholders (see Fig. 1).

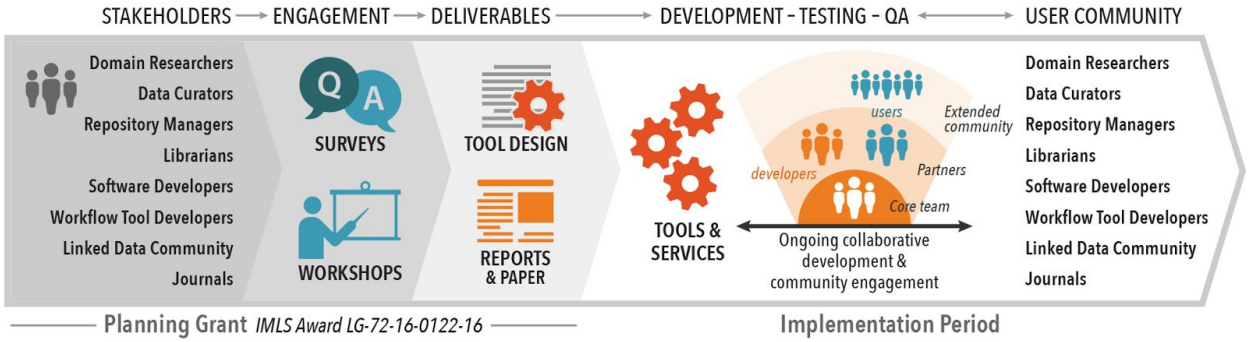


Figure 1: Proposed work in context with the PresQT planning grant.

The project seeks $569,298 to support community open source development and interoperability testing with stakeholders of the top 3 services and features identified as priorities during the PresQT planning effort. The requested funding will enable development of tools and RESTful services, which address the identified needs in the PresQT planning grant and improve reuse of preserved data and software in institutional repository systems. The funding will make such research data more interoperable with repositories like OSF, Science gateway tool HUBzero, and in Emulation as a Service (EaaS) environments like those being developed by project partner, Yale. Project participants from Johns Hopkins University Sheridan Libraries, UC San Diego Library, Yale University Libraries, Purdue University's HUBzero Project, and National Data Service (NDS) will participate as collaborative developers on interoperability testing, and/or as usability stakeholders under subaward. Hesburgh Libraries and the Center for Research Computing (CRC) at University of Notre Dame and subawardees will contribute 1:1 cost-sharing resources including project management, other personnel time, and fringe benefits ($569,298).

Additional organizational partners include Amherst College, CalPoly's Project Jupyter, Center for Open Science, CERN, Confederation of Open Access Repositories (COAR), Fontbonne University, NYU Libraries/ReproZip Project, RDA Metadata IG, Science Gateways Community Institute (SGCI), Tuskegee University, West Big Data Innovation Hub. These organizations will contribute voluntarily as advisors, to the design, and to integration testing (See Table 1). The testing will take place within a digital ecosystem of Fedora, HUBzero, NDS, OSF, and with Emulation as Service solutions being developed by collaborators at Yale.

Project responsibilities: Zheng (John) Wang, Associate University Librarian, Digital Access, Resources and Information Technology will serve as Project Director supported by Richard Johnson, Co-Director of Digital Initiatives and Scholarship & Natalie K. Meyers, E-Research Librarian, Hesburgh Libraries, University of Notre Dame (ND) and lead the activities of dedicated project personnel, collaborators and stakeholders. The Project Director will oversee all aspects of the grant work and coordination between partners, including research, technical work, and communication activities in the broader community. Sandra Gesing, Ph.D. Computational Scientist, CRC, ND will lead the design and implementation of the RESTful web services. She will oversee their integration and testing to assure the quality of the technical outcomes of the project.

Personnel roles, competencies, expertise:

- **John Wang**, Project Director is a librarian and a technologist. Wang oversees three programs in his role as AUL at Hesburgh Libraries: 1)Information Technology and Discovery Services, 2)Resource Acquisitions and Delivery Services, and 3)Digital initiatives and Scholarship. Wang publishes and lectures in the areas of web services, and digital assets management and aspires to seamlessly integrate library experience for users both at online and physical spaces. Previously, he served as Associate Director of Content Division and Director For Digital Assets Strategies at Emory University, and as Head of the Library Web Services at UCLA and was cofounder of the Simul8 Group, student participatory R&D group.
- **Sandra Gesing**, co-investigator and the project's technical lead has perennial experience with web-based development in academia and industry as developer, project leader and/or team leader. Her research interests include science gateways, workflows and distributed as well as parallel computing. She is heavily involved in SGCI, where her role focuses on outreach and community engagement. Gesing will communicate closely with SGCI stakeholders and partners contributing to PresQT. Gesing founded in 2009 the successful annual European workshop series IWSG (International Workshop on Science Gateways). She chairs and coordinates the IEEE technical area on science gateways. She also chairs other related international workshops, and has served on dozens of international conference and workshop program committees as well as on Technical Advisory Boards for Internet2 and NDS.
- **Rick Johnson**, co-investigator project work effort leader will collaborate with Gesing providing input on the tool design, knowledge of library focused data curation technologies, and outreach with repository collaborators. He has experience as Co-Director of Digital Initiatives and Scholarship at Hesburgh Libraries directing the design and development of data curation and digital library solutions. Johnson participates as a presenter at Code4Lib, CNI, Open Repositories, and DLF. He has contributed to DASPOS and the multi-institutional Hydra collaboration. Johnson steered development of an ORCID plugin for Hydra. Johnson also serves from 2015 as a Visiting Program Officer for the Association of Research Libraries' SHARE project..
- **Natalie Meyers** co-investigator and project work effort leader will collaborate with Gesing and perform outreach on use case tests with scientists, expert consultants, ReproZip, and emerging FAIR best practice groups in the library and research data management communities. Meyers is an E-Research librarian who pioneers and provides research data consulting services. Meyers devotes a significant part of her regular time as an embedded e-research librarian and has served as the Bill & Melinda Gates Foundation funded Vector Borne Disease Network(VecNet)[19] digital librarian, as a member of senior personnel on the NSF funded DASPOS project, an integrations manager at the Center for Open Science(COS), and in 2017-18 as Labs Partnerships and Project Manager in the research unit at the COS.
- **Miranda Van Nevel** Project Manager, Hesburgh Libraries, ND will serve as project manager and support the Project Director. She has 3 recent years of experience in research libraries doing project management for technical projects and 10+ years project management experience in industry.

[19] https://www.vecnet.org/

- **Donald Brower**, PhD Digital Library Infrastructure Lead, Hesburgh Libraries, ND will serve as Sr Web Developer and support the Quality Assurance effort. Brower has extensive experience in scientific, industry, and academic application development, recent library and scientific information systems experience and cyber-infrastructure.
- Additionally, two web developers from the CRC and a Quality Assurance specialist from the Hesburgh Libraries at ND are anticipated to contribute to the project. Working closely with research teams, CRC developers are responsive to the unique needs of research software development in an academic setting. They write well designed, testable, efficient code by using best software development practices.
- **Jeffrey Spies**, PhD will consult to the project and advise on service architecture, review specifications, provide input on test strategies, and assist in testing the project's restful services. The Open Science Framework is an extension of Spies' dissertation project. Spies is co-founder of the Center for Open Science and holds a Visiting Assistant Professor position in the Department of Engineering and Society at the University of Virginia's School of Engineering and Applied Science.
- **Elliot Metsger**, Senior Software Engineer, Digital Research and Curation Center (DRCC) at the Johns Hopkins University(JHU) Sheridan Libraries will serve as technical project lead for JHU. Metsger develops software for the Data Conservancy, an organization dedicated to digital archiving and preservation, and is active in the Fedora community. His current work focuses on digital packaging as an enabler of curation and provenance activities.
- **Christine Kirkpatrick**, Executive Director, National Data Service(NDS) will serve as Project administrator for NDS. Kirkpatrick is Division Director for IT Systems & Services at the San Diego Supercomputer Center (SDSC) at the University of California San Diego . Kirkpatrick also serves as deputy director and co-principal investigator of the National Science Foundation-funded West Big Data Innovation Hub.
- **Craig Willis**, Research Programmer, National Data Service(NDS) will serve as technical project lead for NDS. Willis has developed and implemented many data search applications in NDS. Willis is a Sr Research Programmer, NCSA University of Illinois at Urbana-Champaign and a Member of RDA Data Discovery Paradigms IG.
- **David Minor**, Director, Research Data Curation Program, UC San Diego(UCSD) Library will serve as project manager for UCSD. Minor's work on TritonSHARE and Chronopolis, a national-scale digital preservation network that originated with funds from the Library of Congress' NDIIPP Program, and Chronopolis' founding partner role in the Digital Preservation Network (DPN) all will contribute to the project's aims to interoperate with programs that are part of the national digital preservation agenda.
- **Michael Zentner**, Senior Research Scientist, Purdue University & Director at HUBzero Platform will serve as project administrator for HUBZero extending HUBzero's vision to serve communities in academia and industry and interface with preservation systems such as OSF. HUBzero is used as the data preservation platform for Purdue Libraries. Zentner has led many diverse software projects on the technology side in academia and industry and is the Co-PI on SGCI for the Incubator Service Area focusing on sustainability of software projects.
- **Euan Cochrane**, Digital Preservation Manager, Yale University Library will serve as Project administrator for Yale. Cochrane established digital preservation services through Preservica and Emulation as a Service for Yale. He has a particular interest and expertise in software preservation and the use of emulation to maintain access to born digital content.
- **Seth Anderson**, Software Preservation Program Manager, Yale University Library will serve as project lead for Yale. Seth has experience helping diverse orgs like HBO, Carnegie Hall, the Museum of Modern Art, and the Flemish Institute for Archiving (VIAA) with digital preservation and metadata creation, and at Yale develops digital preservation policies and practice for sustainable and scalable long-term preservation and access.

Technical design:
The PresQT tools and RESTful services will be thoroughly designed, following standard-based technologies in the context of accessing data storage and preservation systems. Design goals are:

1. Additional preservation features in tools, workflows and repositories should be easily integrable via standard APIs.
2. 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. In cooperation with stakeholders we plan interoperability testing and usability analysis.
3. The PresQT services will not be standalone solutions but the connection between tools, workflows and databases to existing repositories.

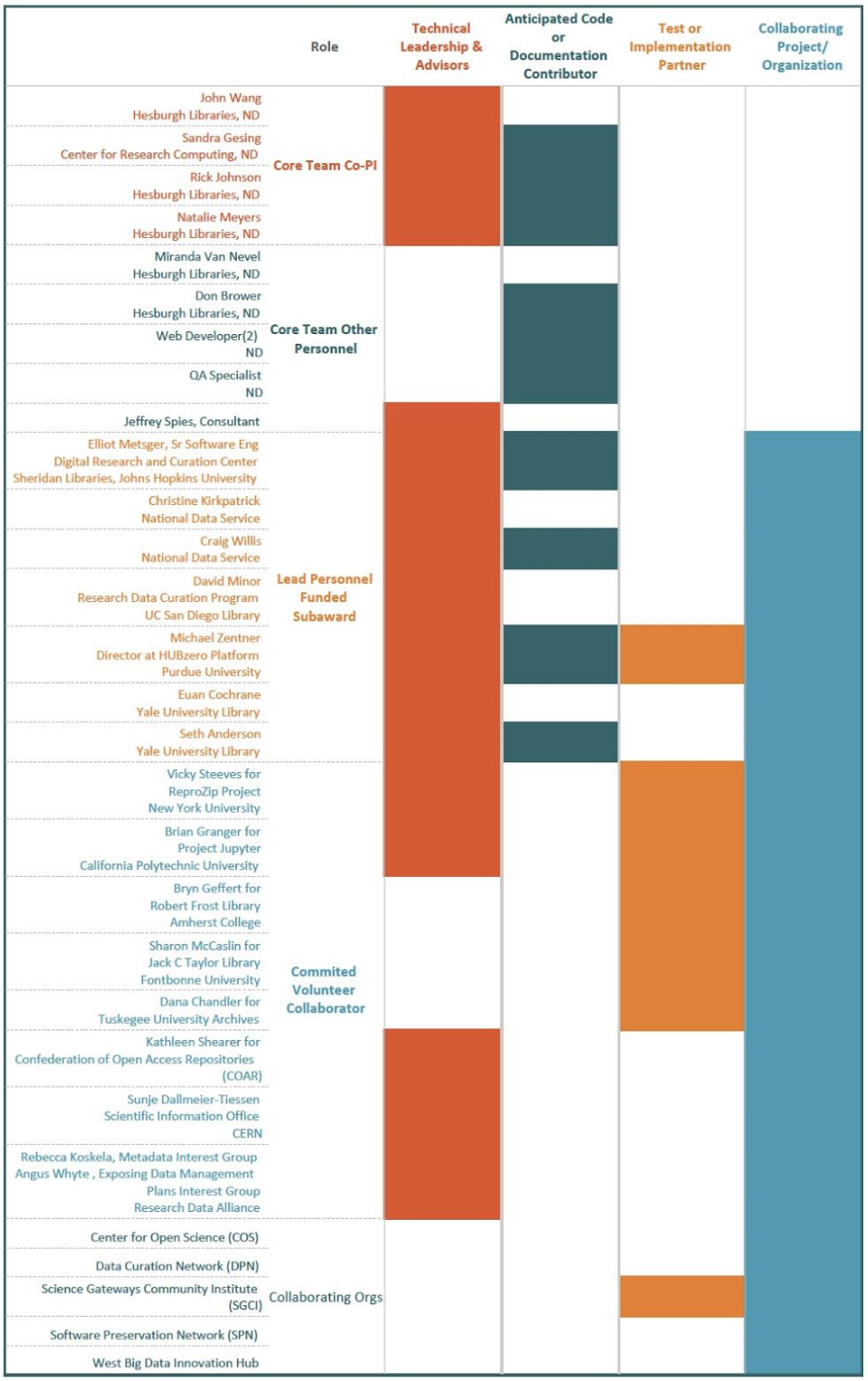


Table 1: Further Detail on PresQT roles and anticipated contributions to various phases of work effort

The PresQT planning effort ranked as most important: services providing the features provenance, workflows, preservation quality, fixity and keyword assignment. Since provenance and workflows are already well supported by existing systems, the design will focus on connecting such tools to repositories and extending the services with features for preservation quality, fixity and keyword assignment.

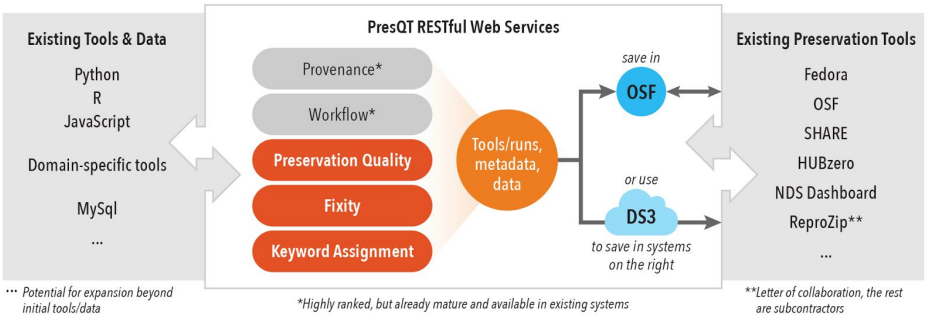


Figure 2: Technical Implementation Diagram

RESTful web services and APIs have evolved to a quasi-standard in web-based systems including preservation systems (e.g. in OSF). PresQT will implement RESTful services for preservation systems widely used in the community and by partners in the project. For example, DS3[20] APIs (Application Programming Interfaces) are offered in the Cloud via Amazon. DS3 offers a security model, data integrity, encryption support, partial data object restore, metadata per data object and infinite scale-out. Such characteristics are essential for the web services we aim to provide: ensuring scalability and supporting addition of vital metadata which can improve the quality of preserved data. Furthermore, at the same time supporting security requirements on the users' side so researchers can determine the granularity of data to share, targeting which groups access their data, or making it completely open access. The PresQT design document will include a best practices section on how to extend existing preservation systems and how to integrate the generally applicable RESTful services created during our development sprints. In addition to describing the services, PresQT aims to include, instructions and examples for using them, and detail the benefits of metadata enhancement during the preservation life cycle.

Project plan: Progress milestones and deliverables will be tracked to achieve the project's intended outcomes. We will use the SCRUM framework for efficiently execute the design and implementation phase. The SCRUM processes can be characterized in the following way:

- A product owner creates a prioritized wish list called a product backlog.
- During sprint planning, the team pulls a small chunk from the top of that wish list, a sprint backlog, and decides how to implement those pieces in text or software.
- The team has two weeks to complete its work, and the ScrumMaster keeps the team focused on its goal.
- At the end of the sprint, the work should be publishable.
- The sprint ends with a sprint review and retrospective.
- Next sprint begins, the team chooses another chunk of the product backlog and begins working again.

The design phase will be kicked-off via a meeting with all partners and result in agreed design documentation. Then development sprints will lead to three prototypes, which will be integrated and tested. The tests will be finalized in a meeting with all partners. The final outcomes of the project will be a release of features, tools and RESTful web services, including the documentation as open source in GitHub with all supporting project documents openly accessible in OSF.

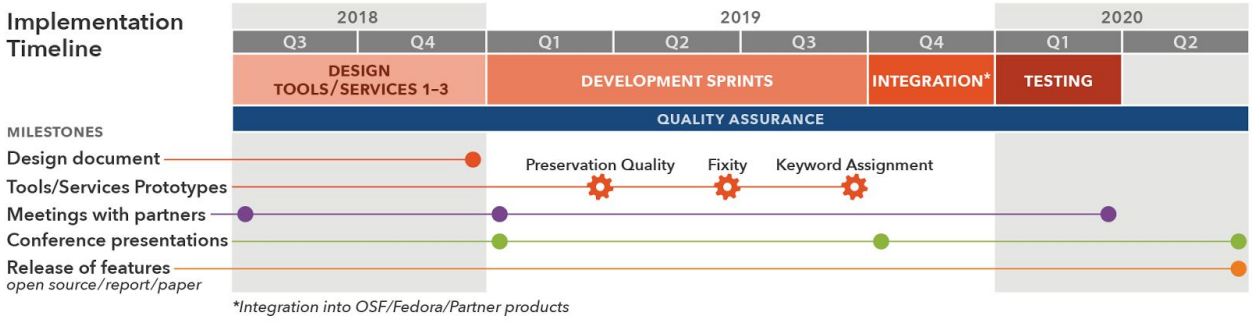


Figure 3: Timeline

[20] https://github.com/SpectraLogic/ds3_python_sdk

Project Communication: To evaluate PresQT progress and measure performance of the different project phases, we anticipate offering monthly calls from Q4/2018 open to the community and sprint reviews from Q1/2019 with bi-weekly demos of progress. The PresQT calls are planned to facilitate interaction in the community, to receive feedback on the design document, and serve as sprint reviews for the implementation, as well as to create a an open forum for questions, and facilitate bug tracking during the integration and testing phases. This community feedback will mitigate the risks of software development to ensure solutions are effective and appropriately targeted. The call content will be shared with the community as YouTube screencasts, a community engagement strategy successfully employed during a similar multi-partner Hydra-in-a-box grant funded project that engaged the community through demos[21]. Thorough documentation on how to use and apply the resulting tools and web services to an existing system will be created and distributed as part of the project.

Projected performance goals and deliverables: The following quantitative and qualitative goals and associated metrics will be used to track and measure our outcome in the different phases of the project.

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	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

Mitigating Potential risk in the work plan: A potential risk for the project could be that the resulting web services are not broadly used by preservation systems and in projects and do not meet the needs of the community. This risk is addressed via integrating a large number of diverse partners with different levels of involvement. We collaborate in a core team with subawardees who have clearly defined implementation goals. Additional partners have committed via collaboration letters to test, communicate and implement the results of the project (see Table 1). Naivete in design, or at the other end of the spectrum, overthought complexity in design, overwhelming implementation steps, or system requirements are examples of three risks further mitigated by our diversity of partners and communication plan. The quality and reasonableness of the PresQT design document, the implementation, integration and testing steps and protocols will be assured via dedicated reviews with a diverse group of technical advisors and partner test organizations (see Table 1). One more example of risk could be if implementation of a necessary web service cannot be accomplished in the given time. PresQT will use SCRUM project management to be quickly aware of blockers in the timeline so the project can address dependency delays accordingly.

[21] https://www.youtube.com/watch?v=kZqZ4vFTjy4

Project activities informed by appropriate theory and practice: The goals of PresQT project are motivated by preservation needs at ND and in the community, based on the findings of the PresQT planning grant, community needs and past project experience. Coordinators at the Hesburgh Libraries and the CRC have extensive combined experience in preservation projects and software development projects as made evident in accomplished projects such as VecNet, DASPOS, OpenMPS[22] and Whole Tale. The technical project design via the SCRUM framework is successfully applied at many Notre Dame CRC web-based agile software projects, e.g. the CRAFT repository[23], Paper Analytical Devices project for identifying suspected counterfeit drugs[24].

Sustainability Plan: We will sustain the work as open source project hope to extend the community built up during the two years of project time, during the preceding planning effort. Attention to stakeholder communication and creation of easy to follow documentation , as well as the YouTube videos can encourage a wider uptake at libraries beyond the funded project period as well as in systems adopted for use by scientific researchers. The public facing APIs and web service dependencies used at ND within the CurateND[25], Fedora and OSF ecosystem are anticipated to be maintained and supported by Duraspace, Notre Dame and COS respectively beyond the project duration. DPN, SPN, RDA, SGCI and NDS are all collaborating organizations similarly anticipated to continue after the PresQT implementation project funding ends which can provide forums for sustaining community engagement. For example, SGCI is funded via NSF and will continue until 2021 with the option to get funding for another five years. SGCI provides projects which may otherwise lack the financial resources or human resources to implement a science gateway or services for science gateways opportunities to apply for up to one year of developer time. Especially smaller institutions and those with repositories which have been not yet been extended for the use of the RESTful web services who are seeking help beyond the funded PresQT project period will be encouraged to apply at SGCI for support utilizing SGCI's short application of 3-4 paragraphs. To date SGCI has been able to support all requests suitable for the science gateway landscape and we are optimistic this responsiveness to community needs will continue.

Diversity Plan

In order to ensure a broad set of concerns are met through our effort, our plan presents a multi-pronged approach to inclusion of different communities with participants from academic libraries at large, small, and medium-sized institutions both public and private and a diverse set of organizational partner communities. Our group of small and medium-sized institutions include Fontbonne University, Amherst College, and Tuskegee University. Fontbonne University is a member of the Catholic Research Resources Alliance (CRRA[26]), and Tuskegee University is a member of the Historically Black Colleges and Universities (HBCUs). Our team of primary investigators is gender balanced and includes input from Notre Dame's Hesburgh Libraries and the university's Center for Research Computing. The composition of our key project staff furthermore reflects a gender balance and healthy community input from preservation, scientific, and library perspectives. tool and repository stakeholders.

Despite the existence of many components and services that facilitate higher quality preservation and sharing, the hitherto disconnected nature of many of these services has made adoption of both complex and more specialist solution(s) out of reach for many libraries, researchers,and digital archivists alike. Our planning grant surfaced many high priority gaps that still exist at both large and small institutions. Our design acknowledges these barriers to entry, facilitates tool adoption and data sharing between diverse and sometimes underserved communities by providing more formal well documented links between tools and services more often used

[22] https://mpsopendata.crc.nd.edu/
[23] https://www.darpa.mil/program/circuit-realization-at-faster-timescales
[24] https://padproject.nd.edu/
[25] https://curate.nd.edu/
[26] http://www.catholicresearch.net/

previously by only larger organizations or by disciplinary specialists in the relative isolation of their respective domains. This is not through a lack of library or researcher interest, as evidenced by the relatively recent establishment of communities like Research Data Alliance (RDA), and SGCI and their burgeoning growth. Furthermore, by targeting and collaborating with existing services and service providers openly offering tools of great value to the community, our project can subsequently increase the impact and adoptability of those enterprise-level services to less resourced institutions.

Our RESTful technology approach lowers the barrier significantly for such smaller less resourced institutions to access enterprise level services preservation, metadata, and sharing services. In order to ensure, we are actively considering the needs of a range of institutions we have engaged smaller to medium-sized partner institutions as testing and adoption partners (Tuskegee University, Fontbonne University, and Amherst College) alongside larger organizations such as Johns Hopkins University and UC San Diego who have greater experience and maturity as open source code contributing collaborators. Throughout our implementation phase, partners will share expertise through bi-weekly virtual demonstrations, and mutually benefit from regular updates as outlined in our project design.

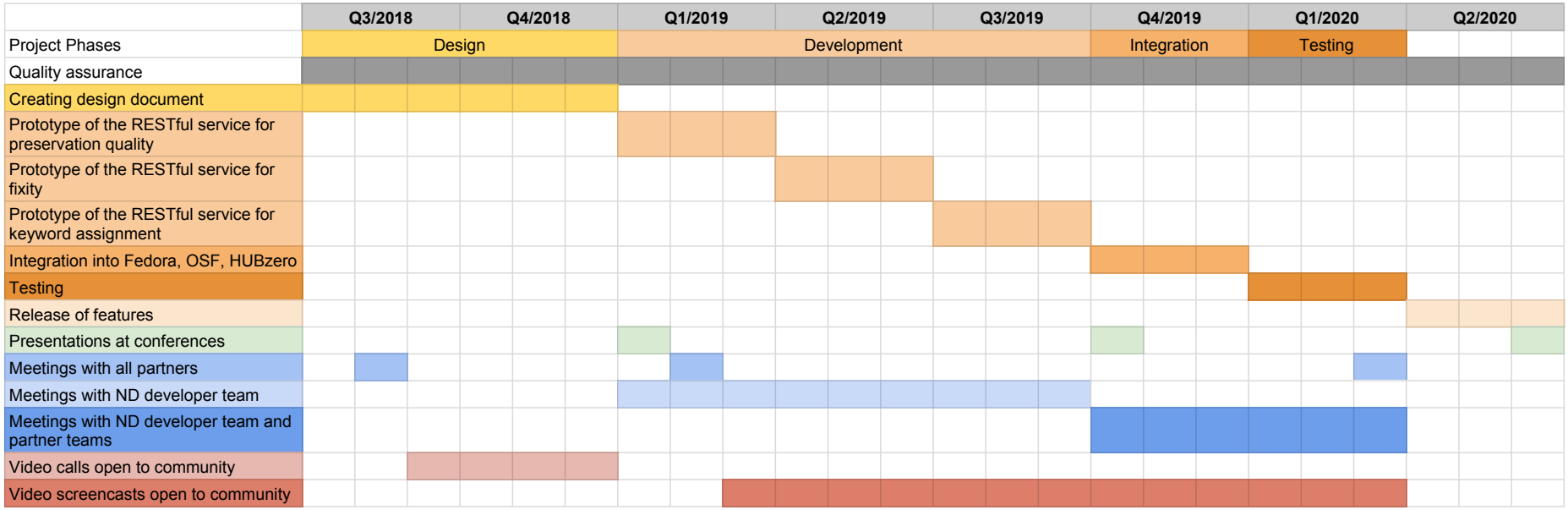
National Impact

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. 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[27], Data-Literature Interlinking Service[28]) will advance reproducible science and ease data re-use. The project will improve and support the national digital platform enabled through collaborations with Midwest Big Data Hub, NDS/NCSA, and improve interoperability of US repository systems nationally and abroad, as well as improve usability of US researchers' data on with international platforms like those at CERN.

In addition to the steps outlined in the Sustainability section of our Project Plan, to ensure increased impact beyond the project phase all PresQT funded software will be made openly available on Github. Project documentation and videos produced throughout the grant period will be shared on the Open Science Framework and YouTube to support continued adopters beyond the funded grant period. Core project participants likewise will continue to engage after the grant period with their associated collaborative communities of common interest such as DCN, RDA, SGCI, NDS, and SPN. This participation will include presentations at community conferences, workshops, and events thereby supporting continued adoption and contributing toward ongoing community interest in improvement of the delivered software tools. We anticipate opportunities to expand on the number of services available through the PresQT service framework through future grants aligned with the needs of software and data preservation communities and we hope also to further expand the number of integrations with active research tools through such opportunities. Finally, to support long term adoption by project partners, the University of Notre Dame is committed to documenting, maintaining and supporting the developed frameworks and services for user communities into the future.

[27] https://www.crossref.org/
[28] https://dliservice.research-infrastructures.eu/#/

ScheduleforCompletion



DIGITAL PRODUCT FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products can be challenging. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

- ☐ Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, you must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

Part I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

We will create software and share it openly on github repository (https://github.com/) with permissive CC0 licensing and/or Apache licensing that allows re-use. We will create a website that is openly accessible, we will continue to share and publish meeting minutes and a software design document on the PresQT Open Science Framework (OSF) project at https://osf.io/d3jx7/ DOI 10.17605/OSF.IO/D3JX7 as well as share linked YouTube videos. These resources will be all be openly licensed for re-use. The created software will be published in GitHub with Creative Commons licensing (CC0) that will be agreed with partners in the design document. CC0 is a well-known license that is designed to be recognized and useable in almost all countries around the world and may therefore be preferred over general statement of public domain dedications or the use of other licenses. We will attempt to publish in open journals, and to negotiate publisher's agreements for open access, when possible. We may be prohibited from openly disseminating information published in some peer-reviewed journals. Otherwise, as the publisher allows, we will make openly available a final, pre-publication version (aka preprint) of papers we write describing our approaches and results. Project Participants and journals may continue to assert copyright in publications and authors will be expected to exercise their right to openly share copies of their final pre-publication federally funded written work and encourage to sign agreements that specifically allow the manuscript to be deposited for public posting in open institutional and disciplinary open access repositories (such as CurateND, curate.nd.edu) as soon as possible after journal publication. As an example, the kind of language that an author or partner institution might add to a copyright agreement includes the following: "Journal acknowledges that Author retains the right to provide a copy of the final manuscript to [repository name, e.g. CurateND] upon acceptance for Journal publication or thereafter, for public archiving as soon as possible after publication by Journal. "

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

The new digital products will be openly accessible as public domain or under a Creative Commons license. We encourage

the re-use of the digital products and impose no restrictions on access. The only exception might be journal papers, which is dependent on the publisher of the journal (see above). But we aim at making a final, pre-publication version available if allowed by the publisher.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

We won't create any products that involve privacy concerns.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and format you will use.

We will create software shared on github, 1 design document, 40 meeting notes, 20 YouTube videos and at least 1 paper. The design document and meetings notes will be documents in OSF, the YouTube videos in MPEG and papers in PDF.

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

For the software we will use github for version control and software release and sharing, for meeting notes and the design document we need a web browser and Google doc, for the videos we will use QuickTime on MacBooks and papers can be created in Google doc or Microsoft Word and then saved as PDF.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

Software as ASCII text
MPEG: The recording resolution is dependent on the display of the computer on which the recording takes place. We assume to use our MacBooks with a resolution of 1440 x 900.
Word documents will be created in Microsoft Winword 2010 for Mac.
The exact PDF standards are dependent on the requirements of a publisher of a paper.

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

We will use SCRUM framework for the project management. While SCRUM was developed for agile software projects, it is also suitable for complex innovative projects. The different phases and products of the project are monitored via SCRUM processes. That is:

- A product owner (PI Johnson, PI Meyers or PI Gesing) creates a prioritized wish list called a product backlog.
- During sprint planning, the team pulls a small chunk from the top of that wish list, a sprint backlog, and decides how to implement those pieces in text or software.
- The team has two weeks to complete its work.
- Along the way, the ScrumMaster keeps the team focused on its goal.
- At the end of the sprint, the work should be publishable.
- The sprint ends with a sprint review and retrospective.
- As the next sprint begins, the team chooses another chunk of the product backlog and begins working again.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of

the federal award (see 2 C.F.R. § 200.461).

We will use github for maintaining software and OSF for public access to and data sharing of the design document, meeting notes, videos and paper(s). OSF is available for free to the public and allow for versioned updates of all digital content created via the project (for software see Part III below). For preservation of software and project documentation after the period of performance we will use CurateND.edu.

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

OSF uses Dublin Core and extensions to Dublin core. Although metadata is arguably not subject to copyright in the United States, to facilitate the use metadata across borders and avoid later arguments or confusion, we will make metadata available under a CC0 license. A simple policy statement effectively noting that to the extent that there is copyrightable expression, that copyright interest is being made fully available through a CC0 grant. Such a statement would acknowledge that there may not be any copyright protection in the metadata while simultaneously ensuring that any doubt is removed over whether such metadata may be shared or reused. It removes legal risk for those who are seeking to use the metadata, particularly because of the existence of a large number of institutional policies relying on a CC0 license for metadata. While the United States does not have any generic protection for databases, other countries do provide such protections. The European Union, for example, has a Database Directive, which provides for fifteen years of protection for computer records. In the EU, the protection provided under the Database Directive is separate from any copyright that may be granted. Russia similarly affords legal protections for databases. Even aside from such legal protections of databases, some countries may apply the “sweat of the brow” doctrine and afford copyright protection to metadata.

OSF uses and supports thesauri based on

- Taxonomy of Research Doctoral Programs from the National Academies
- Classification of Instructional Programs, 2010 edition, from the National Center for Educational Statistics (NCES)
- Medical Subject Headings (MeSH) from the National Library of Medicine
- Law subject headings are informed by the Current Index to Legal Periodicals (CILP) and FindLaw
- Several Business categories come from Cabell’s

This can be customized with institution specific terms if wished by partners.

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

We are using OSF for preserving data and metadata and this information will be maintained via the University of Notre Dame during and after the award period in our preservation repository CurateND.edu. We can transfer assets between OSF and CurateND using packages compliant with the bagit specification.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

The planned meetings, regular calls with the community and publication of results as well as the publication of the software in its different stages will assure a widespread discovery. The software stored in GitHub is usable via different mechanisms such as forking a project, downloading a project or using the GitHub RESTful API to work with a project.

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

The digital content will be publicly available via OSF. Users only need a web browser, a plugin for streaming MPEG videos and a PDF viewer.

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

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

Part III. Projects Developing Software

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

We intend to create RESTful web services for the major features for preservation quality, fixity and keyword assignment. These web services will be first be applicable via commandline in Python and R scripts and are based on the DS3 cloud web services and/or use the OSF commandline tools. The main functions are:

- authentication of users
- registering/creating/editing projects
- storing/editing data and metadata
- package tools via ReproZip
- storing/editing tools
- storing/editing runs

The integration into Fedora, OSF, SHARE, HUBzero and the NDS Dashboard are part of the integration phase of the project and assure that the web services are applied in already widely used preservation frameworks and software packages.
The choice of preservation frameworks and software packages addresses a wide range of audiences from librarians to publishers to researchers. The primary audiences are librarians and researchers who are supported in their daily workflow without the need to switch to different computing environments. Thus, librarians can use their preservation eco-system and researchers can use their tools such as a Python script extended via the web services to fulfill preservation needs of data and software storing information in the institution's preservation eco-system, for example Fedora, and/or in OSF.

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

The goal of our software is to be repository and technology agnostic and pluggable into existing preservation and research eco-systems. Thus, we take into account diverse repositories and preservation tools such as Fedora, OSF, SHARE and ReproZip as well as science gateways such as HUBzero and a dashboard developed by the National Data Service (NDS).
The Software Preservation Network (SPN) coordinates software preservation efforts to ensure the long-term access to software. They connect and engage the legal, public policy, social science, natural science, information and communication technology and cultural heritage preservation communities that create and use software. SPN is collaborating with us. Further collaborators include the Data Curation Network (DCN) Science Gateways Community Institute (SGCI), West Big Data Innovation Hub and Research Data Alliance (RDA). Via involving these different projects and communities, we assure to consider comprehensively diverse aspects in the preservation life cycle and used preservation tools.
The innovation is to extend existing preservation eco-systems with missing services for preservation quality, fixity and keyword assignment. While provenance and workflows are already offered maturely in quite a few preservation systems or workflow management systems, the additional features and

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

We will use Python for the development of the RESTful web services. Python has mature libraries for developing such web services. RESTful web services have evolved to a standard in web infrastructures and are widely used in preservation systems such as Fedora and in dashboards and science gateways. DS3 offers a security model, data integrity, encryption support, partial data object restore, metadata per data object and infinite scale-out. Such characteristics are essential for the web services we aim to provide: it ensures the scalability of the solution while adding vital metadata to increase the quality of preserved data and supporting security requirements on the users' side so that they can decide at which granularity they want to share data and with which groups or make it completely open accessible.
We chose for the integration with the RESTful web services several widely used preservation systems, science gateways and dashboards: Fedora, OSF, SHARE, HUBzero and the NDS dashboard to reach a wide community and serve librarians in their existing preservation eco-systems.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

We chose for the integration with the RESTful web services several widely used preservation systems, science gateways and dashboards: Fedora, OSF, SHARE, HUBzero and the NDS dashboard to reach a wide community and serve librarians in their existing preservation eco-systems.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

Our software is interoperable with several preservation systems. The software is repository and technology agnostic but for its use users need to know how to integrate RESTful web services and to apply a RESTful API.

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

We will use SCRUM framework for the project management. The different phases and products of the project are monitored via SCRUM processes. That is:

- A product owner (PI Gesing) creates a prioritized wish list called a product backlog.
- During sprint planning, the team pulls a small chunk from the top of that wish list, a sprint backlog, and decides how to implement those pieces in text or software.
- The team has two weeks to complete its work.
- Along the way, the ScrumMaster keeps the team focused on its goal.
- At the end of the sprint, the work should be publishable.
- The sprint ends with a sprint review and retrospective.
- As the next sprint begins, the team chooses another chunk of the product backlog and begins working again.

The code will be largely authored in python, shared openly on github.
User documentation will be also managed in github facilitated by Read the Docs documentation hosting service. (https://readthedocs.org/).

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

Diverse Science Gateways
https://crc.nd.edu/index.php/research/gateways

Whole Tale Dashboard
http://wholetale.org/

Diverse software projects of the CRC
https://crc.nd.edu/index.php/software

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you

intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

The created software will be published with Creative Commons licensing (CC0) that will be agreed with partners in the design document. CC0 is a well-known license that is designed to be recognized and useable in almost all countries around the world and may therefore be preferred over general statement of public domain dedications or the use of other licenses. GitHub won't create any cost for preserving the software and we will assure to keep the software up-to-date also after the project duration for our own preservation eco-system.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

The software will be published in GitHub in its different stages during the SCRUM process.

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository: GitHub

URL: https://github.com/

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

A.8 Identify where you will deposit the dataset(s):

Name of repository:

URL:

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?