Abstract

Many cultural heritage institutions have established repository systems to preserve the important digital content either generated or collected by their constituencies. An essential aspect of digital preservation is risk management. To aid in mitigating risk factors associated with lack of geographic diversity, lack of technological diversity, and loss of data related to human activities and systems failures, several distributed preservation services, such as The Digital Preservation Network (DPN), Chronopolis, and the Academic Preservation Trust (APTrust) have emerged. As these services have matured, the problem of tracking data from a local repository to a distributed preservation service has not been resolved. Working with an institution whose current repository is at different stages of development with regards to long-term curation of objects (Northwestern), an institution with a dedicated dark distributed repository that is not integrated with their digital library platform (UC San Diego), and an Advisory Board including the technical lead for Fedora, as well as lead members from DPN, AVPreserve, Archivematica, Hydra-in-a-box, and Chronopolis, this planning grant is designed to articulate the questions many institutions are grappling with surrounding the integration of local services with distributed preservation networks and identify broadly-applicable solutions and design patterns. There are a number of questions which we can immediately identify as part of what needs exploration, some of these are:

- How does one curate objects to ingest into a long-term dark preservation system?
- How does versioning of objects and metadata play out in a long-term dark preservation systems and how to automate these actions?
- How can systems that store data differently be made more interoperable?

To ensure the research is useful beyond the scope of the systems and institutions listed, a targeted group of diverse organizations that represent a variety of types of libraries, museums, archives, and preservation solutions will be surveyed. A subset of initial respondents will be interviewed more fully in order to produce user studies, and both sets of data will be analyzed for a final report. The intended audience consists of cultural heritage organizations who have committed to long-term preservation of their digital assets and wish to mitigate risk by distributing those assets into other preservation systems, as well as is the producers and developers of digital preservation services. Spanning one year in duration, this collaborative endeavor between Northwestern University Libraries and the University of California San Diego Library, will articulate questions many institutions are grappling with surrounding the integration of local services with distributed preservation networks, identify broadly-applicable solutions and design patterns, and propose high-level technical solutions. Ultimately, this report will serve as a guiding document for the technical integration of services, benefiting the producers of digital preservation services and institutions using their services.

Narrative

1. Statement of Need

The IMLS National Digital Platform aims to "bridge gaps between disparate pieces of the existing digital infrastructure, for increased efficiencies, cost savings, access, and services."[1] One such gap currently exists between local repository systems and dark, distributed preservation systems. Many cultural heritage institutions have established repository systems to preserve the important digital content either generated or collected by their constituencies. An essential aspect of digital preservation is risk management. To aid in mitigating risk factors associated with lack of geographic diversity, lack of technological diversity, and loss of data related to human activities and systems failures, several distributed preservation services, such as The Digital Preservation Network (DPN), Chronopolis, and the Academic Preservation Trust (APTrust) have emerged. As these services have matured, the problem of tracking data from a local repository to a distributed services has not been resolved. This planning grant, a collaborative endeavor between Northwestern University Libraries and the University of California San Diego Library, will articulate questions many institutions are grappling with surrounding the integration of local services with distributed preservation networks, identify broadly-applicable solutions and design patterns, and propose high-level technical solutions.

One intended audience for this planning grant consists of cultural heritage organizations who have committed to long-term preservation of their digital assets and wish to mitigate risk by distributing those assets into other preservation systems. This group faces numerous challenges, including but not limited to:

- Distributed digital preservation is costly and often a subset of the entire repository corpus must be selected. For example, DPN members can deposit up to 5 terabytes of data before incurring additional per-terabyte charges beyond their initial membership fee. This selection must be made prior to ingest into the distributed system. As an article about Chronopolis in 2010 duly noted "it is not within the purview of Chronopolis to define the usage needs and models of data users and researchers." (Minor 130)[2] How does an institution further select materials out of a collection of materials which have already been deemed as valuable?

- By their nature, dark distributed systems are not active; once data is deposited, updates and deletions should be infrequent. This can be problematic, given the tendency for digital collections, and especially the metadata describing them, to evolve over time. Part of the problem arises from the way digital collections or assets are acquired, as collections are often accumulations rather than fixed entities. Descriptive metadata changes happen at various points in time, and this dynamic content needs to fit within a logical and technology framework designed for managing static content. Versioning functionality may exist in local repository systems but does not necessarily translate to distributed systems.

[1] https://www.imls.gov/issues/national-issues/national-digital-platform

[2] Minor, D., Don Sutton ; Ardys Kozbial ; Brad Westbrook ; Michael Burek ; Michael Smorul. *International Journal of Digital Curation*, 01 July 2010, Vol.5(1), pp.119-133. http://ijdc.net/index.php/ijdc/article/view/150/212

- There is a dearth of models guiding the implementation and management of multiple copies in multiple systems. The Open Archival Information System (OAIS) is not explicit in detailing the functionality differences between local and distributed preservation systems. One extension of OAIS, the Outer OAIS - Inner OAIS model has been proposed, but remains untested.[3] The management of multiple copies in multiple systems is problematic in numerous ways: systems use different identifiers, local versions are more dynamic and mutable than distributed ones, staff turnover can affect management of distributed materials, etc.

- Lastly, in addition to difficulties associated with managing multiple copies in multiple systems, the actual storage of data in systems differs. For example, for managed content, Fedora 4 stores its binaries in an underlying database and pair-tree like file system structure, but distributed systems like Chronopolis and DPN repackage store data as files in the BagIt File Packaging Format.[4] What implications do these differences for the restoration of data into the original system?

This project intends to further articulate these challenges, identify additional ones through a survey and interviews. The second intended audience for this project is the producers and developers of digital preservation services. The project will benefit its intended audiences by identifying solutions and design patterns for increased interoperability between local repository systems and dark, distributed digital preservation systems. As data grows exponentially in local repositories, it's imperative that we find solutions to these challenges to decrease the risks of losing scholarship and other collections that may be valuable for future generations.

Distributed digital preservation systems are not a recent development; the Lots of Copies Keep Stuff Safe (LOCKSS) network was instituted at Stanford in the early 2000's.[5] The MetaArchive Cooperative, built on the same software as LOCKSS, was formed in 2004.[6] Since these initiatives were established, many institutions have stood up local implementations of institutional repositories. For example, as of last year, 1974 organizations or institutions worldwide have implemented DSpace and 331 have implemented Fedora.[7] And unlike LOCKSS-based systems, newer distributed digital preservations systems like DPN do not require the depositing institution to maintain a server to host content distributed throughout the network. Rather, the depositing institution submits the data and the service provider manages the technology stack. Working with an institution whose current repository is at different stages of development with regards to long-term curation of objects (Northwestern), an institution whose dedicated dark distributed repository is not integrated with their digital library platform (UCSD), the technical lead for Fedora, as well as lead members from DPN, Hydra-in-a-box, AVPreserve, Archivematica, and Chronopolis, this grant will complement and build upon work done in those communities by outlining integration patterns and potential solutions.

2. Impact

Distributed digital preservation is an essential element in an effective digital preservation plan. Materials stored in one location are vulnerable to natural or man-made disasters and can potentially be

[3] Zierau, Eld, and Nancy Y. McGovern. "Supporting the Analysis and Audit of Collaborative OAIS's Using an Outer OAIS-Inner OAIS (OO-IO) Model." In *Proceedings of the 11th International Conference on Digital Preservation*, 209–18. Melbourne, Australia: State Library of Victoria, 2014.http://ipres2014.org/sites/default/files/upload/iPres2014-Proceedings-version%201.pdf.

[4] https://tools.ietf.org/html/draft-kunze-bagit-13

[5] Dobson, C. "From Bright Ideas to Beta Test: The Story of LOCKSS." In *Searcher*, 11. February 2003, 50-53. http://www.lockss.org/locksswp/wp-content/uploads/2011/11/C.Dobson.pdf

[6] https://www.metaarchive.org/the-cooperative

[7] http://duraspace.org/sites/duraspace.org/files/2015%20DuraSpace%20Annual%20Report.pdf

undetectably tampered with. Distributing copies of digital materials over a network with fixity checking at each node mitigates those risks. Establishing the groundwork for increasing interoperability of local and distributed digital preservation systems, the main purpose of this planning grant, will impact the field by providing the foundational design documents from which technical solutions can be devised.

Likewise, users of digital content in the United States will benefit from greater interoperability of local and distributed preservation systems as it increases the likeliness that content will be preserved, and thus accessible, in the case of disaster. The primary product of this planning project is a report summarizing the key integration issues, drawing from interviews and the environmental scan, including user stories and design patterns which can be implemented in a variety of environments as well as recommendations for technical solutions. This report will serve as the guiding document for the technical integration of services, which will benefit the producers of digital preservation services.

Given that this is a planning project, focused on building the foundation for the development of future services, performance indicators are difficult to measure. The project team will rely on the Advisory Board, including digital preservation software system developers, to provide evaluation and feedback on the utility of the survey, survey results, recommendations, and final reports.

This proposed grant project is designed to maximize input and consensus building from a diverse constituency within the library and archive fields. Explained in more detail below, the project will be guided by an Advisory Board of experts to aid in the user study methodologies and final outcomes. Additionally, the project design includes user studies that will survey and interview a diverse set of cultural heritage organizations that represent a variety of types of libraries, museums, archives, and preservation solutions.

3. Project Design

To ensure our research is useful beyond the scope of the systems and institutions included on the project team, we will survey a targeted group of diverse organizations that represent a variety of types of libraries, museums, archives, and preservation solutions. Our Advisory Board is made up of vendors, consultants, and projects which all work with a large number of institutions struggling with these issues and will help ensure that our findings have broad applicability.

There are a number of questions, as identified above, which we can immediately identify as part of what needs exploration, this project is scoped to address the following:

- How does one curate objects to ingest into a long-term dark preservation system?
- How does versioning of objects and metadata play out in a long-term dark preservation systems and how to automate these actions?
- How can systems that store data differently be made more interoperable?

Activities:

1. Environmental/Feasibility Scan (Months 1 - 5)
    - Review planned or suggested development related to this topic.

- Creation and review of a survey and follow up interview questions, which will be vetted by a Library Assessment Specialist at Northwestern University Libraries. This survey will be disseminated to a variety of organizations.
    - We will be sure to make sure our questions address the questions above and allow us to:
        - Understand the breadth of implemented local systems.
        - Identify local workarounds and metadata fixes in place to address these issues.
        - Gather data about local preferences around versioning.
        - Identification of preservation policies and rights issues and how they are reflected in the technology itself.

2. User Studies (Months 6 - 10)
    - We will release the survey to institutions identified as being engaged in digital preservation activities. Our Advisory Board will help identify institutions. We will send out targeted emails to individuals at those institutions and strive for a 15% response rate.
    - Schedule and perform follow up interviews with a subset of the respondents.
    - Perform data analysis and identify patterns in the results.

3. Report (Months 11 - 12)
    - Creation of user stories.
    - Determine recommendations for technical solutions.
    - Write final recommendations report.

4. Meetings & Travel
    - Month 1
        - Key project staff initial kickoff and planning meeting in Chicago/Evanston, IL
    - Month 2
        - Virtual Advisory Board meeting
    - Month 5
        - In-person Advisory Board meeting with key project staff - Chicago/Evanston, IL
    - Month 7
        - Short meeting of key project staff and board members who are attending Open Repositories 2017
        - Set up in-person interviews with identified institutional representatives for in-depth interviews at Open Repositories
    - Month 11
        - In-person Advisory Board meeting with key project staff - San Diego, CA
        - Meeting of key project staff attending Preservation and Archiving Special Interest Group (PASIG) in Fall 2017
        - Possibility of presenting initial draft findings at PASIG in Fall 2017

5. Dissemination of Research Findings, post grant project period
    - Proposals to present our findings will be submitted to a number of conferences including:
        - Open Repositories (OR) 2018

    - Preservation and Archiving Special Interest Group (PASIG) 2018
    - Digital Library Federation (DLF) 2018
    - Coalition for Networked Information (CNI) Fall Meeting 2017
- The final report, as well as the scrubbed data and other related information, like the user stories created through this process, will be ingested into Northwestern University's Institutional Repository.

4. Diversity Plan [if applicable] N/A

5. Project Resources: Personnel, Time, Budget

This one year planning grant project is a collaboration among the Northwestern University Libraries and the University of California San Diego Library. The project will have one principal investigator, Evviva Weinraub, who is the Associate University Librarian and Portfolio Holder for Digital Strategies at Northwestern University Libraries. She currently serves at the Project Co-Director for the Avalon Media System Project, funded previously by IMLS and currently funded from the Mellon Foundation. Weinraub brings her expertise from holding IT Director positions at Tufts University and Oregon State University and from her previous role at Digital Preservation Network (DPN). At Northwestern, Evviva has broad responsibilities for development and implementation of technology services within the libraries primarily focused on digital collections and repositories, metadata and discovery services, and infrastructure and desktop computing. With Avalon, she works with her co-Director to establish overall project direction and priority setting. She has spoken about engagement with open source communities, building experiential learning opportunities, digital preservation, and technological innovation. Evviva, who is qualified to manage the budget for this planning grant in conjunction with the business office at the Northwestern Libraries, will establish project direction with the Advisory Board and organize those in person meetings as well as serve as a sounding board for the other key staff.

Much of the day-to-day work for this planning grant project will be performed by four other employees. Sibyl Schaefer, the Digital Preservation Analyst for Research Data Curation at UCSD Library and Chronopolis Program Manager, brings over 12 years of experience as a project manager, digital archivist, and digital preservation expert. She will dedicate time to all the key activities including the environmental scan, creation and administration of the survey, research into versioning mechanisms, participation in interviews, and will co-author the report. Carolyn Caizzi, Head of Repository and Digital Curation at Northwestern University Libraries, brings over 12 years of experience working on digitization and repository projects to this planning grant. She will primarily oversee the staff working on the project from Northwestern, coordinate the grant activities, and contribute to the report. The Digital Preservation Librarian position at Northwestern University Libraries will also contribute to the all the key activities and be a counterpart to Sibyl. Additionally, a senior digital repository software engineer at Northwestern University Libraries will participate as a technical resource; contributing to the environmental scan activities, survey creation, and interviews. This developer will help analyze data from the surveys and interviews and propose technical solutions for the report which will be vetted and worked through with the Advisory Board. To facilitate the collaboration needed for this project, the key staff will meet virtually every other week to coordinate work and report out.

The Advisory Board members were chosen for their expertise in digital preservation systems and services and in providing these to broader communities either through open source practices and/or through vendor relationships. This Board will provide overall direction for the project and interact with key staff through two in-person meetings and through two or three other virtual meetings depending on need. The Advisory Board will help identify and propose technical solutions for the report. The members identified are:

- Andrew Woods - Fedora/DuraSpace
- David Wilcox - Fedora/DuraSpace
- TBD - Hydra-in-a-Box representative
- Dave Pcolar - Chief Technical Officer, Digital Preservation Network (DPN)
- Michael Ritter - Developer Chronopolis & DPN
- Bertram Lyons - Senior Consultant, AVPreserve
- Justin Simpson - Director of Archivematica Technical Services

The success of this project will be ensured by the expertise of the key staff and Advisory Board, the funding from this grant for travel and the cost share from Northwestern University and University of California San Diego for the key staff time dedicated to the project. Northwestern's share of the proposed budget includes salaries and fringe for Weinraub, Caizzi, the Senior Developer, and the Digital Preservation Librarian position at 5, 5, 10, and 15 percent of their time respectively for the full year. UCSD's cost share of portion of the budget is 7% of Schaefer's time for the full year and space for one of the Advisory Board meetings. All staff identified have the support of their supervisors to work on this project. Northwestern will be providing access to virtual conference software and the survey tool as well as space for most of the person meetings. All of the items for the budget relate to travel for the Advisory Board plus team project meetings, two travel trips for the key project team, and two conference trips for two key staff to present the project updates and findings to the larger digital preservation community. The total requested budget from IMLS for this planning project is $49,114 and the total cost share from Northwestern is $62,561.

M=Month; A=Activities	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
A1: Environmental Scan reviewing planned or suggested development; creation of survey and interview questions												
A2: User Studies survey conducted; follow up interviews with subset; data analysis												
A3: Report Writing Identification of patterns, user stories, technical recommendations												
A4: Key project team in-person meeting												
A4: Advisory Board and key project team in-person meetings												
A5:Conferences for key project staff to conduct project work: Open Repositories and PASIG												

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.

The authors and/or their institutions will hold the copyright of the report. We will assign the attribution 4.0 international Creative Commons license to all outputs (report and scrubbed dataset) so that both vendors and the wider digital preservation community can use it. http://creativecommons.org/licenses/by/4.0/

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.

The final report, as well as the scrubbed data and other related information, like the user stories created through this process, will be ingested into Northwestern University's Institutional Repository.

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.

N/A.

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.

A final report, user stories and technical recommendations as well as scrubbed raw data from the survey and interviews.

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.

Qualtrics software for the survey and am Excel (or a csv or OOXML file) for the output of the dataset.
Basic word processing software for report.

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).

PDF for report; external scrubbed dataset in Excel (and/or csv or OOXML).

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 since the output is one report and one dataset.

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).

Deposit into Northwestern's Institutional Repository (the Hydra application, Sufia v.7). Assets in the repository are preserved and maintained in Fedora 4.

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).

Our metadata librarian will catalog the deposit into the Institutional Repository. The IR uses MODS as the descriptive metadata schema.

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

N/A since the output is one report and one dataset.

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).

N/A since the output is one report and one dataset. However, our Institutional Repository will be crawled by Google making the report easily discoverable.

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).

The report and scrubbed dataset will be available to the world in the Institutional Repository with an attribution international 4.0 Creative Commons license.

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

Our Institutional Repository is launching in June 2016 so we do not have an example for it yet. Northwestern currently uses Handle for persistent identifiers. An example is seen in the Audio Video Repository: http://hdl.handle.net/2166.MEDIA/j82k22d60z.

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.

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.

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.

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

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

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

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

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.

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

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

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.

The purpose of gathering data from a survey (designed administered through Qualtrics) and from one set of follow-up interviews during months 6-10 (May-September 2017) is: to understand the breadth of implemented local systems at academic libraries and other similar institutions; identify local workarounds and metadata fixes to address long term preservation in dark distributed systems; gather information about preferences around versioning; identifying how preservation policies and rights issues are reflected in the technologies institutions have implemented.

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?

The preliminary proposal was reviewed by the Northwestern University IRB and it was determined that this project does not need to be overseen by it.

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, but we will scrub the data so that institutions are not identified if they do not want to be.

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.

N/A.

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. The key project staff will consult with Northwestern Libraries' internal Library Assessment Specialist on how to create a good survey and on how to use the software.

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?

N/A.

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

The data will be deposited in the Northwestern Institutional Repository, which goes live in June 2016.

8. Identify where you will be publicly depositing dataset(s): Northwestern University Institutional Repository

Name of repository: coming June 2016
URL: coming June 2016

This Sufia based repository application is currently being implemented and will go live in June 2016. The branding for this repository and the URL are currently under review.

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

This data management plan will be reviewed in May 2016 and September 2016 in case adjustments are needed. Additionally, it will be reviewed in December 2016 to verify the deposits were made.

Original Preliminary Proposal

Title: Beyond the Repository: Integrating Local Preservation Systems with National Distribution Services

Type: Planning Grant (1 year // $50,000)

Abstract

Many institutions that are grappling with long-term preservation are navigating unchartered territory when it comes to integrating their repository systems where they are actively managing content with a dark, distributed, preservation system like Chronopolis or DPN. Working with an institution whose current repository is at different stages of development with long-term curation of objects (Northwestern), an institution with a dedicated dark distributed repository that is not integrated with their digital library platform (UCSD), the technical lead for Fedora, as well as lead members from DPN, Hydra-in-a-box, and Chronopolis, we will articulate questions many institutions are grappling with surrounding the integration of local services with distributed preservation networks, and identify broadly-applicable solutions and design patterns. To ensure our research is useful beyond the scope of the systems and institutions listed, we will survey a targeted group of diverse organizations that represent a variety of types of libraries, museums, archives, and preservation solutions. A subset of this group will serve as an ongoing advisory board. This planning grant will inform future projects and potential grants aligning with the IMLS focus theme of enhancing and building interoperable tools.

There are a number of questions which we can immediately identify as part of what needs exploration, some of these are:

- How does one curate objects to ingest into a long-term dark preservation system?
    - How does one signify during ingest that a certain object is DPN bound, or bound for other preservation services?
- How does versioning of objects/actively managed content play out in a long-term dark preservation system and how to automate these actions?
    - How often should versions get pushed to system? Is it the incremental change that is stored or the whole new version of the object?
- How is metadata stored with the object in the dark preservation system and how does it get leveraged in the versioning process?
- How do the technical solutions implemented stay flexible enough to sustain changes in human developed preservation policies that are affected by our economy?

Activities:

- Environmental/Feasibility Scan (Month 1 - 6)
    - Review planned or suggested development related to this topic
    - Targeted survey/interviews with a variety of organizations to better understand the breadth of implemented local systems
    - Research into versioning mechanisms

- User studies (Month 6 - 12)
    - Preservation policies and rights issues and how they are reflected in the technology itself
- Travel (Month 1, Month 6, Month 11)
    - Travel for identified individuals to meet in person

Deliverables:

- A report summarizing the key integration issues, drawing from interviews and the environmental scan
- User stories and design patterns which can be implemented in a variety of environments
- Recommendations for technical solutions

Project Resources: Personnel, Time, & Budget

Over the course of the year, the project team will follow agile processes to manage the diverse partnership identified for this project. The advisory team will meet virtually once per month to report out progress, blockers, and to make sure we are on track to meet our project goals and timelines. The team working more directly on the identified issues, may meet more regularly. The entire team will meet at least twice, to outline questions and goals. We will also set aside times at meetings where the majority of the team members will be in attendance, like CNI, Open Repositories, and PASIG.

We are requesting funding ($25,000) for travel for ten to twelve members of the research team to attend face-to-face meetings hosted at Northwestern University or University California, San Diego. These funds will cover airfare, hotel, and per diem for all members of the team. If there are funds left over, we will use this money to send members to Open Repositories 2017 to share preliminary results. We are requesting funding ($11,765) to cover staff costs directly or to pay for backfill of existing duties. The indirect cost rate is 36% (Other Sponsored Activity) totaling to $13,235 of the budget.

<u>Personnel</u>

- Evviva Weinraub (**Primary Investigator**) AUL for Digital Strategies
- Sibyl Schaefer - Digital Preservation Analyst for Research Data Curation at UCSD, Chronopolis Program Manager. (**UCSD Project Lead**)
- Other identified staff at Northwestern and UCSD
- Andrew Woods - Fedora/DuraSpace
- David Wilcox - Fedora/DuraSpace
- Michael Giarlo - Hydra-in-a-Box
- Dave Pcolar - Digital Preservation Network (DPN)
- Michael Ritter - Developer Chronopolis & DPN (University of Maryland)