Abstract: ***Planning a Community-Created Data Rescue Toolkit***

Recent changes in the operating procedures for scientific research agencies who were previously responsible for collecting, managing, and preserving data has sparked concern about sustained, long-term access to scientific research data. Communities of data professionals, researchers, librarians, and those who rely on analysis of data to inform decisions, responded with initial efforts to assume the role of data stewards, working to ensure continued accessibility and preservation. The sudden nature of the situation means that initial efforts to fill the void have been fractured and ad hoc, lacking coordination across groups of diverse stakeholders. The expertise in data management and preservation from Sheridan Libraries and other research libraries provide necessary skills and resources to contribute to efforts underway. We propose a project to coordinate distributed efforts for government data preservation with libraries providing the central, organizing structure and streamlined entry points for participants. With the proposed planning grant we will develop a blueprint for a toolkit of resources for preserving and providing access to data. As an outcome of the planning we will share documentation of our process as a model for distributed, collaborative, community-driven effort to preserve government data.

The community defines the toolkit as a collection of resources and materials, including human expertise, from which users can pick and choose items to aid them in their work. To plan the toolkit, we will hold a meeting of stakeholders at Johns Hopkins University in Fall 2018. We have recruited collaborators from a variety of institution types to represent the diverse needs and interests on the topic, including: small, medium, and large academic institutions; public libraries; research organizations; data archives; government agencies; and collaborative and consortial data projects. Many of our collaborators have particular interest in environmental data, but we also have participants with expertise in government information, public services, multi-institution collaborations, and data archiving. To elicit further engagement, we will describe the toolkit in a white paper that we will share with our community of collaborators and then with the wider field for public comment. We will also include a network map to visualize diverse efforts and collaborations for preservation of government data. By uniting a group who will gather and generate the materials necessary to inform individuals on how to engage, we create a core network and a distributed partnership of co-creators to share and shape efforts over time.

Beyond the benefits of the toolkit itself, the process of planning presents a unique opportunity to create a model for distributed, collaborative, community development across diverse stakeholders. While toolkits and community collaborations are common, documented models for working together inter-institutionally for the benefit of public, sustainable access to federal data and information will fill a gap and facilitate future endeavors. The community planning model will benefit other information and data professionals, inter-agency collaboration efforts, and interdisciplinary work that addresses the needs of multiple stakeholders.

We see a burgeoning need to develop and deploy a sustainable solution for long-term preservation and access of data produced and maintained by federal government agencies. The JHU library is well positioned with an experienced data services team to expand the National Digital Platform by facilitating coordinated involvement among libraries and other organizations to connect and support the robust efforts already underway. This planning proposal represents an integral step in sustaining access to essential data across disciplines. The outcome of this collaboration, where libraries provide technological and social infrastructure, will serve as a model for other data focused collaborations.

Planning a community-created data rescue toolkit

Johns Hopkins University (JHU) proposes a one-year planning grant (May 1, 2018 through April 30, 2019) to coordinate distributed efforts for government data preservation and with libraries providing the central, organizing structure and streamlined entry points for participants With the proposed planning grant we will develop a blueprint for a toolkit of resources for preserving and providing access to data. As an outcome of the planning, we will share documentation of the process as a model for distributed, collaborative, community-driven process to preserve government data.

Statement of National Need

Government Data Needs

Recent changes in the operating procedures for scientific research agencies previously responsible for collecting, managing, and preserving data has sparked concern about sustained, long-term access to scientific research data. Communities of data professionals, researchers, librarians, and those who rely on analysis of data to inform decisions, responded with initial efforts to assume the role of data stewards, working to ensure continued access and preservation.

The proposal recognizes the need for sustainable access to government data to inform future research that will benefit society. Additionally both researchers and data have diverse characteristics requiring specific attention from different types of professionals and institutions. The initial community focus and current scope emphasizes environmental data while acknowledging that needs extend to other forms of government data. We hope to extend the work to other domains in subsequent phases. There are effective, essential practices taking shape to preserve, care, and prevent loss or depreciation of data, but they are often isolated and could be supported and enhanced by others focused on similar initiatives. Government data needs include the union of meaningful efforts while empowering those who have differing levels of resources and expertise through collaboration.

Summary of National Efforts

Recent national contributions from diverse groups reflect varying levels of resources, infrastructure, and expertise. A major participant, Data Refuge[1], helped spread the word and hosted several events. Arising soon after, Libraries+ Network[2] engaged library professionals in the discussion. Groups are producing specific tools for preservation activities and new initiatives continuously appear. Coordination of efforts is essential in building the national capacity to ensure continued access to data for everyone.

In spring 2017, Data Rescue Boulder, a network of volunteers working with over 3 million datasets, approached the Sheridan Libraries at JHU seeking library-based support to provide long-term stewardship for collected government data. Data Rescue Boulder affirmed that the expertise in data management and preservation from Sheridan Libraries and other research libraries provide necessary skills and resources. The groups recognized the need for connections between existing efforts and in May 2017 presented an initial workflow to the ARL community positioning libraries as

[1] Data Refuge. http://www.ppehlab.org/datarefuge

[2] Libraries + Network. https://libraries.network/

providers of infrastructure and guidance, connecting participants to data experts. In July 2017, JHU hosted a preliminary information gathering meeting of data professionals to discuss the necessity of a coordinated system and standard resources for preservation efforts across diverse communities. Meeting participants agreed that a data rescue toolkit would provide a way for libraries and organizations to participate in varying capacities, by providing guiding resources and advising volunteers. Notes from this meeting are publicly available on the Open Science Framework (OSF).[3]

Need for a Community Toolkit

Diversification of the data rescue efforts across communities and institutions large and small, for profit, and not, will ensure that data and documentation accurately serve the full spectrum of citizen users. It is from community input that we can develop effective community resources. The community defines the toolkit as a collection of resources and materials, including human expertise, from which users can pick and choose items to use as aids in their work. A toolkit presents itself as a way to share a dynamic group of materials that users can select from, contribute to, and share with others. By uniting a group who will gather and generate the materials necessary to inform individuals and organizations on how to engage, we create a core network and a distributed partnership of co-creators from which to share and shape efforts over time.

The proposed planning meeting will be the first step to define and create the toolkit that will provide quality mechanisms and choice materials for distributed groups to work with data. We have identified benefits of the toolkit for communities who need sustainable access to federal data through time. These include:

Researchers: Concern exists among researchers and data professionals at academic institutions around the country. One of our activities in the initial information gathering was outreach to Earth science researchers and those in related disciplines. Through personal interactions with government data users, we were able to get an understanding of data sets, documentation, and products they utilize regularly.

Individuals impacted by the data: Earth science research and environmental data have significant social impacts. The more access researchers and stakeholders have to these data, the more work they are able to do that influences the lives of those throughout the nation, some of whom will use data as evidence of financial need or in litigation activities. How we understand and allot resources is often tied up in environmental data. The data and information science community identifies this through work with researchers who serve specific populations.

Information and data professionals: Those who work in information professions bring an understanding of long-term preservation for government data. The toolkit will collocate information resources and a network of people sharing expertise, perspectives, technical standards, and methods for working with the data and related materials. In our previous

[3] JHU Hosted Data Rescue Event July 2017 https://osf.io/vgc3q/

information gathering meeting we identified needs for working with environmental data in need of preservation by assessing the landscape of the issue and communicating with the community of collaborators. For example, our interactions with the distributed group Data Rescue Boulder gave the group insight into the issue at multiple institutions of a variety of types. Major barriers to participation cited in the discussions include: the need for technical understanding; access to infrastructure and tools; and community development. The toolkit will provide the mechanisms to address these needs.

The Benefits of the Planning Process

Beyond the benefits of the toolkit, the process of planning presents an opportunity to create a model for distributed, collaborative, community development across diverse stakeholders. While toolkits and community collaborations are common, documented models for working together inter-institutionally for the benefit of public, sustainable access to federal data and information is a significant contribution to future endeavors. For example, groups planning future collaborative work in support of social science, geospatial, or medical data can use and extend the shared model and learn from the challenges we encounter. Capturing the planning process for the community toolkit as a model will benefit other information and data professionals, inter-agency collaboration efforts, and interdisciplinary work to address the needs of multiple stakeholders.

Building and Extending Scholarship and Theory

The planning proposal aims to leverage community skills for social benefit. This is often talked about, modeled, and envisioned in literature, but in practice, proves to be extremely difficult. In our proposal we have a concrete start connecting communities as evidenced by expressed involvement of our collaborators and the success of past information gathering and environmental scanning efforts. The theoretical basis for organization of this work is Participatory Action Research (PAR), which unlike positivist and humanist ideologies seeks to reframe knowledge and knowing as an interactive process that includes the voices and perspectives of the people who will work with the material and benefit from continued quality access.[4] We embrace the multiplicity and fluidity of solutions that will shape the planning process and the content for the toolkit.

Our organization and process documentation is reinforced by and builds on work in distributed Knowledge Management (KM) addressed by scholars such as Haythornthwaite et al.'s 2006 discussion of bridging practices.[5] These processes also come to the fore in environmental management with principles of co-management where resource users and government engage in a knowledge partnership: sharing responsibilities, and encountering opportunities for social learning.[6] Finally, we build on issues of sustainability in preservation to construct a model and a dynamic resource that can be utilized not just for our focus in environmental data preservation, but also for multiple types of preservation and problem solving. We find examples of this work in current efforts

[4] Chevalier, J. M., & Buckles, D. (2013). *Participatory action research: Theory and methods for engaged inquiry*. Routledge.

[5] Haythornthwaite, C., Lunsford, K. J., Bowker, G. C., & Bruce, B. C. (2006). Challenges for research and practice in distributed, interdisciplinary collaboration. *New infrastructures for science knowledge production*, 143-166.

[6] Berkes, F. (2009). Evolution of co-management: role of knowledge generation, bridging organizations and social learning. *Journal of environmental management*, *90*(5), 1692-1702.

such as the Data Curation Network[7], and the Big Ten Academic Alliance Geoportal[8]. We include individuals from these initiatives in our proposed meeting collaborators and take into account strides they are making in distributed work and information science practices.

Project Design and Deliverables

Stakeholder Meeting

Our project will bring together existing communities dedicated to preserving government data to develop a toolkit to coordinate and aid in preservation efforts. We will hold a two-day planning meeting of stakeholders in Fall 2018 at Johns Hopkins University in Baltimore, MD. The first day will be devoted to developing the implementation charter, strengthening our existing relationships, welcoming new collaborators, and defining the components of the toolkit. The toolkit will contain a suite of social and technical resources to facilitate participation, including: an inventory of existing tools; a directory of individuals and organizations; and information on metadata. Specific characteristics and additional components will emerge through planning and contributions of meeting participants. Once in place, individuals will break out into the areas where they wish to contribute to solidify the individual pieces that will make up the components. In the second day we will determine roles and resources needed for developing the defined components, and solidify tasks, timelines, and communication schedules. Finally, the team will clearly lay out the communication touch points for the coming year.

After the meeting the project director and manager, aided by an LIS graduate student, will assess outcomes, document the planning process, and map the community as a way to visualize the network of input in various areas. The following objectives for the meeting include:

- To plan the toolkit, designate roles for execution, and establish working groups and a timeline for execution.
- To bring together individuals from multiple domains to work on the planning and contribute to the project.
- To define a method for leveraging community infrastructure to benefit community participation to serve the public.

Prior to the planning meeting related objectives may arise. In our delineation of these objectives, key staff on the project are operating under the assumption that groups across the country are invested in working to preserve government data. Additionally we assume that the methods of archiving and preservation we put forth will aid in sustainable access through time if mindfully administered with communication from our network of collaborators.

Committed Collaborators

The following collaborators are committed to participating in the project and attending the planning meeting in fall 2018. Please find the Letters of Commitment attached to the proposal.

[7] Data Curation Network. https://sites.google.com/site/datacurationnetwork/
[8] https://geo.btaa.org/

Andrew Battista	New York University
Andrew Janco	Haverford College
Andrew Johnson	University of Colorado
Brendan O'Brien*	Environmental Data Government Initiative (EDGI)/Data Refuge
Brian Wee	Massive Connections
Catherine Morse	University of Michigan
David Bleckley	ICPSR - Data Lumos
Fernando Rios*	University of Arizona
James Ng*	University of Notre Dame
Jessica Barrientos	Westminster Public Library
Jim Gillispie	Johns Hopkins University
Joan Saez*	Data Rescue Boulder
Justin Schell*	University of Michigan/ EDGI/Data Refuge
Lisa Johnston	University of Minnesota/Data Curation Network
Matt Spitzer*	Center for Open Science
Megan Potterbusch*	George Washington University
Michael Zarafonetis	Haverford College
Ruth Duerr*	Ronin Institute/Federation of Earth Science Information Partners (ESIP)
Ryan Mattke	University of Minnesota/Big Ten Geospatial Data Project
Stephen Balogh	New York University
Stephen Diggs*	Research Data Alliance/Scripps Institution of Oceanography
Vivian Hutchison	United States Geological Survey

* Individual or associated institutional representative in attendance at the JHU workshop, July 2017

In addition to the committed collaborators, we plan to recruit other collaborators to fill the following roles: an LIS graduate student; an additional public library collaborator; and an additional representative from a government agency.

Toolkit Blueprint and White Paper

In the month following the planning meeting, the project director and manager will draft a blueprint of the toolkit in a white paper. The white paper is our method of communicating the plans in a structured way for public comment and peer evaluation. It will include definitions of the toolkit and

the intended audiences, name participants, delineate toolkit elements, and take inventory of needed resources and roles for creation. The paper will layout two potential timelines: one with and one without additional IMLS funding. Finally, the paper will outline intended outcomes, outreach plans, and a sustainable toolkit management strategy. After the initial draft, we will submit the document to the collaborators for advising, editing, and additions. To visualize the network of contributors and their participation efforts as a reference for future connection and understanding, key staff will create a network map and share it with the community. After a month of review and synthesis by the meeting collaborators, we will share this white paper for feedback through multiple channels such as professional listservs and information networks.

Organization to Support ongoing Community Collaboration

In order to facilitate sustained collaboration from the stakeholder communities, our planning document and material sharing space will be public through the the open source collaboration platform the Open Science Framework (OSF). Built by the not-for-profit group, the Center for Open Science, the OSF is both a workflow system and a flexible repository that allows for coordination of efforts throughout the entire lifecycle of a project.[9] This choice represents our underlying motivation that representatives coming from diverse groups will directly contribute to the structuring of the plan through a horizontal style of group governance and contribution.

One recognized risk is that levels of commitment from collaborators are at the mercy of their time to volunteer on the project. The collected letters of commitment from our participants formalize our existing connection with them and indicates their willingness to devote time to the project. We also understand that an intended, subsequent IMLS national leadership grant may not receive funding. As a result, we will focus on building a strong collaborative foundation, and on the outcomes of the planning as the production of a model to be used by the broader information science, data science, and knowledge management communities for future distributed collaborative work. Important work can still be achieved by establishing this strong and well-networked community. We can shift and adjust our efforts to move forward based on available resources.

The planning process model provides a structure, templates, and resources from the full meeting experience. This includes a brief introduction to the event and the core philosophy with resources on PAR and diversity planning, considerations for budgeting, invitation materials, agenda and activity templates, code of conduct samples, checklists, a network mapping tutorial, outreach plans, and our own lessons learned. By making the planning products public and sharing materials through scheduled outreach activities, we open the conversation for others to join the project and make it their own. These actions support our PAR perspective. Throughout the planning period we will encourage collaborators to get insight from the individuals with which they work and serve, and to bring in additional perspectives they find missing from the framing, semantics, and organization of the resources.

The eventual toolkit will be a dynamic, living project. We will make this explicit and observe the way others use it, as this may inform the quality and saliency of the project through time. We have scheduled assessment and synthesis of feedback into the timeline provided in the accompanying Schedule for Completion.

[9] The Open Science Framework. https://cos.io/our-products/open-science-framework/

Timeline and Tracking

We will track our progress throughout the year by the output and assessment of our documentation and stakeholder feedback and the notes of the virtual meetings for the working groups we formed. Using the OSF shared space for collocation of collaboration materials allows the group to maintain a visual understanding of the status of various efforts at any point and keeps each team accountable for meeting specific goals and objectives that they have defined. Consistent communication via these stable channels of participation are the foundation for the tracking and evaluation of our community efforts.

The audience for the toolkit and related planning models are wide reaching: institutions large and small, academic universities, not for profit agencies, public libraries, and anyone who has a significant stake in government data use in the future. The planning process model targets groups interested in implementing similar efforts so that they may employ it for their own collaborative planning. Both products will need to be discoverable and accessible by these audiences and easily appended.

Our motivation for the focus on PAR stems from our commitment to engage and empower representatives across communities of use along with data and information service professions, so that recognized audiences and potential participants see themselves reflected in the information resources offered. This focus extends to underserved populations. By providing mechanisms of support for varying types of information service professionals, we aim to make entry points for institutions with limited resources, and institutions in underrepresented communities to use this toolkit and/or the planning model in working with their patrons in a participatory capacity. In order to make this possible, we are focusing on the language and materials and engaging public libraries and representatives in the communities who will be able to provide insight. Johns Hopkins is in a prime location for facilitating engagement of underrepresented communities and we look forward to making connections for input as the planning progresses.

Communication and Outreach Plan

All of the materials that we develop as a team will be publically available on the OSF. Using this interface, interested individuals can make comments, and take the structure and content to extend it in a similar way to Github's "forking" functionality. We chose the OSF because it is free, open to anyone, and exemplifies the qualities of data and information preservation and stewardship that will sustain the resources through time. We will write the white paper outlining the toolkit blueprint after the planning meeting. This allows for group sharing and comments via professional organization listservs, ensures that the community is aware of the existence of the plan, and provides an opportunity for interested individuals to provide insight and get involved. Finally, we plan to share our work with related communities via formal presentations at larger venues frequented by a broad cross section of the information science population, for example: the American Library Association (ALA) and the Coalition for Networked Information (CNI). This is another way to connect in person with interested parties, receive feedback, and answer questions face to face.

Sustained Management, Resource Allocation, and Evaluation

The project director and project manager (Mara Blake and Reid Boehm, respectively) will lead and manage the project, including organizing the meeting and to keeping track of communication, sharing, and timing. Blake and Boehm will facilitate greater community participation in shaping and influencing the project by engaging the committed collaborators, taking group suggestions consideration and reaching consensus in each of our teams of work. We have chosen participants from the research community who we believe represent the perspectives of our audiences and have the expertise and capacity to facilitate the work. Blake and Boehm included CVs demonstrate their background and expertise with data preservation and collaborative data projects, in addition to the list of contributors and letters of expressed commitment.

The year timeline will give the community an appropriate amount of time to plan, hold the meeting, and conduct follow up discussions, and generate material that will tie up any loose ends in preparation for the implementation period. Our budget documentation includes funding for collaborators to travel to the meeting at JHU, and professional transcription, as well as considering travel costs for a presentation at a venue such as ALA or CNI. We also request funding for an LIS graduate student to participate in and contribute to the project. Finally, a remaining amount will go to a portion of the salary for the project lead and project manager.

We will use a gantt chart outlining our schedule for completion as a way to track our progress. The interface allows for us to make notes, assign tasks, and keep a log of any changes to the plan. Part of the initial meeting will involve the creation of clearly defined goals and objectives and delineated team expectations. Scheduling consistent, regular communication as a group will ensure that we stay on the same page and in sync with our agreed upon schedule.

As we evaluate our progress we will confirm that our project objectives are easily accessible in the shared place on the OSF where the group can match expectations with the delivered content. This space will be structured prior to the meeting and we will make sure that all collaborators are aware and able to contribute to the documentation throughout the process.

Having the plan and related documentation open to the public, along with sharing to the community in a large presentation venue, allows the group to gain input from stakeholders that otherwise might not have a chance to connect. Participants in the planning process will also have opportunities to express successes, frustrations, and reservations during and after the meeting and related activities.

Diversity Plan

The collaborators and participants in the planning grant will reflect the diverse communities involved with preserving government data. We have recruited collaborators from a variety of institution types, including: small, medium, and large academic institutions; public libraries; research organizations; data archives; government agencies; and collaborative and consortial data projects. Leading up to the planning grant, we also plan to recruit one additional representative from a public library and one from a government agency. We have strength in collaborators that are particularly interested in environmental data, but we also have participants with expertise in government information, public services, multi-institution collaborations, and data archiving. We are also requesting funding for an LIS graduate student to participate.

These diverse perspectives will help us account for the groups who will be served by this project, which include: small academic universities and community colleges, not for profit organizations, public libraries and city government, citizen stakeholders, and those who use public Earth science data to inform their work and life. These different groups were suggestions based on our conversations with data professionals around the country who work in these different areas and/or collaborate with others who work in these areas.

Our budget request includes full funding for the travel-related costs of meeting attendance to ensure that all participants can attend regardless of other available support. We hope this will facilitate full participation of collaborators who are often underrepresented at these types of meetings and will make our meeting as accessible as possible, particularly for representatives from public libraries and non-profits.

Although our collaborators are primarily drawn from those already engaged in the work of preserving government data, we plan to design the toolkit explicitly to facilitate the inclusion of other interested, underrepresented groups. Our collaborators and potential collaborators are individuals with different perspectives from our research domains and areas of work. We understand the need standard language and terminology that allow us to communicate effectively. We will account for this and build accessible language and ontologies into our toolkit by fully explaining our intent, disambiguating unclear terms, and linking to broader explanations as we can.

Our project relies on the successful execution of a productive meeting that brings together diverse stakeholders. To ensure unimpeded participation, we will use a written Code of Conduct and share it with participants ahead of the meeting. We will use the Code4Lib Code of Conduct[10] and tailor it to our needs. Additionally, we will utilize the meeting role assignments from the Ada Initiative to promote on-topic and inclusive conversation.[11] We will also create name tags for participants that specify their preferred pronouns and encourage participants to openly share any other specific needs as individuals.

National Impact

We see a burgeoning need to develop and deploy a solution for the long-term preservation and access of data produced and maintained by federal government agencies. The JHU library is well positioned with an experienced data services team to expand the National Digital Platform by facilitating coordinated involvement among libraries and other organizations to connect and support the robust efforts already underway. This planning proposal represents an integral step to sustaining access to essential data across disciplines. The outcome of this collaboration, where libraries provide technological and social infrastructure, has potential to endure as a model for other data related needs.

Transforming Practices

[10] 2017 Code4Lib Code of Conduct http://2017.code4lib.org/conduct/.

[11] Ada Initiative. Running your unconference discussions effectively. https://adainitiative.org/2013/10/02/running-your-unconference-discussions-effectively-adacamp-session-role-cards/

As information professionals we may not always be able to take time to think about and document how we plan and to organize work that brings multiple perspectives to the table in functional planning sessions. Our hope by sharing this model for working with a distributed and diverse group is that others will see the benefits of participatory work, thinking not just about outcomes and deliverables, but about how the process itself can create more informed, accessible, and relatable content. We will further influence nascent scholars in the field by mentoring an LIS graduate student as part of project.

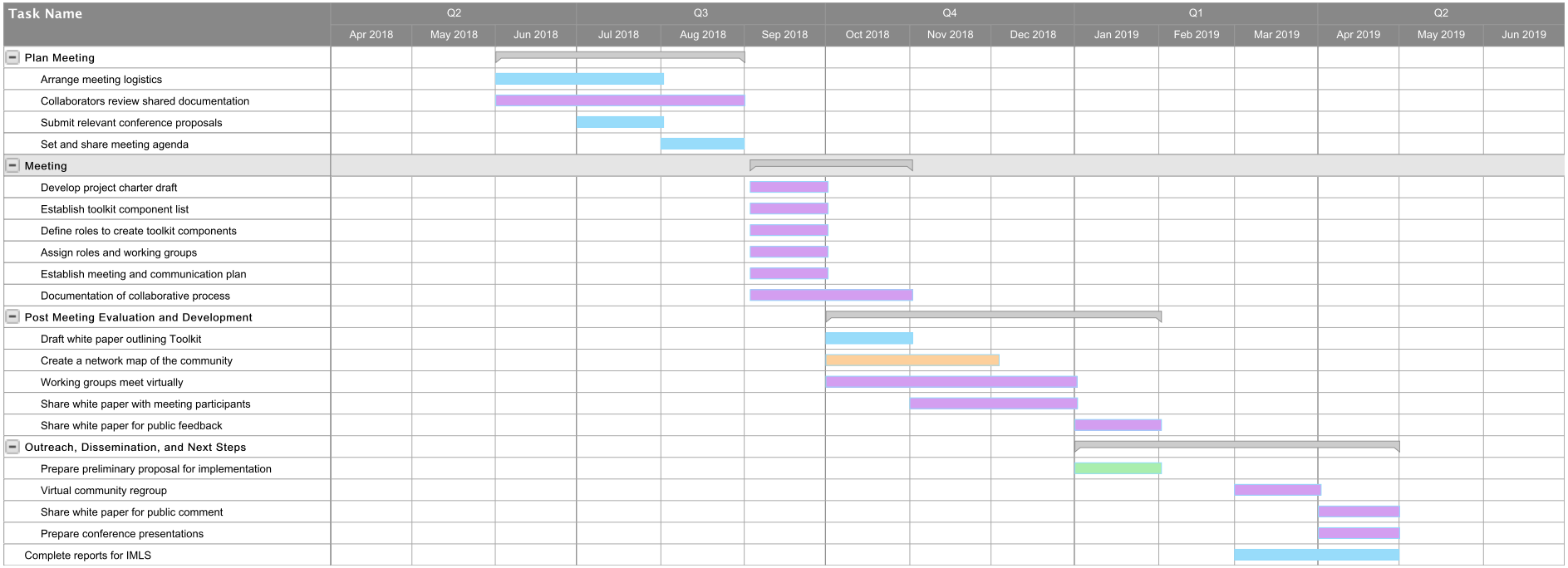
Adaptability

In our plans we are thinking carefully about formats, usability, soliciting feedback, presenting at a conference or meeting and framing materials and documentation as dynamic models. We will work with related communities and all interested entities to ensure that they are able to use both the process model and the toolkit, that they understand it, see their needs reflected in it, and that there is a network of information professionals that they can contact to ask questions. Sustained dialogue and active outreach will allow for maximum use and effective problem-solving.

Products for the Benefit of the Field and Beyond

We propose a project to coordinate the distributed efforts around data rescue and with libraries providing the central, organizing structure, streamline entry points for participants. In creation of a systematic planning model for working with a group of diverse and distributed professionals to solve a problem or create a project that will reflect a multitude of stakeholders we can facilitate future community collaboration. This model and the related documentation, shared on the OSF public sharing platform and advertised in multiple arenas for comment and use, will be available as a living document where we can learn from continued interactions and iterations.The toolkit that comes out of the planning process will be an integral part of unifying the preservation efforts to bridge the gaps in support that the community has identified.

Schedule for Completion



Exported on January 8, 2018 4:11:17 PM EST

DIGITAL PRODUCT FORM

Introduction

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

Instructions

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

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

Part I: Intellectual Property Rights and Permissions

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

We intend to use a non-restrictive open source license for the software we are developing. Johns Hopkins University has applied a variety of non-restrictive licenses. We will work with our Tech Transfer department to determine the most appropriate one for this project. Johns Hopkins University will own the rights to the white paper, but it will be disseminated freely through its institutional repository and GitHub.

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

Johns Hopkins University will own the prototype and white paper. We intend to make the prototype available for unrestricted open access and use, likely limited to non-commercial uses. Potential users of the tool will be notified of any conditions of use through a license agreement, which will be digitally provided prior to use. The white paper will be freely accessible through JScholarship, the Johns Hopkins University institutional repository (https://jscholarship.library.jhu.edu/). A citation for the paper will be provided.

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

The prototype will likely be a locally-hosted application to mitigate privacy breeches or other unauthorized uses. If we decide to incorporate code from pre-existing software libraries, we will abide by the licensing terms of that software and incorporate those terms into our software license.

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

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

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

We will produce two products: a white paper and a software application prototype intended for public release. The white paper will likely exist in PDF form. The functional requirements we will develop to inform the creation of the prototype will likely drive the ultimate format that the prototype takes. While it would be premature at this time to commit to a particular standard for the code base, a number of existing perceptual hashing libraries are written in Python, which will be fully investigated alongside other options presented during information gathering and prototype development stages.

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

As detailed in the narrative, part of the project will involve selecting a service provider to create the prototype, so it would be premature at this time to provide more information; however, we have compiled a list of likely competencies and skills that the software developer will have. The software developer's work will take place on their own computer equipment; his or her work in progress will be stored on a server accessible to the project team. The white paper will be created using a standard word processing application on Johns Hopkins University computer equipment.

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

PDF for the white paper; format for the prototype will be informed by the functional requirements and advisory board input that are part of the project.

B. Workflow and Asset Maintenance/Preservation

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

To oversee the entire project, we intend to create a project charter that employs standard project management practices such as scope of work, stakeholders, and milestones. We intend to create a detailed Request for Proposal when recruiting the software developer. We will sign a contract with the developer that will outline scope of work, outcomes, and a timeline for completion. The Advisory Board will meet on a regular basis with the software developer to gauge progress. The Project Manager will serve in the primary oversight role of the developer's work.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

The white paper will be stored in our institutional repository. Both the white paper and the prototype will be stored in Johns Hopkins University Sheridan Libraries' GitHub. Both digital products will be preserved locally in accordance with the Sheridan Libraries' digital preservation procedures.

C. Metadata

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

This cannot be answered at this time, as the functional requirements created as part of this project will inform what standards we use. However, we intend to apply widely-adopted standards whenever possible. It is likely that existing products that perform digital work (such as BitCurator) will inform our decisions. Using the tool will produce metadata; we will evaluate and select a standardized reporting format for logging usage activities.

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

Metadata and associated digital assets will be preserved on the Sheridan Libraries' networked storage system. Any major products developed in the course of the project will also be shared through GitHub.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the

digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

The white paper and prototype will be shared through our GitHub space. We also intend to make announcements at the end of the project directing interested parties to this space. The white paper will be cataloged in the Johns Hopkins University Sheridan Libraries catalog, following library standards.

D. Access and Use

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

The white paper and prototype will be openly available online through our GitHub space, accessible via standard web browsers.

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

Examples relevant to this project include:

National Digital Stewardship Residency project documentation: https://github.com/jhu-archives-and-manuscripts/homewood-photo

The Archaeology of Reading: https://archaeologyofreading.org/ (Sheridan Libraries responsible for software development; AOR is a collaboration with the Center for Editing Lives and Letters at UCL and the Princeton University Library)

JScholarship institutional repository: https://jscholarship.library.jhu.edu/

Part III. Projects Developing Software

A. General Information

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

This project will develop a prototype for a tool to identify near-duplicate digital images based on their visual content. One of goals of this planning grant is to identify the functional requirements that the software would perform. Initially identified functions are: allowing a user to point to a directory of images and recursively comparing the visual content of those images to one another in a many-to-many scenario; alerting the user to groupings of images ("clusters") most likely to have near-duplicate content; allow for transparency as to how clustering was achieved and allow for adjustments in sensitivity/threshold values; provide reporting in a standardized format that can be packaged with archival information packages (AIPS) with the option to store hashes for later querying; run locally without having to upload images to an off-site server.

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

Existing software in the BitCurator suite partially perform similar functions to the tool we intend to prototype. FSlint (https://wiki.bitcurator.net/index.php?title=Identify_and_delete_duplicate_files) allows users to find duplicate files and make selection decisions within the application. It also has the ability to group content with similar file names and allows the user to adjust the sensitivity to create clusters. FSlint does not employ a method for finding duplicate or near-duplicate files based on their visual content, such as perceptual hashing. The tool sdhash (https://github.com/sdhash/sdhash) allows for the comparison of two "arbitrary blobs of data" based on the similarity of binary data; however it does not take perceptual similarities into account and only compares in a one-to-one ratio.

Existing perceptual hashing tools partially perform a major function of the software we intend to create, which is to compare digital images based on their visual content. Libraries such as ImageHash (https://github.com/JohannesBuchner/imagehash), pHash (https://www.phash.org/) and Blockhash (http://blockhash.io/) focus on finding exact duplicate content through one-to-one comparisons, whereas our tool intends to work for significantly different use cases by identifying near-duplicate content and creating clusters. These libraries implement various underlying perceptual hashing algorithms and will require further testing to see which algorithms, when used

alone or combined, will be most applicable to our software. Further, we intend to produce a graphical user interface, and there are none associated with these libraries.

B. Technical Information

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

This cannot be answered at this time, as the functional requirements created as part of this project will inform what standards we use. However, we intend to apply widely-adopted standards whenever possible. It is likely that existing products that perform digital processing work (such as those in the BitCurator suite) will inform our decisions. Many perceptual hashing libraries are written in Python, so that language will be of strong consideration.

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

The prototype will complement existing technologies for supporting digital forensic workflows, such as the BitCurator suite of tools. It is our hope that the prototype will also advance the image processing field by having a locally-run, easy-to-use tool for identifying near-duplicate images, extending uses and implementations for existing perceptual hashing libraries, and complementing the work of computer vision products such as Google's Cloud Vision API without the necessity to rely on machine learning.

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

Because the functional requirements will address dependencies, and the creation of those functional requirements are to be developed as part of the grant, we are not at liberty at this time to detail what dependencies will exist. However, the Sheridan Libraries has expertise in Java, Ruby, and Python programming languages, and Linux systems for running software. It is likely that the prototype will align with these existing technologies.

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

The prototype will come with instructions on how to install and use the tool. This documentation will be shared through our GitHub with the application. Because this is a prototype, we don't expect to refine the documentation or the features of the product unless we decide to seek funds to develop a more robust application based on the prototype.

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

Johns Hopkins University Archives and Manuscripts: https://github.com/jhu-archives-and-manuscripts

Johns Hopkins University Sheridan Libraries: https://github.com/jhu-sheridan-libraries

Data Conservancy: https://github.com/DataConservancy

RMap Project: https://github.com/rmap-project/rmap

As the prototype developer will be selected during the grant period, we cannot yet provide examples of any software tools they have created.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

We intend to use a non-restrictive license for the software we are developing. Johns Hopkins University has applied a variety of non-restrictive licenses. We will work with our Tech Transfer department to determine the most appropriate one for this project.

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

The prototype, source code, and associated documentation will all be made available through the Sheridan Libraries' GitHub. The white paper will link to this repository.

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

Name of publicly accessible source code repository: Johns Hopkins University Sheridan Libraries

URL: https://github.com/jhu-sheridan-libraries

Part IV: Projects Creating Datasets

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

Part IV is not applicable for this project.

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

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

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

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

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

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

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

Name of repository:

URL:

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