Research Data & Software Preservation Quality Tool Planning Proposal

University of Notre Dame

Brief Project Description: We seek funding for stakeholder engagement and a collaborative planning effort to enhance reproducibility and more open sharing of research data through open source development of a *Research Data & Software Preservation Quality Tool.* Such a tool would provide for reuse of preserved software applications, improve technical infrastructure, and build on existing data preservation services.This tool will fill an essential niche in the technical stewardship portfolio, and its collaborative open source development will improve and support the national digital platform.

Statement of Need

Today's researchers can explore and analyze scenarios and explore hypotheses more quickly than ever before as computation is now interwoven with science. The creation of chemical compounds can be simulated before touching a physical lab. We can model the interaction of biological organisms to better forecast reaction to changes in environmental conditions. Disaster response tools and corresponding openly available data and models support saving lives and resources. The software, data, and platforms that are the part and parcel of such scientific endeavors can create efficiencies and foster rapid mutual progress when shared between scientists and information systems. Such promises of open data and interoperable systems have prompted many government agencies and funding bodies to mandate data sharing. However, as more and more scientific research is born digital utilizing complex computational resources that can simulate and analyze a dizzying array of possible scenarios, preserving and sharing research becomes an increasingly challenging effort. To reuse data, it is often necessary to have access to corresponding workflow, software, and complex computational environments that may have been custom built for a research project. Even with the most willing researchers, preparing such data for reuse can present a tremendous barrier to sharing. Depositing data can be quite labor intensive. Metadata enhancement, provenance reconstruction, reformatting and data documentation efforts can impede timely and complete data sharing. Researchers and their parent institutions often respond reluctantly or incompletely to the funder and publisher mandates for data and software sharing or are overwhelmed with the task being experts in their domain but not necessarily specialists for 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. Even the best data archiving and sharing methods can vary dramatically from lab to lab, from one institution to another, as well as between disciplines, countries and regions with their policies and mandates. Reuse and reproducibility are jeopardized in either case.

Research on counter-norms argues that more than goodwill is needed to shift practices to align more closely with reproducibility[1]. As research is increasingly born digital inside complex workflows and archived in a heterogeneous manner, it becomes imperative to better plan tools that can foster and facilitate researchers and repositories to utilize best practices and standards to preserve their data, software, and methods for better interoperability and re-use.

Today's scientists and scientific data curators face a challenge to enhance reproducibility and enable more open sharing of reusable research data. Recent attention to scientific reproducibility has increased awareness that many of today's experiments can not be easily reproduced. The Center for Open Science's Reproducibility Project: Psychology(RP:P) was a collaborative effort of 270 contributors to

[1] Anderson, Martinson, & DeVries, 2007

replicate 100 important findings in the psychology literature. RP:P published its findings in Science magazine on August 28, 2015[2]. After the results have been published, people from around the world engaged in conversations about the impact of this study on reproducibility and transparency. Those conversations and subsequent reproducibility studies have made it abundantly clear that the difficulty of reproducibility is not isolated to psychology. Monya Baker reported recently in *Nature* that "More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments" [3]. In computational science, workflows can be defined as *a sequence of connected steps in a defined order based on their control and data dependencies*. If suitable data is available, saved workflows should be a promising vehicle for propagating reusability of scientific methods and thus reproducibility. But, even in the case of something as tightly defined and purposed for reproducibility as workflows, a study of the team around the social marketplace MyExperiment for sharing Taverna workflows illustrated that as high as 80% of workflows may not be reproducible or reusable out of the box[4].

These reproducibility challenges above indicate that as data sharing mandates from funders and agencies mature, so too should preservation systems and techniques likewise evolve. 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."[5] Our proposed project personnel recently organized and hosted Container Strategies for Data & Software Preservation[6], a successful two-day Linux container centric workshop. This workshop sponsored by the NSF-funded Data and Software Preservation for Open Science (DASPOS) project[7] allowed participants to explore container solutions together. Presentations & discussions at the workshop indicate that interoperable software preservation tools are becoming mature enough to be better integrated with data sharing repositories which in turn can enable reproducible research. DASPOS is one of our early-committed planning project participants and brings the disciplinary perspective of high energy physics where data sharing requires preserving analysis alongside shared data. In the DASPOS commitment letter, Mike Hildreth, the Primary Investigator writes: "A tool like the one suggested in this proposal will be a vital ingredient to repository function and public access, and something like it was identified as a target for development and support as an integral part of the open data ecosystem."

In this spirit, our project complements existing repository infrastructure, aims to more deeply integrate workflow and software preservation tools and expands our own and/or our early-committed participants' previous work with an aim toward data preservation that facilitates scientific re-use and experimental reproducibility.

---

[2] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science (80-.).* 349, 6251 (August 2015), aac4716–aac4716. DOI:http://dx.doi.org/10.1126/science.aac4716

[3] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 7604 (May 2016), 452–454. DOI:http://dx.doi.org/10.1038/533452a

[4] Jun Zhao, Jose Manuel Gomez-Perez, Khalid Belhajjame, Graham Klyne, Esteban Garcia-Cuesta, Austin Garrido, Kristina Hettne, Maree Roos, David De Roure, and Carole Goble. Why workflows break understanding and combating decay in taverna workflows. In EScience (e-Science), 2012 IEEE 8th International Conference on, pages 1–9. IEEE, 2012.

[5] Bruce Alberts et al. 2015. Self-correction in science at work. *Science (80-.).* 348, 6242 (June 2015), 1420–22. DOI:http://dx.doi.org/10.1126/science.aab3847

[6] https://osf.io/y9mpx/

[7] https://daspos.crc.nd.edu/

Impact

The proposed project effectively addresses several timely data reuse issues and will have a lasting impact on the field by affording researchers and data curators with methods to better represent digital workflow methodologies, improve data and software provenance, automatically enhance metadata, perform schema validation, improve file format recognition, interoperability, data integrity and ultimately facilitate scientific reproducibility. Interoperability of the proposed data preservation and quality tool with existing platforms and solutions such as these aforementioned can improve 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 University of Notre Dame actively contributes to the Hydra Fedora Open Source repository platform code with which we've built CurateND, our institutional repository,[8] and the Vector-borne disease network digital library[9] as well as other collaborative tools that facilitate interoperability. Notre Dame's work with container and virtualization solution, Umbrella, provides an early proof of concept for bringing computing infrastructure closer to preserved objects in interoperable repositories and data commons[10]. Notre Dame has also successfully conducted a research project to integrate the National Data Service (NDS)' computational dashboard with the OSF, enabling researchers to push code and data files seamlessly from our IR or OSF storage to the NDS. We will soon embark on a project to more deeply integrate our IR with the OSF allowing users to push projects and files to our IR for long-term institutional preservation. In the process of developing the above solutions, we have sought software preservation solutions like container and emulation techniques to further ensure reproducibility and re-use.

The willingness of leading institutions and developers to engage in a collaborative planning effort for the proposed Data and Software Preservation Quality tool acknowledges that the time has come for next steps. The urgency with which stakeholders feel the need to address the challenges of reproducible science and the wide-ranging audience for such a proposed tool led us to seek this planning grant. Our proposed project design allows for input, consensus building, and buy-in from others regionally, domestically, and internationally in and/or outside the field. In addition to DASPOS, the Center for Open Science(COS) has pledged their willingness to participate as a dedicated project partner, and we have early commitments of participation from: the Scientific Information Service at CERN, The Research Data Alliance (RDA) Interest Group on Virtual Research Environments (VRE IG), RDA Interest Group on Metadata (Metadata IG), the Science Automation Technology Laboratory at the USC Information Sciences Institute, as well as Cal Poly and Project Jupyter (one of the most popular open-source software tools used for reproducible science). We are pleased to also have a pledge of participation from the Midwest Big Data Hub as well as the Confederation of Open Access Repositories (COAR), and from Michael Witt, as the Head of the Distributed Data Curation Center of the Purdue University Libraries, which includes his leading roles in the Data Curation Profiles, the Purdue

[8] https://curate.nd.edu/
[9] https://dl.vecnet.org/
[10] https://osf.io/s5e2b/

University Research Repository (PURR), DataCite, and the re3data.org registry of research data repositories. This project's potential to bring such domestic and international platforms, tools and experiences together gives us an opportunity to expand and improve digital content and services in the United States and abroad facilitating global scientific progress.

Projected performance goals and outcomes:

We will measure our performance by reporting on how well we are meeting our deadlines, workshop attendance count, and also by publicly sharing indicators like survey participation rate, as well as views and downloads through the metrics supplied on our project's site on the OSF where project resources will be shared. As our outreach expands, and our number of resources made available to the community increases, so too should our views and downloads. In addition to sharing our report and its administrative and technical project plans for implementing a Research Data Quality tool we will also summarize our findings in a paper we will present at conferences like Open Repositories 2017 and/or publish to a wider community through a journal like: the CODATA *Data Science Journal*, the *International Journal of Digital Curation(IJDC),* the *Journal of Librarianship and Scholarly Communication* and/or *Journal of eScience Librarianship* (JESLIB).

Our project targets for these performance measures will be a survey participation rate of >35%, workshop attendance of >65% of invited attendees with a desired participation count of 16-30 attendees at each workshop, increasing project site visits and page views from an estimate of 50/100 in the first quarter to 200/400 in the last quarters of the project. We would also expect to see and measure downloads increasing for our project report following attention garnered through conference presentations of our findings/and or publication of journal articles related to our project. We will measure success toward these targets through recordkeeping, survey and web analytics.

Tangible products expected to result from this project include**:**

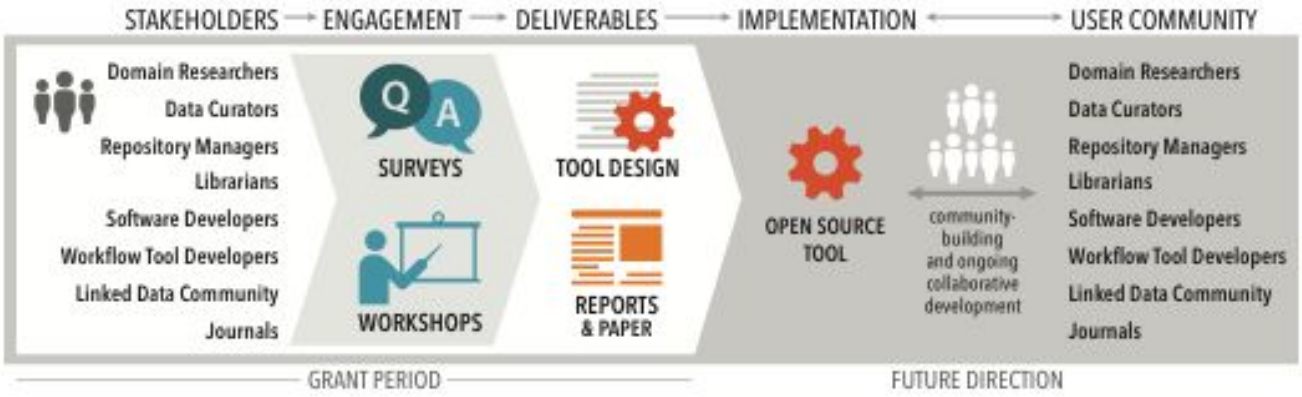
- A project web presence on the OSF
- A survey questionnaire & survey dataset
- Open workshop agendas, proceedings, and workshop reports
- A project final report
    - Administrative plan for bringing up a Data and Software Preservation Quality Tool
    - Technical Plan for implementing Data and Software Preservation Quality Tool
- One or more Open Access published papers

The outcomes of this project have the potential for successful, widespread adoption, integration, and adaptation. The resultant report and its administrative and technical plans will be of great benefit to multiple institutions and constituencies and actionable across a range of systems and funding levels. Our project should provide a tangible value to the library and data preservation fields through cost-savings achieved through collaboration on tool planning and interoperable open source design.

Project Design

Goal & Objectives: The project's goal and objectives are to develop technical and administrative implementation plans for an open source Data and Software Preservation Quality Tool that can be jointly developed by interested parties.

The activities to implement the project: Conduct stakeholder engagement through outreach to user communities and tool providers (survey, workshops, panels), conduct an evaluation of related systems, openly and transparently report all proceedings, collaboratively author and edit a report of findings, prepare technical and administrative plans.



Stakeholder Engagement: The collaborative planning effort will engage many stakeholders expecting the proposed to tool leverage existing platforms, solutions and linked data. We will seek multi-institutional expert opinion including journal, funder, software developer, repository manager, and data curators' perspectives. We will do outreach to stakeholders, conduct stakeholder panels at co-located conference opportunities, and convene stakeholders at two project workshops. We will work together with stakeholders to document interoperability opportunities and prioritize a feature set. Engaging other repository stakeholders and NDS developers during this planning grant will help us plan a tool that is interoperable and repository agnostic. Additionally, our stakeholder engagement will be informed by usability aspects. As mentioned above, the target user group of the resulting tool includes audiences who are not specialists in data and software preservation and thus, the design of the tool needs to reflect these users' needs, not only on a feature level but also on a usability level. In this context, we will be conducting our effort during the planning phase so that future tool development usability aspects will be compatible with standards such as CISU-R suggested by the Usability Group of NIST (National Institute of Standards and Technology)[11]. CISU-R defines three levels of compliance for usability.

> Level 1: Context of use must consider individually: stakeholders, intended user groups, the main goals for each group, the intended technical or computing environment, the intended physical or social environments, scenarios of use specifying tasks in context and any prerequisite documentation or training materials.

[11] http://zing.ncsl.nist.gov/iusr/documents/CISU-R-IR7432.pdf

Level 2: Measures must include: performance measures, e.g., achieving user goals and satisfaction measures via known questionnaires.

Level 3: The test method specifies how it is planned to evaluate that the requirements are met.

While we are concerned in the planning phase with the design of an intended tool and not the usability testing of an existing tool, the consideration of usability aspects and measures from the beginning will positively influence the design in a way that the different user groups are involved not only in feature requests but usability requests.

Survey: We will conduct a questionnaire based survey at the beginning of stakeholder engagement to gauge community priorities and expand our outreach (see *Questionnaire Brainstorming,* a sample question Appendix provided as SupportingDoc1.pdf). Information gathered through the proposed survey will be shared openly and inform our subsequent workshop focus areas.

Evaluation of Systems: We will ongoingly conduct an evaluation of systems related to our project topic, and do continuously outreach to constituent user communities and tool providers.

Panels and Workshops: During our outreach efforts we will convene panels, and hold workshops which will be preserved and shared transparently using OSF for Meetings.

Report: The output of our survey, evaluation of systems, panels and workshops will form the basis for a collaboratively authored planning report, which will identify priorities, potential roadblocks, and competing strategies. Our final report will cover:

- Identification of stakeholder prioritized requirements (must have, phased, nice to have)
- Identification of ways the tool improves preservation data quality and interoperability (potentially using format recognition, bit-level preservation, linking to format registries)
- Consideration of containerization & virtualization methods in context of data curation and interoperability with repositories
- Consideration of workflow tools & E-Notebooks to improve preservation and better enable reproducibility (Pegasus, Taverna, Jupyter/IPython)
- Consideration of metadata automation methods that enhance preservation and quality of the output of computational models and processes
- Consideration of linked data opportunities with data citation resources, format registries, authority files and PID systems

Technical and Administrative Project Plans: The report will be supplemented with detailed technical and administrative project plans. The technical and administrative project plans will be actionable by stakeholders along a continuum of need, capacity, and future funding scenarios.

The roles and commitments of partnering organizations: The Center for Open Science (COS) will be a dedicated partner organization, participating in monthly check-ins, attending, presenting and

participating in the workshops, and participating actively in the preparation of the final report. COS' role will be focused on reproducibility and interoperable data sharing aspects of the project as well as through provision and support of the project's use of the Open Science Framework (OSF) to store, share, and collaborate on project components. The OSF provides an application framework and platform for integrating and connecting services across the research lifecycle. Other organizations have agreed to be participating partners, committing to attend one or more workshops and contribute to collaboratively authoring workshop reports and the final project report.

The proposed project will directly and immediately benefit all stakeholders, all of whom will be able to access, further develop and take action on the project output at their own institutions and within their own disciplines. Interested stakeholders who self-identify as potential collaborators for jointly developing the proposed tool will be able to pursue joint open source development and funding to support such an effort with a plan that has been informed by an expert, well-informed constituent community.

Project Resources

Personnel:

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 convene workshops, engage with stakeholders, and lead authorship effort on the final report/project plan. The Hesburgh Libraries will contribute cost-sharing resources including project management, other personnel time, and fringe benefits. Sandra Gesing, Ph.D. Computational Scientist, Center for Research Computing, ND will be responsible for usability aspects of the project, technical outreach, and authoring of the Technical Project Plan. Gesing will integrate researchers from the beginning of the project ensuring the application of their input to the design, features and layout of the planning report and resulting platform. Richard Johnson will collaborate with Gesing providing input on the tool design, knowledge of library focused data curation technologies, and outreach with repository collaborators.

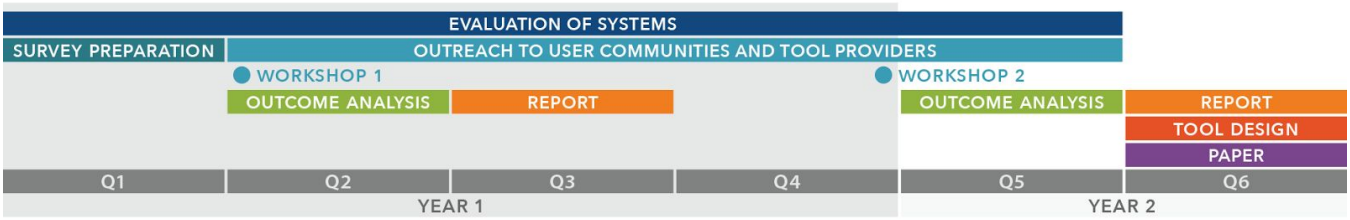
The proposed *Research Data & Software Preservation Quality Tool* will provide for reuse of preserved software applications, improve technical infrastructure, and build on existing data preservation services. Given our current interoperability efforts with National Data Service, Umbrella and Hydra/Fedora as described in our impact statement, our work plan and our partnerships with Center for Open Science and SHARE, we are in a unique position to lead a planning effort to co-develop and integrate tools or tool-suites that better represent digital workflow methodologies, improve data and software provenance, automatically enhance metadata, perform schema validation, improve file format recognition, interoperability, data integrity and ultimately facilitate reproducibility. This tool will fill an essential niche in the technical stewardship portfolio, and its collaborative open source development will improve and support the national digital platform.

Dedicated partner participation effort from the Center for Open Science (COS) will focus on data sharing and reproducibility. The project will reimburse COS for personnel costs to participate in

monthly check-ins, for stakeholder engagement travel, and for dedicated hours effort during review phase of the final report.

Time:

The timeline below depicts how the proposed project will be phased to maximize the opportunity for input from stakeholders and allow for a thorough technical evaluation of systems.



We propose to begin the project with a planning phase to include workshop planning, survey preparation, and evaluation of existing systems beginning no later than December 2016. We then propose at the start of the 1st quarter of 2017 to identify and broadly survey potential stakeholders regarding must-have, phased and nice-to-have feature requirements of the proposed data and software quality tool, to invite participation in coming project workshops, and begin outreach to user communities and tool providers.

A questionnaire will be developed by the project leads with input from early-committed stakeholders. The questionnaire will be administered through Qualtrics, a survey tool, to an audience of potential domestic and international respondents, including the 20,000 users who are part of the Science Gateway Institute's client base. Connection to the Science Gateway audience will be facilitated through the involvement of Dr. Gesing. Survey response data will be aggregated, analyzed and made available on the project website and inform discussion during workshops especially related to feature prioritization.

We will conduct the first stakeholder workshop of the project inviting those identified here in the workshop proposal, as well as others who self-nominate during the initial survey, seeking a broad range of input. The workshops will be designed to solicit participants' input and engage in collaborative planning. The goal will be to move toward defining a tool to fill gaps, better represent digital workflow methodologies, improve data and software provenance, automatically enhance metadata, perform schema validation, improve file format recognition, interoperability, data integrity and ultimately facilitate scientific reproducibility. The first workshop will focus more on "What" features are needed, the second workshop will be concerned more with "How" to create a feasible tool that can be widely used and accepted by the community. Workshop speakers will make short presentations, participate with attendees in facilitated round-table discussions and panels, and engage in brain-storming sessions that will engage stakeholders to share actionable input that can inform tool development. Participants will explore system architecture alternatives, user experience opportunities, and prioritize functional requirements for the proposed tool.

A report of the first workshop proceedings will be issued in the third quarter of the project and will inform planning for the second workshop.

Continued outreach to user communities and tool providers will be ongoing through the project's first year as will evaluation of existing systems.

The second workshop will be held toward the conclusion of the project's first year and/or beginning of the fifth quarter.

Collaborative authorship and editing of the final report, technical design, and any associated papers will take place in the sixth and final quarter.

Budget: We are seeking $45,000 to cover elements of time for dedicated personnel, convene workshops & panels, and meet with stakeholders. Stakeholder engagement will be essential to the project's success. Stakeholder engagement funds will be used to cover participant travel, accommodation, meals and meeting costs. Cost details are provided in our budget and explicated in our provided budget justification document.

Communications Plan:

Audiences: Gathering input from our early commitment named project participants, survey respondents, and other stakeholders and communities will be essential to conducting an informed, collaborative development effort. In addition to stakeholders named already, input from workflow tool developers like the Pegasus Project, VisTrails and/or the myGrid team alongside consideration of container and virtualization capabilities will help us provide for software reuse. Interaction with the Library of Congress' Sustainability of Digital Formats effort, the RMap Project, the Earth Science Information Partners (ESIP) and outreach through the Research Data Alliance (RDA)'s relevant topical Working and Interest Groups will help us explore linked data approaches and platform opportunities that can further enable interoperable software preservation. Planning ahead for interoperability with data citation platforms like DataCite, re3data.org, SHARE, Thomson Reuters' Data Citation Index (DCI), and others will be part of our strategy. Engagement with SHARE will further provide access to automatic and curation-based metadata enhancement tools.

Outreach: We plan to reach our audiences through organization listserv announcements, the project survey, the workshops, open sharing and promotion of the project proceedings and workshop reports through OSF during the project, as well as through the open publication of the final report, and one or more journal articles.

Community Building: The project promises to strengthen opportunities for collaborative software development, help like-minded organizations jointly develop solutions across national and disciplinary borders, and to help strengthen the research data repository community by providing a forum for discussion and action related to software preservation and reuse. There is potential for project participants to become more deeply engaged in Research Data Alliance Activities and vice versa, as well as for Scientific Research Information systems and tool developers to strengthen ties with academic library-grown repositories and approaches. There is potential for project participants to take the project's resultant technical and administrative project plans forward in the next phase to work together to jointly develop the proposed tool as a funded research project, an RDA working group activity.

Means to measure audience engagement and outcomes: We will rely primarily on survey response rates, workshop participation, and web analytics to measure audience engagement and outcomes.

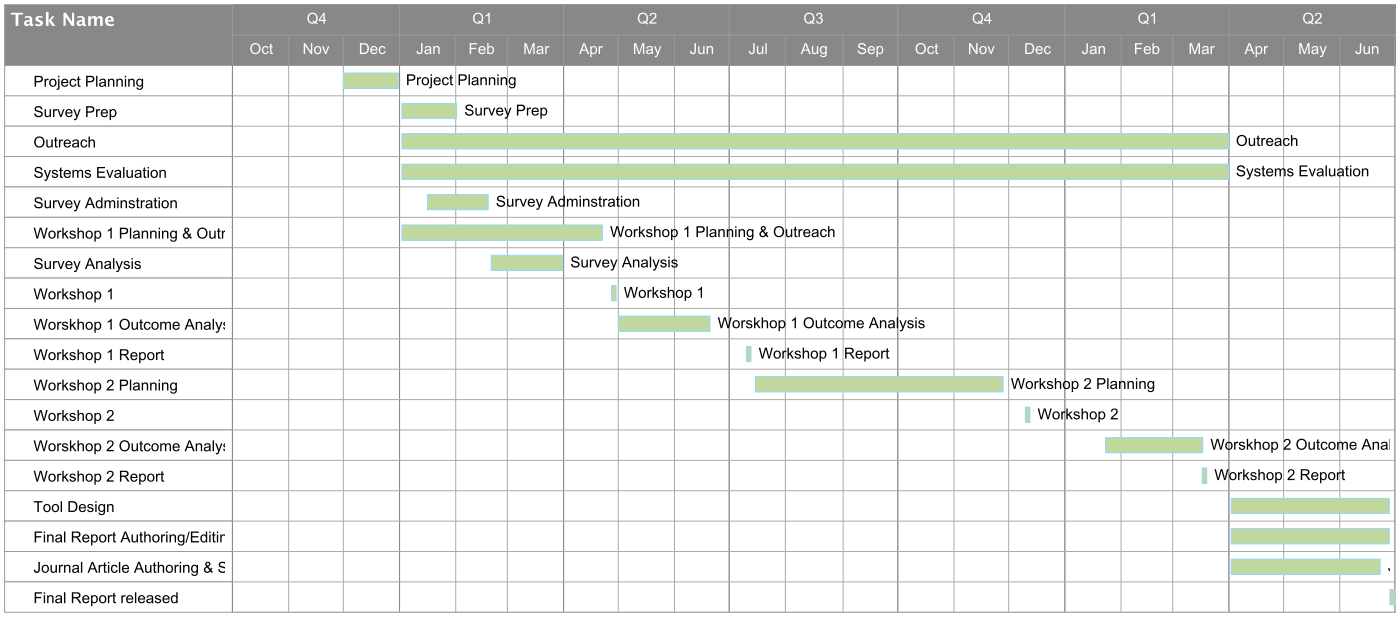
Staff assignments for outreach, promotion and dissemination: Project personnel led by Project Director John Wang will jointly pursue outreach during the early phase of the project as stakeholders are further identified, then support services from Notre Dame's Center for Social Research will be utilized to conduct the project survey in Qualtrics. Notre Dame's Center for Research Computing communications services will be utilized to support project personnel to plan, promote, organize and provide hospitality communications for the workshops. Project personnel, led by Project Director John Wang, will coordinate joint-authoring of workshop reports which will be shared openly on the OSF. Project Personnel, with Dr. Sandra Gesing in a lead role, will pursue evaluation of systems, preparation and collaborative authoring of the project's Technical and Administrative Project Plans. Project participants, with Project Director John Wang in a lead role, will collaboratively author the final project report with project personnel and dedicated personnel from the Center for Open Science serving as section editors.

Technical and Administrative Project Plans: The final report will be supplemented with detailed technical and administrative project plans. The technical and administrative project plans will be actionable by stakeholders along a continuum of need, capacity, and future funding scenarios.

Summary Statement

The proposed project's openly available reports, data collection, survey as well as workshop outputs will foster the sustainable design of the intended Data and Software Preservation Tool. Via the integration of the diverse stakeholders and user communities from the beginning, an open-source culture and community around the tool's adoption can be fostered, which can lower the hurdle for contributing researchers or developers' experience with open-source tools published in a finalized design. Thus, we envision that this planning grant has the potential to broaden adoption of such a tool for diverse frameworks. It will be designed in a way that it cannot only serve explicitly selected user communities and preservation tools but be modular, easily extensible and flexible enough that it fits a wide range of existing preservation tools and workflow systems.

Schedule of Completion



DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded research, data, software, and other digital products. The assets 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 is not always straightforward. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and best practices that could become quickly outdated. Instead, we ask that you answer a series of questions that address specific aspects of creating and managing digital assets. 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

If you propose to create any type of digital product as part of your project, complete this form. We define digital products very broadly. If you are developing anything through the use of information technology (e.g., digital collections, web resources, metadata, software, or data), you should complete this form.

Please indicate which of the following digital products you will create or collect during your project (Check all that apply):

	Every proposal creating a digital product should complete …	Part I
	If your project will create or collect …	**Then you should complete …**
☑	Digital content	Part II
☐	Software (systems, tools, apps, etc.)	Part III
☑	Dataset	Part IV

PART I.

A. Intellectual Property Rights and Permissions

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the intellectual property status of the content, software, or datasets you intend to create? Who will hold the copyright? Will you assign a Creative Commons license (http://us.creativecommons.org) to the content? If so, which license will it be? If it is software, what open source license will you use (e.g., BSD, GNU, MIT)? Explain and justify your licensing selections.

A.2 What ownership rights will your organization assert over the new digital content, software, or datasets and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users about relevant terms or conditions.

All digital content will be shared openly on the Open Science Framework for the benefit for the community at large. We will impose no conditions on access and use beyond those in the Creative Commons Attribution-ShareAlike 4.0 International License under which we will license the project content.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

The only content we will manage that may involve privacy concerns will be our survey data set. We will obtain rights from our survey participants to share the survey data under Creative Commons Attribution-ShareAlike 4.0 International License . We will not share personally identified physical address information if collected in the survey. We do intend to share identifying information at the organizaitonal/institutional level as collected by the survey with permission of respondents.

Part II: Projects Creating or Collecting Digital Content

A. Creating New Digital Content

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

We will create a Project web site on the Open Science Framework(OSF) at: https://osf.io/d3jx7
We will create a Survey questionnaire and collect survey data utilizing Qualtrics, an industry leader providing online survey software to which the University of Notre Dame has arranged a campus site-license for current ND faculty, staff and students which we will be able to use for the proposed project. We will share the survey data as standard ASCII csv delimited data .
We will collaboratively author project reports using GoogleApps like Google Docs and Google Sheets and share them on the OSF project site.

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

We anticipate using OSF, Qualtrics, and Google Apps as listed above. We may use Lucidcharts (https://www.lucidchart.com) and/or Balsamiq(https://balsamiq.com/).

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

txt, pdf, csv, png(72-200 dpi), jpg(72-200dpi)

B. Digital Workflow and Asset Maintenance/Preservation

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

N/A.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance (e.g., storage systems, shared repositories, technical documentation, migration planning, 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 CFR 200.461).

We will share the project proceedings, data, and reports on the Open Science Framework (OSF). We will archive the project and any pre-prints on CurateND, the University of Notre Dame's Institutional repository.

C. Metadata

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

For our survey data we will provide metadata and generate documentation that conforms with Data Documentation Initiative(DDI).

For our archived project data on CurateND we will use dublin core metadata standard.

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

Our metadata will be preserved on OSF and CurateND and maintained using each repository's metadata editing and management features .

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of digital content created during your project (e.g., an API (Application Programming Interface), contributions to the Digital Public Library of America (DPLA) or other digital platform, or other support to allow batch queries and retrieval of metadata).

Both OSF and CurateND have APIS that support batch queries and retrieval of metadata and both are indexed by SHARE .

D. Access and Use

D.1 Describe how you will make the digital content 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).

Digital content will be made accessible to the public on-goingly through the active life of the project on the OSF (https://osf.io/d3jx7) and will also be preserved and made accessible to the public at project's end on CurateND (https://curate.nd.edu/) Both systems are accessible via standard web browsers.

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

CurateND https://curate.nd.edu/
VecNet Digital Library https://dl.vecnet.org/
Container Strategies for Data and Software Preservation Workshop https://osf.io/y9mpx/
Technology Enhanced Research http://www.internationalinnovation.com/technology-enhanced-research/

Part III. Projects Creating Software (systems, tools, apps, etc.)

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) this software will serve.

N/A This is a Planning Grant no software will be created during this project.

A.2 List other existing software that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

N/A This is a Planning Grant no software will be created during this project. However our project will include an evaluation of existing systems to be included in our final report.

B. Technical Information

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

N/A

B.2 Describe how the intended software will extend or interoperate with other existing software.

N/A

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

N/A

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

N/A

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

N/A

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under an open-source license to maximize access and promote reuse. What ownership rights will your organization assert over the software created, and what conditions will you impose on the access and use of this product? 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 any prohibitive terms or conditions of use or access, explain why these terms or conditions are justifiable, and explain how you will notify potential users of the software or system.

N/A

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

N/A

C.3 Identify where you will be publicly depositing source code for the software developed:

N/A

Name of publicly accessible source code repository:
URL:

Part IV. Projects Creating a Dataset

1. Summarize the intended purpose of this data, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

Survey data will be collected in first quarter 2017 using Qualtrics to store respondents' answers to a questionnaire we will administer as part of stakeholder engagement.

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?

Our proposed survey and its data collection is related to systems not human subjects, therefore our survey is unlikely to need IRB approval, however we will be developing our questionnaire in consultation with Notre Dame's Center for Social Research. If CSR advise us on the basis of our questionnaire content that we need to to seek IRB approval, CSR will assist us in preparing our submission and we can submit online in Dec 2016 or Jan 2017 to get approval prior to conducting our survey in February 2017. The board meets monthly.

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).

No.

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.

We will collect speaker and particpant consent agreements for photography/videography during workshops . This activity and the records collected are managed and maintained by campus video services and the Center for Research Computing's communication specialists' team. We may collect consent/participation agreements digitally for sharing of our survey data using Qualtrics. We will maintain a record of consents collected and note that consents are on file in the metadata related to project files.

5. What 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).

Qualtrics.

6. What documentation (e.g., data documentation, codebooks, etc.) 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?

We will associate our survey questionnaire and documentation with the survey dataset using DDI.

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

Digital content will be made accessible to the public on-goingly through the active life of the project on the OSF (https://osf.io/d3jx7) and will also be preserved and made accessible to the public at project's end on CurateND (https://curate.nd.edu/) Both systems are accessible via standard web browsers.

8. Identify where you will be publicly depositing dataset(s):

Name of repository: Open Science Frameworks
URL: https://osf.io/d3jx7

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

We will review the data management plan during project planning and during q5 of the project.

Planning a Research Data & Software Preservation Quality Tool

Brief Project Description: We seek funding for stakeholder engagement and a collaborative planning effort to enhance reproducibility and more open sharing of research data through open source development of a *Research Data & Software Preservation Quality Tool.* Such a tool would provide for reuse of preserved software applications, improve technical infrastructure, and build on existing data preservation services. Given our current interoperability efforts with National Data Service, Umbrella and Hydra/Fedora as described in our impact statement, our work plan and our partnerships with Center for Open Science (COS) and SHARE, we are in a unique position to lead a planning effort to co-develop and integrate tools or tool-suites that better represent digital workflow methodologies, improve data and software provenance, automatically enhance metadata, perform schema validation, improve file format recognition, interoperability, data integrity and ultimately facilitate reproducibility. This tool will fill an essential niche in the technical stewardship portfolio and its collaborative open source development will improve and support the national digital platform.

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. Depositing data can be quite labor intensive. Metadata enhancement, provenance reconstruction, reformatting and data documentation efforts can present significant barriers to timely and complete data sharing. Curators engaged near the end of the research life cycle often receive incomplete metadata, at-risk formats, and a paucity of data documentation. Reuse and reproducibility are jeopardized in either case. Research on counter-norms argues that more than goodwill is needed to shift practices to align more closely with reproducibility[1]. The proposed effort bridges gaps we have encountered between existing digital library infrastructure, repositories and software reuse. ND actively contributes to the Hydra Fedora Open Source repository platform with which we've built CurateND[2] and the Vector-borne disease network digital library[3]. We have also integrated National Data Service (NDS) computational dashboard with COS's Open Science Framework (OSF). The OSF provides an application framework and platform for integrating and connecting services across the research life-cycle. Our recent work with container and virtualization solution, Umbrella, provides an early proof of concept for bringing compute infrastructure closer to preserved objects in interoperable repositories and data commons[4]. Interoperability of such a tool with existing platforms can improve 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.

Leadership and Budget: We are seeking $45,000 to cover time of dedicated personnel, convene workshops & panels, and meet with stakeholders. Stakeholder engagement will be essential to the project's success. Stakeholder engagement funds will be used to cover participant travel, accommodation, meals and meeting costs. Zheng (John) Wang, Associate University Librarian, Digital Access, Resources and Information Technology & Natalie K. Meyers, E-Research Librarian, Hesburgh Libraries, University of Notre Dame (ND) will serve as the Co-Project Directors and lead the activities

[1] Anderson, Martinson, & DeVries, 2007
[2] https://curate.nd.edu/
[3] https://dl.vecnet.org/
[4] https://osf.io/s5e2b/

of dedicated project personnel, collaborators and stakeholders. The Project Directors will convene workshops, engage with stakeholders, and lead authorship effort on the final report/project plan. The Hesburgh Libraries will contribute cost-sharing resources including project management, other personnel time, and fringe benefits. Sandra Gesing, PhD Computational Scientist, Center for Research Computing, ND will be responsible for usability aspects of the project and technical outreach. Gesing will integrate researchers from the beginning of the project ensuring the application of their input to the design, features and layout of the planning report and resulting platform. Dedicated partner participation from the Center for Open Science (COS) will focus on data sharing and reproducibility. The project will reimburse COS for personnel costs to participate in monthly check-ins, for stakeholder engagement travel, and for dedicated hours effort during review phase of the final report.

Work Plan: The collaborative planning effort will touch engage many stakeholders because the proposed tool leverages existing platforms, solutions and linked data. We will seek multi-institutional expert opinion including journal, funder, software developer, repository manager, and data curators' perspectives. We will convene stakeholders in panels at workshops to document interoperability opportunities and prioritize a feature set. Engaging other repository stakeholders and NDS developers during this planning grant will help us plan a tool that is interoperable and repository agnostic. Input from workflow tool developers like the myGrid team and consideration of container and virtualization capabilities will help us provide for software reuse. Library of Congress' Sustainability of Digital Formats effort, the RMap Project, the Research Data Alliance (RDA)'s Data Type Registries Working Group, and RDA's Research Data Provenance Working Group present linked data approaches and opportunities that can further enable interoperable software preservation. Interaction with data citation platforms like SHARE, DataCite, re3data.org, and Thomson Reuters' Data Citation Index (DCI), and others will be considered during planning. Engagement with SHARE also provides access to automatic and curation-based metadata enhancement tools. Gathering input from the above various stakeholders and communities will be essential to planning a collaborative development effort.

Projected performance goals and outcomes: Information gathered during panels, workshops and meetings will be preserved and shared transparently using OSF for Meetings. This input will form the basis for a collaboratively authored planning report which will identify priorities, potential roadblocks, competing strategies, and provide detailed technical and administrative project plans for creating a *Research Data & Software Preservation Quality Tool.* This grant will fund at a minimum:

- Identification of stakeholder prioritized requirements (must have, phased, nice to have)
- Identification of ways the tool improves preservation data quality and interoperability (potentially using format recognition, bit level preservation, linking to format registries)
- Consideration of containerization & virtualization methods in context of data curation and interoperability with repositories
- Consideration of workflow tools & E-Notebooks to improve preservation and better enable reproducibility (Pegasus, Taverna, Jupyter/IPython)
- Consideration of metadata automation methods that enhance preservation and quality of the output of computational models and processes
- Consideration of linked data opportunities with data citation resources, format registries, authority files and PID systems