Always Already Computational: Library Collections as Data

Abstract

University of California Santa Barbara, University of Pennsylvania, University of North Carolina Chapel Hill, Penn State University, Emory University, and Texas A&M University seek $100,000 for a National Forum Grant under the National Digital Platform funding priority to hold a series of meetings of librarians, archivists, museum professionals, researchers and practitioners, and technologists. These meetings will support development of a strategic approach to developing, describing, providing access to, and encouraging reuse of library collections that support computationally-driven research and teaching in areas including but not limited to Digital Humanities, Public History, Digital History, data driven Journalism, Digital Social Science, and Digital Art History.

Predominant digital collection development focuses on replicating traditional ways of interacting with objects in a digital space. This approach does not meet the needs of the researcher, the student, the journalist, and others who would like to leverage computational methods and tools to treat digital library collections as data. For example, a Digital Humanities researcher engages in term frequency visualization, topic modeling, and network analysis across thousands and sometimes even millions of items. Collection managers have no common framework to guide transformation of collections to meet this type of use. Where the few computationally-amenable digital library collections have been developed, they are each prepared, described, and made accessible differently and are typically difficult to discover. Furthermore these few examples are primarily focused on provision of text data, whereas the scope of data extends to images, moving images, sound, web archives, and beyond. Finally, as a technical and social consideration, library repository platforms, both sole institution and aggregative efforts, are not currently developed to enable ready access to collection objects at scale, nor do they align collections with a data reuse paradigm. Although the Digital Public Library of America, Stanford University, and DuraSpace have received IMLS funding to build an out-of-the-box digital collections solution, it does not extend to facilitate development of library collections that support computational use.

The National Forum Grant will support a two and a half day librarian, archivist, museum professional, researcher and practitioner, and technologist summit at the University of California Santa Barbara in Spring 2017. In addition to the summit, grant funds will be allocated to supporting three meetings at library and technology conferences and two meetings at research community conferences to share, as well as iteratively refine, *Always Already Computational: Library Collections as Data* work. In keeping with the themes that emerged in the National Digital Platform report, this project embraces the notion of "radical and systematic collaborations." At each of the meetings that follow the National Forum meeting, the project team will incorporate community contribution in substantive ways by attaching formal working groups and signatories from each meeting. In doing so the project team will commit more fully to meaningful, vested, and systemic collaboration with peers throughout the country at institutions large and small.

IMLS support will enable development of a community focused strategic direction (November 2016 - November 2018) that leads to **(1)** creation of a library collections as data framework that supports pragmatic collection format transformation and documentation, **(2)** development of computationally amenable library collection use cases and user stories **(3)** identification of methods for making computationally amenable library collections more discoverable through aggregation and other means, (**4)** and articulation of guidance, in the form of functional requirements, that will support development decisions relative to technical feature integrations with existing initiatives like IMLS's Hydra-in-a-Box.

Always Already Computational: Library Collections as Data

University of California Santa Barbara, University of Pennsylvania, University of North Carolina Chapel Hill, Penn State University, Emory University, and Texas A&M University seek $100,000 for a National Forum Grant under the National Digital Platform funding priority to hold a series of meetings of librarians, archivists, museum professionals, researchers and practitioners, and technologists. These meetings will support development of a strategic approach to developing, describing, providing access to, and encouraging reuse of library collections that support computationally-driven research and teaching in areas including but not limited to Digital Humanities, Public History, Digital History, data driven Journalism, Digital Social Science, and Digital Art History.

---

1. Statement of Need

Over the past twenty years a large number of library collections have been digitized. Reel-to-reel tape, photos, and manuscript pages now fill digital repositories throughout the country. Combined with an increasing flow of born-digital items, digital library collections have come to represent a rich community resource for users that seek to find and watch a film, look at a photo, or read a page. Yet a focus on replicating traditional ways of interacting with collections in a digital space does not meet the needs of the researcher, the student, the journalist, and others who would like to leverage computational methods and tools to treat digital library collections as data. For example, a Digital Humanities researcher engages in term frequency visualization, topic modeling, and network analysis across thousands and sometimes even millions of items.

Collection managers have no common framework to guide transformation of collections to meet this type of use. Where the few computationally-amenable digital library collections have been developed, they are each prepared, described, and made accessible differently and are typically difficult to discover. Furthermore, these few examples are primarily focused on provision of text data, whereas the scope of data extends to images, moving images, sound, web archives, collection metadata, and beyond. Finally, as a technical and social consideration, library repository platforms, both sole institution and aggregative efforts, are not currently developed to enable ready access to collection objects at scale, nor do they align collections with a data reuse paradigm (Stodden and Miguez 2014; Padilla 2016). Although the Digital Public Library of America, Stanford University, and DuraSpace have received IMLS funding to build an out-of-the-box digital collections solution, it does not extend to facilitate the development and sharing of library collections that support computational use. Before existing and potential projects can confidently move forward to serve the computational needs of researchers, grounds must be set for meaningful collaboration between librarians, archivists, museum professionals, researchers and practitioners, and technologists to develop pragmatic solutions salient to the challenges of developing computationally amenable collections.

Despite lack of concerted library development in this area, disciplinary and professional communities forge ahead with research and creative production. Types of use exhibited by these communities include but are not limited to text analysis, image analysis, sound analysis, and network analysis. Orientation to the full scope of *academic* use types can be gained through in-depth analysis of data use practices across disciplines as represented in core Digital Humanities journals (Padilla and Higgins 2016), by reviewing works at the annual global Digital Humanities conference (Weingart 2016), and by studying edited volumes that have to this point effectively compiled a broad range of research in this space (Gold 2012; Gold and Klein 2016; Burdick, Drucker, Lunenfeld et al 2012; Schreibman, Siemens, Unsworth 2016). Keeping pace with this research, major scholarly societies like the Modern Language Association and the American Historical Association have moved to adapt to scholarly production that is expressly digital by issuing guidance on digital scholarship evaluation in the context of tenure and promotion decisions (MLA 2012; AHA 2015). Outside academia, an uptick in data driven approaches by journalistic entities

as diverse as the New York Times, Quartz, and Buzzfeed extends commitment to computational work and reproducible analysis from a primarily academic activity to a broader field of social concern (Singer-Vine, 2016). It is also worth noting that academic and non academic work in this space appears to be converging, as exhibited by an upcoming conference, "Digital Humanities + Data Journalism" at the University of Miami. As activity across these increasingly fluid spaces expands, individuals and organizations operating outside of cultural institutions seek to fill data infrastructure and data standards gaps. With funding from the Knight Foundation, projects like Dat explore leveraging Git to facilitate collaborative work with data, and organizations like the Open Knowledge Foundation dedicate resources to developing semi-self-documenting data packages to enable more streamlined data sharing (Ogden 2016; Open Knowledge Foundation 2016).

Computational activity like the kind described above is contingent on the availability of collections that are tuned for computational work (Hughes 2014). Suitability is predicated on form, integrity, and method of access (Padilla 2016). Consider as a use case, a project that entails text analysis of many thousands of works. Generally speaking, digital collections provide access to images of book pages in the PDF format, along with a series of other image-based formats. This form of data is well-suited to a reading experience, but given that data instantiated in a TXT file are more readily analyzed by a common suite of text analysis tools, access to TXT derivatives would be ideal. With respect to integrity, consideration turns toward standards for describing collections that extend past typical bibliographic description to adapt research data documentation conventions meant to support reproducibility. These considerations include: indication of data modification, data quality, and algorithmic dependency. Methods of access are currently widely divergent, spanning simple provision of compressed collection objects in ZIP files (University of North Carolina Chapel Hill), exposing a static collection website that can be crawled using a tool like wget (University of Pennsylvania), leveraging Github for text collection access (Indiana University), provisioning an API (Digital Public Library of America), enabling FTP access to collections, delivering physical hard drives with data on them (Price 2014), mediating computational processes performed on collection data through a platform (Bonn 2016), to facilitating data access through use of torrent technology (Academic Torrents). There is clearly no consensus driven best practice in generating, describing, and providing access to computationally amenable library collections. In lieu of establishing these practices, institutions run the risk of misplaced investment of resources that foster the creation of an irregular, ultimately disorienting data access environment.

While approaches to preparing library collections that are readily amenable to computational approaches are nascent, preliminary models have been advanced by project team institutions. Critical thought in this area of work is reflected in an emergent scholarly literature and is an increasingly resonant area of interest within the library community (Padilla 2016; Varner 2016, Padilla and Higgins 2014). For example, at Digital Library Federation (DLF) 2015, a working session led by this project team on "Digital Collections as Data: Re-Packaging, Re-Mixing, and Sharing Collections for New Forms of Scholarship" drew nearly 100 participants. Subsequently, a separate but complementary DLF Digital Library Pedagogy group was formed, which in turn drew more than 150 members in fairly short order (Dickson and Kelly 2016). As work in this area of need begins to scale to institutions large and small throughout the country, we are at an opportune moment. A gathering of key stakeholders is needed to craft a strategic direction that leads to **(1)** creation of a library collections as data framework that supports pragmatic collection format transformation and documentation, **(2)** development of computationally amenable library collection use cases and user stories **(3)** identification of methods for making computationally amenable library collections more discoverable through aggregation and other means, (**4)** and articulation of guidance, in the form of functional requirements, that will support development decisions relative to technical feature integrations with existing initiatives like IMLS's Hydra-in-a-Box.

2. Impact

Few attempts have been made to prepare and provide access to collections that are tuned for computational use. Hathitrust Research Center (HTRC) represents perhaps the most well developed and well resourced of existing approaches. With its research Portal, application programming interface, high performance computing backend, and dedication to research and advocacy that supports the ability to leverage computational methods on in-copyright works, HTRC does a broadly heterogeneous research community a considerable service. However, this model and its associated set of personnel and infrastructure do not extend to consider how institutions large and small throughout the country can develop their own collections and infrastructure to meet a range of computational uses in light of resources particular to local context. This project aims to fill that gap.

The National Forum Grant will support a two and a half day librarian, archivist, museum professional, researcher and practitioner, and technologist summit at the University of California Santa Barbara in Spring 2017. In addition to the summit, grant funds will be allocated to supporting workshops at library and technology conferences as well as research community conferences to share and iteratively refine, *Always Already Computational: Library Collections as Data* work. In keeping with the themes that emerged in the National Digital Platform report, this project embraces the notion of "radical and systematic collaborations." On the library side, participant selection will span university, college, and special libraries. On the technologist side, participant selection will span institutions like the Digital Public Library of America, the Internet Archive, organizations like George Mason University's Center for History and New Media, as well as projects like Hydra-in-a-Box, and the Knight Foundation funded Dat. On the researcher/practitioner side, participant selection will span scholars operating in the Humanities and Social Sciences engaged in computational research, as well as leading data-driven journalism practitioners from publications like *Buzzfeed News* and *Quartz*.

At each of the workshops held across seven conferences that follow the National Forum meeting, the project team will incorporate community contribution in substantive ways by attaching formal working groups and signatories from each community. In doing so the project team will commit more fully to meaningful, vested, and systemic collaboration with peers throughout the country at institutions large and small. Formal contribution to the project will not be limited by opportunity to attend any of the workshops that follow the National Forum meeting. Virtual contribution to the project will be enabled and encouraged throughout the project via the project website. While the activity of this project will be initially framed by a centralized group of representative experts, at the conclusion of the project it will have been strategically decentralized and attuned to the realities and attendant priorities of a larger, nationally dispersed group.

3. Project Design

3.1 Identification of participants

The project team recognizes that the challenges described in this project require multifaceted solutions, developed and diversified according to community investment in varied stages of library collection development, description, access, and use. This recognition is aligned with intention through strategic identification of representatives that can expertly frame community interests at the National Forum meeting. Initial identification of key representatives spans librarians, archivists, museum professionals, researchers and practitioners, and technologists. Within each group diversification of possible contribution will be evaluated with care. In what follows, community groups and possible contributors are identified.

Librarians, Archivists, Museum Professionals
Librarians, Archivists, and museum professionals have a deep understanding of collection production, description, access, and use. Potential invitees include [redacted]
Collectively this group will bring to bear depth of experience in developing collections for computational use, data curation services, dynamic crowdsourcing projects predicated on library collections, next generation web archiving services, digital public history projects, and in depth user studies that focus on evaluating computational use of collections.

Researchers and Practitioners
In order to develop impactful solutions it will be essential that the project encourages substantive collaboration with users engaged in computational utilization of collections. Invitations to representatives of Humanities and Social Science disciplines as well data driven journalism practice will be extended. Potential invitees include [redacted] Collectively this group represents a wide range of perspectives on the practice of doing computational work predicated on access to heterogeneous digital objects.

Technologists
Any conversation about digital collection development must include the individuals responsible for building systems intended to support their use, as well as parallel projects that hold potential to inform that work. Potential invitees include [redacted] Collectively this group holds expertise in a range of open source tools, platforms, and technology projects, bringing awareness of the process required to incorporate computational use cases into the shared technical infrastructure and user interfaces of libraries, archives, and museums.

3.2 Project Performance Goals & Outcomes

The combined activity of this project will lead to **(1)** creation of a library collections as data framework that supports pragmatic collection format transformation and documentation, **(2)** development of computationally amenable library collection use cases and user stories **(3)** identification of methods for making computationally amenable library collections more discoverable through aggregation and other means, (**4)** and articulation of guidance, in the form of functional requirements, that will support development decisions relative to technical feature integrations with existing initiatives like IMLS's Hydra-in-a-Box.

3.3 Structure and Agenda

National Forum Agenda

In early 2017, the project team will convene a group of stakeholders for a two and a half day national forum at the University of California, Santa Barbara. The goal of this meeting will be threefold: **(1)** to articulate computationally amenable library collection use cases; **(2)** to initiate a collection of best practices that support developing, describing, and providing access to computationally amenable library collections; **(3)** to create a dynamic feedback structure that enables a wide range of communities to shape and play a primary role in the production of final project outputs. Outputs produced during the forum will include computationally amenable library collection use cases; best practices for the transformation of library collections to meet computational use; recommendations for increasing discoverability of collections to avoid siloing; and guidance for identifying and developing technical features that encourage collection access and use. Collaboratively developed outputs will reflect the range of expertise at the meeting and will help the project team create a series of resources that hold high potential for resonating with community need. As initial stakeholders for their respective communities, forum participants will play a crucial role in shaping the questions as well as participation mechanisms that will be used after the forum to encourage iterative community refinement and ultimate finalization of project outputs.

Day One - *Articulate use cases*
During the first day of the meeting, group members will reflect on core challenges to developing, describing, and providing access to computationally amenable library collections. Group members will have been prompted to frame their thinking in this space along four questions that align to data definition, description, access, and reuse. Individual group members will be pre-selected and asked to share specific solutions they have developed that speak to one or more of these questions with the larger group during the meeting. Initial group activities will directly inform development of use cases that **(1)** model stakeholder concepts of data definition in the context of library collections, **(2)** apply to a range of potential user communities **(3)** introduce optimal library data formats to suit particular use types, **(4)** integrate data sharing practices with descriptive and technical solutions that support them, **(5)** work through varying forms of data access (api, bulk download, torrent), **(6)** gather tools that support collection transformation, **(7)** consider rights issues impeding use of collections as data, and **(8)** articulate a range of data discovery scenarios. Addressing these issues will help establish a shared understanding of the issues at work in this space, and will surface meaningful overlaps and distinctions across various groups invested in the development and use of library collections as data.

Day Two - *Initiate best practices*
During the second day, the group will focus on initiating best practices for resolving challenges attendant to developing, describing, and providing access to computationally amenable collections. By drawing on use cases developed during the first day of the forum, as well as project participant experience in developing and/or making use of collections, the group will generate an initial library collections as data framework geared toward pragmatic collection transformation and documentation, identify methods for making computationally amenable library collections more discoverable through aggregation and other means, and begin to articulate functional requirements that hold potential to support development decisions relative to technical feature integrations with existing initiatives like IMLS's Hydra-in-a-Box. These will not be complete solutions. Rather, forum activity is meant to frame and seed the component parts of the project under the assumption that sustained and heterogenous feedback garnered over the duration of the project, in in-person and virtual contexts, is a requirement to achieve development of

complete solutions. By committing to this ideal, forum participants align the project with systemic community-vested collaboration.

Day Three - *Create participation structure, develop framing questions*
In the final day, the group will provide guidance to the project team on the development of participation mechanisms and framing conversations that will be used post-forum, across virtual meetings and conferences, to encourage iterative community refinement and ultimate finalization of project outputs. Keen attention will be paid to identifying and structuring possible ongoing contributions that will serve to advance the goals of the project in forum participant communities. Participation mechanisms will include platforms like a project website with community annotation function, open Google documents, and web forms. Framing conversations will align with the goal of taking initial project activity fostered by a centralized group of representative experts and transitioning that work to a point where it is strategically decentralized and attuned to the realities and attendant priorities of a larger, nationally dispersed group.

Follow-up Feedback Activities

Within a month of the initial meeting, the draft library collections as data framework created at that meeting, framing questions, collection of use cases, and national forum white paper will be published to the project website. The project website will utilize an annotation feature to allow open community annotation of draft project outputs. The website will be used as a hub to gather feedback and use cases from a broad range of community members. In addition to the annotation function on the project website, the website will provide access to forms that prompt and collect structured feedback aligned with framing questions produced during the national forum, as well as community suggestions for functional requirements and perspectives that inform that development of user stories that can be used to substantiate use cases developed during the national forum. The call for feedback will be made on listservs, Twitter, and with directed requests to groups whose voices may have been underrepresented within the initial forum. This call will be repeated over the following 18 months as the document is iteratively refined and shaped to meet the needs of a broader community.

In addition to the project website and broad outreach activities, the project team will facilitate project workshops at seven conferences, where members of the project team will partner with community members to engage in constructive evaluation of initial project outputs. The workshops will be designed to raise new questions and to help refine or correct project outputs. Workshop activity might point to new approaches, to the need for new tools, or result in calls for new forms of documentation that are more salient to particular audiences. The conveners of each workshop will write up a summary of the outcomes to be shared on the project website, so that the framework and documentation can evolve in a transparent manner over the course of project activity.

Final Publications and Outputs

During a closing meeting of the project at DLF 2018, the project team will present the final set of project outputs: library collections as data framework, use cases and user stories, functional requirements for technical solutions that support library collections as data, methods for making these types of collections more discoverable, and a summative white paper. Depending on the needs that emerge from the community during the year, tutorials for the creation of datasets and specifications may also be published.

3.4 Project Timeline

Phase	Duration	Activities
National Forum Planning	November 2016- March 2017	Frame forum invitation
		Identify forum participants
		Invite forum participants
		Develop National Forum schedule
		Develop National Forum website
National Forum Event	Spring 2017	Develop initial project outputs
		Place initial project outputs on National Forum activity website for public comment and feedback
National Forum output refinement, transition to community inflected outputs	June 2017- March 2018	Digital Humanities - July 2017
		Society of American Archivists - July 2017
		DLF - November 2017
		AHA - January 2018
		MLA - January 2018
		Code4lib - Winter/Spring 2018
		Open Repositories - June 2018
Virtual community solicited refinement and final output production	June - October 2018	Prompt and incorporate last community refinements into final product outputs
Final output presentation	November 2018	DLF - November 2018

4. Diversity Plan [if applicable]

Every effort will be made to diversify participation in the forum, and subsequent activities, with the scope of the work critically shaped by those perspectives as well as prior scholarship that outlines the dangers of re-creating bias in the cultural record via unintended canonization. Accordingly, the project will frame questions of dataset creation, description, contextualization, access, and re-use in ways that allow for development of library collections data that lend themselves to supporting traditionally underrepresented narratives (Earhart 2012). Sustained effort

will be dedicated to sharing and encouraging opportunities for underrepresented groups to play a primary role in shaping project outputs.

5. Project Resources: Personnel, Time, Budget

Personnel

Thomas Padilla, Humanities Data Curator at the University of California Santa Barbara will serve as PI. Thomas is responsible for developing best practices for curating Humanities research data, producing library collections developed for computational use, shaping repository policies and features to encourage computational use of collections, fostering data curation practices across Humanities disciplines, and advancing Digital Humanities services. Thomas publishes, presents, and teaches widely on Humanities data, data curation, and data information literacy.

Laurie Allen, Assistant Director for Digital Scholarship at the University of Pennsylvania Libraries within the Teaching, Research & Learning Directorate will serve as Co-PI. The Digital Scholarship group is responsible for coordinating data curation & management activities, supporting digital humanities and digital methodologies, and undertaking digital publishing initiatives. Before joining the Penn Libraries in February 2016, Laurie worked as Coordinator for Digital Scholarship at the Haverford College Libraries.

Patricia Hswe, Co-Department Head of Publishing and Curation Services at The Pennsylvania State University Libraries, will serve as Co-Investigator. She contributes vision and strategy for a suite of digital scholarship services aligned with the Libraries' mission and core values for teaching, learning, and research. Her chief areas of attention are user services for ScholarSphere, Penn State's institutional repository, which preserves and makes broadly accessible the intellectual assets of the University's faculty, students, and staff; consultation on data management for faculty and students seeking guidance and information on best practices and standards for managing their research data; and product ownership of cultural heritage object collections that have been digitized or are born-digital.

Stewart Varner, Digital Scholarship Librarian at the University of North Carolina, Chapel Hill will serve as Co-Investigator. He earned his Ph.D. in American Studies from Emory University and his MLIS from the University of North Texas. Dr. Varner led the Doc South Data project at UNC which made hundreds of texts available for simple download as plain text in order to encourage their use in text analysis projects. He frequently writes and presents on digital humanities and libraries.

Sarah Potvin, Digital Scholarship Librarian at the Texas A&M University Libraries will serve as Co-Investigator. Based in the Office of Scholarly Communication, she coordinates a portfolio of digital humanities, digital collection, and library technology projects, with attention to platforms, standards, public and open access, and community efforts to promote interoperability. Her research and service work is primarily focused on digital humanities.

Elizabeth Roke, Digital Archivist and Metadata Specialist in the Rose Library at Emory University will serve as Co-Investigator. Primarily focused on preservation, discovery, and access to digitized and born digital assets from special collections, Elizabeth works on a variety of technology projects and initiatives related to repository development, metadata standards, and archival description. Elizabeth is particularly interested in linked data approaches to description.

Time

Phase	Duration	Activities
National Forum Planning	November 2016- March 2017	Frame forum invitation
		Identify forum participants
		Invite forum participants
		Develop National Forum schedule
		Develop National Forum website
National Forum Event	Spring 2017	Develop initial project outputs
		Place initial project outputs on National Forum activity website for public comment and feedback
National Forum output refinement, transition to community inflected outputs	June 2017- March 2018	Digital Humanities - July 2017
		Society of American Archivists - July 2017
		DLF - November 2017
		AHA - January 2018
		MLA - January 2018
		Code4lib - Winter/Spring 2018
		Open Repositories - June 2018
Virtual community solicited refinement and final output production	April - October 2018	Prompt and incorporate last community refinements into final product outputs
Final output presentation	November 2018	DLF - November 2018

Budget

Funding requested for this project is $100,000.00 to cover the cost of organizing and hosting a two and a half day summit at University of California, Santa Barbara. In addition to the summit, funds will support attendance at library and research conferences in order to share as well as iteratively refine *Always Already Computational:*

Library Collections as Data work. Primary costs will be allocated to summit travel, lodging, and meals, and secondary cost will be allocated to conference attendance to facilitate project workshops.

6. Communications Plan

The intended audience for this project is broad, as it aims to develop outputs that are vested in community interest in computational uses of library collections as data. Stakeholders include librarians, archivists, museum professionals, technologists, data practitioners, and researchers from a range of institutional contexts. Given this range, the communications plan is designed to extend outward over the course of two years to engage and incorporate the perspectives of multiple audiences. Forum participants will frame and initiate project outputs and the project team will transition that work over the remaining 16 months to produce a product that is significantly shaped by community input.

Initial national forum meeting outputs will be published to the project website where they will be available for shared annotation and response. The website will be used as a hub to gather feedback and use cases from a broad range of community members. In addition to the annotation function on the project website, the website will provide access to forms that prompt and collect structured feedback aligned with framing questions produced during the national forum, as well as community suggestions for functional requirements and perspectives that inform that development of user stories that can be used to substantiate use cases developed during the national forum. Throughout the project, the project team will also hold occasional twitter chats to answer questions and collect thoughts, and will post updates about the project across social media channels.

Holding structured workshops at disciplinary conferences such as Digital Humanities, MLA, and AHA will help produce targeted feedback from scholarly audiences, while workshops at Open Repositories, DLF, Society of American Archivists, and Code4lib will ensure that the framework is developed in collaboration with the producers of digital collections as well as the creators and maintainers of collection platforms. After each workshop, the project team members and interested workshop participants who attended the conferences will write up and share a summary of the workshop outcomes, and will encourage workshop participants to continue contributing to the project through the website and feedback collection forms. After the final workshop, as the team works collaboratively on refining the framework and accompanying documents, the project team will host at least one virtual feedback session to invite forum and workshop participants, as well as the broader community to participate in the process.

The final results of the project work, including the library collections as data framework, use cases and user stories, functional requirements of technical solutions to support library collections as data, methods for making these types of collections more discoverable, and summative white paper will be presented at DLF 2018.

University of California, Santa Barbara

Always Already Computational: Library Collections as Data

11/2016-11/2018	2016		2017												2018										
	Nov	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**
Planning phase																									
Frame invitation																									
Identify participants																									
Invite participants																									
Develop National Forum schedule																									
Develop National Forum Website																									
National Forum																									
Develop draft framework and documents																									
Publish forum outputs on National Forum activity website for public comment and feedback																									
Publish National forum activity white paper on website for public comment and feedback																									
Feedback and Refinement phase																									
Web feedback collection and outreach																									
Digital Humanities 2017																									
Society of American Archivists 2017																									
Digital Libraries Federation 2017																									
American Historical Association Annual Meeting																									
Modern Languages Association Annual Meeting																									
Code4lib - Winter/Spring 2018																									
Open Repositories 2018																									
Virtual Feedback sessions																									
Final Output Presentation Phase																									
Digital Library Federation 2018																									
Web Publication and dissemination of final version																									
Closure phase																									
Forum Materials added to Penn's Scholarly Commons Repository, linked from website																									

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.

All materials will be assigned a Creative Commons License CC-BY-NC 4.0. The project team is committed to publishing all materials produced in the course of this project under these terms, to the extent legally permissible and subject to any obligations to third parties.

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 University of California, Santa Barbara will not assert ownership rights over digital products produced in the course of this project, to the extent legally permissible and subject to any obligations to third parties. These products will include a website and a collection of documents that will include a white paper, a framework, functional requirements, use cases, and user stories. Published materials will be copyright by the authors but will be made openly and freely available via the University of Pennsylvania's Institutional Repository, ScholarlyCommons http://repository.upenn.edu/. The project website will be

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

The project team does not plan to collection information that involves privacy concerns. The project team will engage a broadly diverse set of stakeholders, and in so doing will make clear the intentions of the project and the commitment to sharing project outputs with the public under the terms of the CC-BY-NC 4.0 license. Where sensitivities are encountered the project team will anonymize collected information.

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.

The project website will be created as a simple website with an annotation feature such as hypothes.is built in to allow for open contribution. When documents are collaboratively authored (during the forum, workshops, and active planning sessions of the project team), they will be drafted in google docs. After each collaborative authoring session is complete (for example, after the forum is over), project team members will be responsible for moving the documents created onto the website as pdf's, markdown, and html documents.

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.

With the exception of google docs, all software used to produce content will be open source, and the work to create the website will be undertaken by project team members who have experience in web development and design.

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, CSV, DOCx, markdown

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).

The project team will have monthly meetings throughout the project timeline, with more frequent meetings as needed in preparation for the forum, for workshops, and for the final presentation. As the larger part of the project timeline will be spent after the forum gathering feedback and refining the documents created earlier, community feedback will help ensure that the project is continuing on track.

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).

During the project, the digital assets will be stored on the project website hosted by the University of Pennsylvania Libraries. At the end of the project, all documents produced for publication through the project website will be stored in the the Penn Libraries ScholarlyCommons website. In addition a web archive of the site will be created and stored in the Penn Institutional Repository as well. The University of Pennsylvania Libraries will continue maintaining the project website for at least one year after the project end. After the first year the site will be maintained for as long as both the Penn Libraries and a majority of project team members agree that it is useful, after which point project content will be removed, and urls

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).

When the content is removed from the project website and stored in the ScholarlyCommons repository, Dublin Core metadata will be attached.

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

N/A

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

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).

All materials produced in the course of the project will be made openly available online, primarily through the project website. As documents are developed by the project team, and in concert with community partners, drafts will be created in Google Docs and transitioned by the project team to the website for further refinement and publication.

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

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.

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?

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).

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.

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).

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?

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

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

   Name of repository:
   URL:

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

Original Preliminary Proposal

Always Already Computational: Library Collections as Data

Michigan State University, University of Pennsylvania, University of North Carolina Chapel Hill, Penn State University, Emory University, and Texas A&M University seek $100,000 for a National Forum Grant under the National Digital Platform funding priority to hold a series of meetings of librarians, technologists, and researchers. These meetings will support development of a strategic approach to developing, providing access to, and encouraging reuse of library collections that support computationally-driven research and teaching in areas including but not limited to Digital Humanities, Public History, Digital History, data driven Journalism, Digital Social Science, and Digital Art History.

Statement of Need

Over the past twenty years a large number of library collections have been digitized. Reel-to-reel tape, photos, and manuscript pages now fill digital repositories throughout the country. Combined with an increasing flow of born-digital items, digital library collections have come to represent a rich community resource for users that seek to find and watch a film, look at a photo, or read a page. Yet a focus on replicating traditional ways of interacting with collections in a digital space does not meet the needs of the researcher, the student, the journalist, and others who would like to leverage computational methods and tools to treat digital library collections as data. For example, a Digital Humanities researcher engages in term frequency visualization, topic modeling, and network analysis across thousands and sometimes even millions of items. Collection managers have no common framework to guide transformation of collections to meet this type of use. Where the few computationally-amenable digital library collections have been developed, they are each prepared, described, and made accessible differently and are typically difficult to discover. Finally, as a technical and social consideration, library repository platforms, both sole institution and aggregative efforts, are not currently developed to enable ready access to collection items at scale, nor do they align collections with a data reuse paradigm. Although the Digital Public Library of America, Stanford University, and DuraSpace have received IMLS funding to build an out-of-the-box digital collections solution, it does not extend to facilitate development of library collections that support computational use.

While approaches to preparing library collections in this vein are nascent, preliminary models have been advanced by project team institutions.[1] Critical thought in this area of work is reflected in an emergent scholarly literature and is an increasingly resonant area of interest within the library community.[2] For example, at Digital Library Federation 2015, a working session led by this project team on "Digital Collections as Data: Re-Packaging, Re-Mixing, and Sharing Collections for New Forms of Scholarship," drew nearly 100 participants. As this work begins to scale to institutions large and small throughout the country, we are at an opportune moment. A gathering of key stakeholders is needed to craft a strategic direction that **(1)** generates pragmatic guidance for local collection transformation efforts that are widely discoverable, **(2)** affords the possibility of integration with other collections, **(3)** informs development of technical features that can be integrated with existing initiatives like IMLS's Hydra-in-a-Box, and **(4)** has the foresight to consider innovative perspectives that hold potential to support development of unanticipated solutions.

Project Plan

The grant will support a two and a half day librarian, technologist, and researcher summit at Michigan State University in Spring 2017. In addition to the summit, grant funds will be allocated to supporting two meetings at

[1] See Michigan State University - http://www.lib.msu.edu/dh/humdata/ , University of North Carolina Chapel HIll - http://docsouth.unc.edu/docsouthdata/ , University of Pennsylvania - http://openn.library.upenn.edu/

[2] Padilla, Thomas. (Forthcoming, 2016) "Humanities Data in the Library: Integrity, Form, Access." D-Lib Magazine. March/April 2016., Varner, Stewart. (Forthcoming, 2016) "*DocSouth Data: Open Access Data for Digital Humanities."* In K. Smith (Ed) *Open Access and the Future of Academic Libraries*. Lanham, MD: Rowman and Littlefield Publishing Group., Padilla, Thomas and Devin Higgins. "Library Collections as Data: The Facet Effect." *Public Services Quarterly*, Vol. 10, Issue 4, 2014.

library and technology conferences and two meetings at research community conferences in order to share, as well as iteratively refine, *Always Already Computational: Library Collections as Data* work. In keeping with the themes that emerged in the National Digital Platform report, this project embraces the notion of "radical and systematic collaborations." On the library side, participant selection will span university, college, and special libraries. On the technologist side, participant selection will span institutions like [redacted] On the researcher/practitioner side, participant selection will span scholars operating in the Humanities and Social Sciences engaged in computational research, as well as leading data-driven journalism practitioners [redacted]. Participant selection will also extend to include experts with complementary expertise from organizations [redacted]. The combined activity of this project will lead to **(1)** development of use cases that inform collection format transformation and documentation, **(2)** collaborative production of a framework that can guide collection manager efforts to transform and document collections to meet computational use, and **(3)** identification of methods for making computationally amenable library collections more discoverable through aggregation and other means.

Relevance to National Digital Platform Priority

With its focus on developing nationally scalable practices that improve the functionality and discoverability of computationally amenable library collections, *Always Already Computational: Library Collections as Data* aligns directly with the National Digital Library Platform funding priority emphasis on, "documentation and system interoperability across digital library software projects." Project activity is predicated on "radical and systematic" librarian, technologist, and researcher collaborations. Conscious and sustained engagement along these axes will help to ensure development of practices and solutions that will guide librarians throughout the country, at institutions large and small, as they approach development of collections that support computationally inflected academic inquiry, professional practice, and creative exploration.

Performance Goals and Outcomes

Project activity will lead to creation of a framework for guiding development of digital library collections that are readily amenable to computational exploration. This framework will make possible immediately pragmatic gains for the wider digital library community as they work to create additional facets of value for the collections under their care. In addition, the framework will extend to inform approaches to enhancing collection discoverability, collection integration with existing initiatives like Hydra-in-a-Box and DPLA, as well as development of technical and social solutions that encourage collection access and reuse.

Personnel

Thomas Padilla, Digital Scholarship Librarian at Michigan State University will serve as Principal Investigator. Laurie Allen, Assistant Director for Digital Scholarship at University of Pennsylvania Libraries will serve as Co-Principal Investigator. Stewart Varner, Digital Scholarship Librarian at University of North Carolina Chapel Hill Libraries, Patricia Hswe, Digital Content Strategist and Head ScholarSphere User Services at Penn State University Libraries, Elizabeth Roke, Digital Archivist at Emory University, and Sarah Potvin, Digital Scholarship Librarian at Texas A&M University are Co-Investigators.

Budget

Funding requested for this project is $100,000.00 to cover the cost of organizing and hosting a two and a half day summit at Michigan State University. In addition to the summit, funds will support two meetings at library and technology conferences and two meetings at research community conferences in order to share as well as iteratively refine *Always Already Computational: Library Collections as Data* work. It is anticipated that the primary cost will be allocated to summit travel, lodging, and meals, and secondary cost will be allocated to conference attendance.