Abstract

Systems Interoperability and Collaborative Development for Web Archiving

As the Web becomes the medium of record for many types of information, web archiving is an increasingly vital function of memory institutions. In furtherance of the National Digital Platform priority of IMLS, this project seeks funding to expand national web archiving capacity by undertaking research that will build a foundation for collaborative technology development, improved systems interoperability, and an Application Programming Interface (API) based model for enhanced access to, and research use of, web archives. The project team consists of Internet Archive (Archive-It) as lead institution partnering with Stanford University Libraries (both DLSS and LOCKSS), University of North Texas, and Rutgers University. This two-year research project will outline successful community models for cooperative technology development work, prototype and test API-based interoperability, and explore how interoperability can enable new access models, improve discoverability, and expand shared digital services.

The practice of archiving web content poses severe programmatic and technical challenges to institutions, given its considerable infrastructure and engineering requirements; the dynamic, evolving nature of the Web; and the sheer volume and complexity of born-digital materials being collected. These high barriers to entry help explain why Archive-It (https://archive-it.org) is the overwhelming web archiving solution of choice. Yet, the ease of entry into web archiving via Archive-It has not spurred investment in local preservation or sustained collaborative technology development.

Given this landscape, the assembled project partners would pursue design-based research and development to prototype API-enabled interoperability and research and test economic and community models to support collaborative development. In the process of this work, and with broad community input, the project will answer a number of key questions, primarily: What are the functional and technical requirements of API-based system interoperability for web archiving and what economic and community models can inform and support such work? This project aims to answer these questions by focusing on defining and piloting an Export API that matches the community-driven success of Archive-It as a national-scale digital platform with key institutions with extensive web archiving, digital library, and web-data research experience.

Project outputs include two open-source Export APIs (by Archive-It and LOCKSS) that will be documented, disseminated, and implemented in at least alpha-level code on multiple platforms by project partners. Project partners will also create open-source local utilities for testing ingest of data using the Export APIs. The project team will publish a report on its research into effective models for collaborative development for web archiving and share its learned lessons on strategies for more effective community building. In collaboration with the community, the project team will also sketch out an array of other promising web archiving APIs for future work. The project will also host a National Symposium on Web Archiving Interoperability (and publish a summary report) to help coalesce a growing, yet undeveloped, community and organize it for better coordination of local systems, research needs, and technology development.

The proposal supports the National Digital Platform funding priority by increasing access to the shared services and infrastructure of an existing platform while freeing local capacity to focus on local services and sustainability. Archive-It's status as an national web archiving infrastructure ensures broad impact and project work will lay the groundwork for future collaborative development and systems interoperability that will expand access, increase local preservation, and improve the discoverability and use of web archives.

Systems Interoperability and Collaborative Development for Web Archiving

Internet Archive (Archive-It), Stanford University Libraries (Digital Library System and Services & LOCKSS), University of North Texas Libraries, Rutgers University

Project Narrative

1. Statement of Need

The primacy of the Web for the publication and dissemination of information has made web archiving an increasingly necessary collecting mode for libraries and archives. The practice of archiving web content, however, poses severe programmatic and technical challenges to institutions, given its considerable infrastructure and engineering requirements; the dynamic, evolving nature of the Web; and the sheer volume and complexity of born-digital materials being collected. In light of the challenges facing the web archiving field, this proposal seeks funding for research on an effective community model for collaborative technology development for web archiving, the prototyping of an application programming interface (API) for export of web archive data, and planning for a longer-term community project that builds on, and sustains, both of those efforts.

The Web is a critical source for the materials that cultural heritage organizations have traditionally concerned themselves with collecting, curating, preserving, and making accessible, such as art, government information, gray literature, institutional legacy, journalism and politics, cultural activity, and scholarly work. In turn, significant public audiences are utilizing archived web materials. For instance, journalists utilize archived web material as a key source for researching articles (Deuze, 2005); researchers and academics are also drawing on archived web material for academic research (Weber, 2012; Bennett, 2005). The Web is also a critical source of novel primary sources for which there may or may not be readily-identifiable historical analogs, such as social and interactive media. Libraries and archives have appropriately responded by archiving the Web in growing numbers – 38% of the organizations responding to the 2013 National Digital Stewardship Alliance (NDSA) Web Archiving Survey had started their web archiving programs within the previous two years and the number of Archive-It subscribers grew 20% in the last year alone (from 2014 to 2015).

These are welcome trends; but a challenge as unavailing as preserving the Web needs as much engagement as possible. Web archiving is difficult in a way that inarguably benefits from a growing number of libraries and archives building these collections. However, web archiving is also hard in a way that demands that the organizations doing it work not merely in parallel but more purposefully together, if the field is to move forward. The desire for collaboration has inspired forums such as the International Internet Preservation Consortium (IIPC), the Society of American Archivists (SAA) Web Archiving Roundtable, and the community of Archive-It Partners. Prior efforts set a foundation for this proposal, but there is a dearth of work with regards to technological interfaces for accessing archived web material. Similarly, there is a need for a stronger community framework to support archiving and subsequent access.

A key area for collaborative technology development in web archiving is systems interoperability, enabled by standardized APIs (Honey and Herring, 2009). Without standard interfaces, individual institutions have had to make extensive downstream customizations of core tools, thus diminishing already-limited collective development resources. The monolithic nature of current web archiving systems constrains the flexibility of libraries and archives to disentangle and enhance discrete functions of the lifecycle and do so with the confidence that the improvements will be reusable by anyone else (Dagger et al. 2007). A comparison to the relative progress made in other areas of digital library development is instructive; the most mature technologies (e.g., open source projects such as Fedora, Hydra, and Blacklight) have been developed collaboratively and with community and interoperability in mind.

Evidence of the need for collaborative technology development for interoperable web archiving systems is supported by three observations about the state of the field: the web archiving community needs to be able to re-tool more efficiently to keep pace with the Web; fractional resourcing and the concomitant internal focus of program development have inhibited specialization that could lead to more breakthrough innovations in scalability and access; and a huge part of the web archiving community lacks a framework for participation in collaborative technology development.

The Web is constantly growing, changing, and becoming more dynamic, at a pace that does not allow for backlogging its capture and processing (Tien, 2013; Agata et al, 2014). A more deliberate focus on the community in technology development will lead to more sustainable projects, and an API-based architecture will allow organizations interested in specific functionality to focus on and contribute to isolated components. The web archiving community in general, but smaller operations in particular, does not have an effective framework for collaborating on technology development, yet most of the new and future members in the community are at the low-end in terms of resource allocation. These organizations are typically Archive-It users with fractional staffing who have so far had little opportunity to inform the direction of technology development (NDSA Web Archiving Survey, 2013). An effective community model will enlist this “fat, long tail” of web archiving stakeholders to inform requirements and specifications, support technology development as a group, and free up local capacity for the exploration of new access models and other experimentation. In doing so it will further the National Digital Platform goals of increasing access to the shared services and infrastructure of an existing platform while freeing local resources to focus on new services and sustainability. This will lay the groundwork for future collaborative development and interoperability that has the potential to expand access and discoverability, increase local preservation, and build the community.

2. Impact

Broadly, this project will have wide ranging impact in terms of improving Web archiving technology for librarians and archivists, forming a sustainable community for collaborative development, and improving the ability of digital libraries to deliver content. The project team will research effective community models to provide an appropriate framework for this work and put it into practice for the joint development of prototype APIs for export of web archive data. In addition to testing a more broadly-envisioned web archiving API

ecosystem, a standalone export API has the potential to more immediately improve the limited participation in distributed local preservation and help advance nascent explorations into new access tools.

The concentration of development work on enhancing Archive-It maximizes the potential impact. With more than 350 partners, Archive-It is the most popular web archiving solution for the vast majority of U.S. web archiving institutions and its share of the market continues to grow over time (NDSA Web Archiving Survey Report, 2013). Archive-It, however, is not a local preservation solution. Notwithstanding this caveat, most organizations do not download their data, even for local storage, let alone for preservation (NDSA Web Archiving Survey Report, 2013). Machine interfaces for exporting and then importing data locally, such as will be developed in this project, will make this easier, laying the groundwork for other potential APIs and facilitating integration with other local discovery and access services.

2.1 The Impact of Improving Web Archiving Technology for Librarians

Providing a blueprint for systems interoperability will capitalize on unused fractional development resources across institutions. Such consolidation of resources and strategic development will have significant positive impact by improving digital archiving services on a national scale by allowing easier integration of their web archives with other preservation and collection management systems such as DPN and ArchivesSpace. The involvement, as partners and advisory board members, of the research and web science communities ensure even broader exposure to, and benefit from, the access mechanisms made possible because of the project's research and testing of standard programming interfaces. Catalyzing these types of new access models will help web archiving programs better demonstrate use and better advocate for the value of their web collections.

2.2 The Impact on the Web Archiving Community

The project will allow for broad external input and guidance though a number of mechanism that include: the formation of a Technical Working Group to explore APIs; feedback from the national and international community via partners' roles (and coordination of affiliated meetings) in professional organizations like IIPC, NDSA, Preservation and Archiving Special Interest Group (PASIG), and Coalition of Networked Information (CNI); the hosting of a National Symposium; and through open conference calls, communications, and engagement with regional working groups and researchers.

2.3 The Impact of Improving Access for Researchers

An export API will provide improved researcher access to web archive data. The dominant access paradigm of Wayback "replay" of archived web content may lead prospective researchers to suppose that more data-centric services are not available or feasible. Those researchers who inquire about research data services face significant challenges and may additionally have to build or cobble together custom solutions to work with the data in the way it is provided. Standardized APIs for data delivery will engender more consistent expectations of the access services provided by web archives and reduce the amount of custom tooling that researchers will have to maintain or deploy. As a JISC report noted, "tools should be sharable and easy for researchers and librarians to implement." (JISC,

Researcher Engagement with Web Archives, 2010). This project's research and testing will help operationalize the community's ability to provide such tools.

2.4 Project Outcomes

The project will measure success in a number of ways both quantifiable and qualitative. Quantifiable metrics and performance indicators include an increase in the number of non-partner staff participating in, or contributing to, project activities via listservs, blog posts, conference presentations, formation of working groups, and the inclusion of the project's research focus in national meeting agendas. Other quantifiable impact metrics include non-project institutions expressing interest in helping define interoperability specifications and use cases and doing API testing, expressions of interest from affiliated systems (such as Archivematica, ArchivesSpace, DPN, others) in pursuing better interoperability via APIs, and increased local preservation of WARC files by Archive-It partners. More qualitative metrics include an increase in incidence of web archiving technology sessions in digital library (but not web archiving-focused) communities, the emergence of other APIs informed by the project's research, proposed partnerships from the researcher community on tools and interfaces possible via programmatic API access, and propagation of the project's work in the international web archiving community.

Tangible, formal products will result from this research project. These include documented guidelines for a community and economic model for sustained collaborative technical development for web archiving, a research paper outlining a broader API-based web archiving ecosystem, and a white paper summarizing the results and outcomes of a national symposium focused on community building, interoperability, and access. Technical tangible projects emerging from this research include tested export APIs in two environments and three "import" utilities testing local consumption of those APIs. Community building, both on a technical level via research and on a broader national level via a national symposium, will provide value by helping coalesce a growing, yet undeveloped, community and organize it for better coordination of local systems, research needs, and cooperative development. Project benefits will be sustained via continued technical work and API-focused conference working groups beyond the project.

3. Project Design

Goals and Objectives

The project is designed around three primary research questions:

1. *What are the attributes of a community model that can support sustainable and broad-based collaborative web archiving technology development?* The community modeling aspect of the research has a twofold intention: exploration and then demonstration of a suitable framework for collaborative technology development for interoperable web archiving systems. The project will be conducted in the spirit of the community it hopes to catalyze, e.g., with transparent processes and communications, openness to participation, and structures to lower the barrier to contributing. Part of the community model is also to understand how the economics of web archiving affect organizations' capacity to participate in collaborative technology development.

2. *What are the community needs and possibilities for the planned open API to facilitate transfer of web archive data between distributed systems and what other prospective APIs does it point to?* Several discrete and broad-impact use cases informed the choice of starting with the export API: distributed preservation and use of web archive data generated through Archive-It; getting web archive data into or out of LOCKSS for preservation or access uses; and delivering data to researchers. This is not an exhaustive accounting of the potential of an export API and certainly does not reflect the possibilities of a more broadly-realized API ecosystem. Other use cases, new operational models, and hybrid architectures may be enabled. These possibilities will shape a roadmap for work on web archiving APIs beyond the duration of the grant.

3. *How can better interoperability of web archiving systems support new forms of access and research use?* The web archiving lifecycle and the internal operational needs of web archiving programs are comparatively well-understood, relative to the data needs of researchers. The research value of web archives will not be unlocked by merely providing a mechanism for researchers to obtain raw web archive data. Understanding both the diversity and the distribution of different research disciplines' web archive data requirements will inform what features may be encompassed by the export API that is the subject of this grant and what may be queued up for future work.

Project Activities

The project will pursue design-based research. The project team will iteratively refine both the collaboration framework and the prototype API to optimize for effective participation and API utility. The starting point of this process is learning more about and from the web archiving community. This means answering such questions as: Who are the current and prospective members of the community? What is the range of shared and divergent interests and objectives for stakeholders of different types (e.g., small organizations, big organizations, unaffiliated enthusiasts, curators, developers, researchers, legal deposit institutions, archives, libraries, etc.)? How can those interests best be served by a more interoperable web archiving ecosystem? How can different stakeholders be enlisted to help realize that ecosystem? Information will be gathered and use cases solicited in direct dialogue with the web archiving community and by mining the research literature. Key strategies for community engagement will be technical experts invited to participate in the API Technical Working Group meetings and the cross-section of representative stakeholders invited to participate in the national symposium event (see appendix).

Feedback from this project will inform the attributes of the collaboration framework most likely to engender long-term participation for the larger initiative beyond the project group. It will also ensure that the API Technical Working Group's initial specification and subsequent refinements of the export API for web archive data are based on tangible and, hopefully, heretofore unconsidered use cases. The project team will implement the prototype API as part of diverse systems, both as a demonstration of the power of interoperability and to create tools that may be used for identical purposes in multiple environments throughout the community. Specifically, Internet Archive will implement the export functionality of the API for Archive-It and SUL will implement the same for LOCKSS. SUL, Rutgers University, and UNT will also implement tools for the

consumption and ingest of web archive data from the export API. They will then exercise the Archive-It and LOCKSS export APIs to test the facility of the data exchange and look for opportunities for improvement. At the conclusion of the two-year project, the export API will have been documented, disseminated, and implemented in at least alpha-level code on multiple platforms. The project team will write up a report on its research into effective models for collaborative technology development for web archiving and share its learned lessons on strategies for more effective community building and engagement. In collaboration with the community, the project team will also sketch out an array of other promising web archiving APIs for future work. The project will feature a national symposium at its midpoint to convene the web archiving community, demonstrate the benefits and status of project work, seed a community for Year 2 and post-grant cooperation, and to reorient the community's focus towards improved access models and tools, collaborative development, and systems integration.

Roles and Commitments

The product of the project is in some sense its method, so all of the partners will work to seed the collaboration that will be continued and expanded. Each partner will also take on additional roles and commitments. As a foundational institution and preeminent service provider in the web archiving community, Internet Archive will lead the community building, outreach, and canvassing efforts. Internet Archive will also be the primary organizer and host for the national symposium, provide research support for the community modeling, generate API requirements, help to co-design the export API, and implement and test the export API on Archive-It. As a founding member of the Hydra and Blacklight communities, SUL will lead the research on community models as well as participate in community building, outreach, and canvassing efforts. SUL will also help to generate API requirements, co-design the export API, implement and test both export and import functionality in LOCKSS, and implement and test import of web archive data from both Archive-It and LOCKSS into the Stanford Digital Repository (SDR). With a program with significant technical experience in web archiving, UNT will help to generate API requirements and will implement and test import of web archive data into local systems. With a research group versed in working with web archive data, Rutgers University will help to ensure that researcher use cases are centrally considered in the specification of the export API. Rutgers University will help to generate API requirements and will implement and test import of web archive data from Archive-It into local research data infrastructure.

Preliminary Work

While perhaps not specifically anticipating this project, there are a number of intersecting initiatives and activities. The successful collaborative development in the Hydra, Fedora, and Blacklight communities will be mined for insights into bootstrapping an analogous community framework for web archiving. On the technology development side, Internet Archive has developed multiple APIs to improve access services and interoperability with distributed systems. As part of a grant from the Mellon Foundation, Internet Archive and New York University (NYU) will partner to build an API to facilitate integrated presentation of multimedia content preserved in NYU's repository within the Archive-It access interface. Internet Archive has a public API exposing the underlying index of

archived links in Wayback Machine and Internet Archive and SUL staff have also participated in the IIPC-funded Memento Aggregator and Memento Profiling projects.

A growing number of workshops and programs are focused on building relationships between web archiving organizations and researchers, including the British Library's BUDDAH, the L3S Research Center's Alexandria (both advisory board members), and Rutgers University's National Science Foundation grant (#1244727) for ArchiveHub that resulted in the WIRE Workshop. The IIPC Access, Harvesting, Preservation, and a prospective new API-specific Working Groups present interest-segmented communities keen on addressing specific web archiving challenges. While these dispersed or localized efforts lay preliminary groundwork and will inform this project, they have lacked coordinated development efforts, been internally focused, or were developed with marginal community input – all characteristics this proposal seeks to address through better collaboration between people, institutions, and systems. The project partners will test the applicability of its approach both through local ingest of web archive data served via the export API from both Archive-It and LOCKSS, through project activities around collaborative development, and through solicitation of broader web archiving community and researcher feedback to inform its iterative research and development.

4. Project Resources: Personnel, Time, Budget

The two-year project will include personnel from all four partner institutions, with Internet Archive and SUL taking lead roles in project work and outcomes. Internet Archive, as the lead institution, will take overall responsibility for the grant. The requested IMLS funding will primarily subsidize technical work. This includes technical research and the development of the candidate API at Internet Archive and sub-contracts to SUL for API development and Rutgers and UNT for contributing to technical specifications and local API testing. Grant funds will also be used to cover travel costs for the Technical Working Group and provide travel stipends to attendees of a national symposium. Additional funds will contribute to documentation, project management, and community building.

Project Leads

Internet Archive: **Jefferson Bailey**, *Director of Web Archiving Programs*, ([redacted]) will serve as the project's Project Director and will oversee all aspects of the grant work and coordination between partners, including research, technical work, and community building activities. At Internet Archive, Jefferson oversees Archive-It, web archiving programs, research services, and collaborative projects, including IA's current participation in Mellon and IMLS funded work. He has been PI on IMLS-funded grants, has experience in digital preservation and archives, and serves on the IIPC Steering Committee. **SUL**: **Nicholas Taylor,** *Web Archiving Service Manager* ([redacted]), will help gather use cases, generate requirements and specifications, and design, document and test the project APIs. He will contribute to the community modeling efforts, serve as SUL's local project manager, and support the Project Director. He has five years of experience in research libraries managing and supporting web archiving services and technologies, serves as Co-Chair of the IIPC Access Working Group, and is an active participant in NDSA and SAA.

Technical and Research Contributors
Internet Archive: **[to be hired, position description in appendix]**, *Senior Software Engineer* ( ) will be responsible for all Internet Archive technical development including determining technical specifications, functional requirements, data modeling, development of project API, technical support for testing, and iterative coding. *Engineering Project Manager*, ( ) will manage Internet Archive's engineering work on the project. **SUL**: *Tom Cramer*, Chief Technology Strategist and Associate Director of Digital Library Systems, ( ) will help gather use cases, help specify the export APIs and a sketch of the larger overall API framework, translate the specifications to technical work, and contribute to the economic and community modeling efforts. He is a founder of the Hydra Project, and active contributor to Blacklight, and is Co-Director of PASIG and on the DuraSpace Board of Directors. *David Rosenthal*, LOCKSS Chief Information Scientist, ( ) will vet requirements, participate in the specification of the export API, and develop software to manifest the API and/or test it in local systems. He is founder of the LOCKSS Program, a distinguished computer scientist, and has written widely on digital preservation economic models. **UNT**: *Mark Phillips*, Assistant Dean for Digital Libraries, (included below) will coordinate UNT's role in the project, including technical specifications and testing the project's export API. He rebuilt UNT's digital library system and has lead multiple IMLS-funded digital library projects. **Rutgers University**: *Matthew Weber*, Assistant Professor of Communications, (included below) will coordinate Rutgers' testing of the project's export API from a research perspective. His work includes large-scale, data-driven study of web archives and he manages the NSF-funded "Utilizing Archival Resources to Conduct Data-Intensive Internet Research."

Project Management Contributors
Internet Archive: *Lori Donovan*, Senior Program Manager, Archive-It, ( ) will manage the project's national symposium, help with project communications and reporting, coordinate the participation of the Archive-It community, and contribute to project management. Lori has seven years experience working on Archive-It and in technical product management, community building, outreach, and education. *Jacques Cressaty*, Director of Finance, will manage the grant finances and has overseen all accounting at Internet Archive since 2001.

Additional Project Engineering and Testing:
SUL: .
UNT: .
Rutgers: .
Costs and time commitments are detailed in partner subcontract documents in Appendices.

Technical Working Group: Core project staff and engineers, plus Robert Sanderson (Technical Collaboration Facilitator, SUL (Memento Project)), Andy Jackson, (Web Archiving Technical Lead, British Library), Stephen Abrams (Associate Director, CDL,

UC3). [redacted] Costs are detailed in the Budget Justification.

Advisory Board: Helen Hockx-Yu (British Library), Wolfgang Nejdl (L3S Research Center), Ed Fox (Virginia Tech), Kent Norsworthy (University of Texas), Jimmy Lin (University of Maryland). The Advisory Board will evaluate project work twice a year via conference call or webinar. No funds are requested.

Budget

The Budget and Budget Justification provides a detailed summary of requested grant funds and cost sharing for project activities.

5. Communications Plan

The project's communication plan builds on the existing leadership roles of the project partners in the national and international web archiving community and leverages those roles to broadcast results, solicit feedback, and build a base for sustained community participation. The four key audiences are the national (and, by extension, international) web archiving community; the broader library, archives, and curatorial communities; digital library developer groups; and finally, the researchers and web scientists utilizing web archives and desiring new access tools. The communication plan will reach these audiences through numerous methods: conference presentations; a national symposium; ongoing participation in professional working groups; use of open communication channels including a wiki, listserv, and public conference calls; open-access publication of research and open-source code; and outreach via social media, newsletters, and blogging.

A major aspect for communicating to key audiences will be through presentations at professional conferences, such as IIPC, CNI, SAA, Code4Lib, iPres, PASIG, JCDL, and CurateGear. Archive-It and SUL convene their own conferences as well: the annual Archive-It Partner Meeting (held in conjunction with SAA) and SUL's LDCX conference. These will provide additional opportunities to publicize project work, solicit focused feedback, and build community. The grant's project plan also includes both a Technical Working Group to focus on API specifications and a national-scale symposium to involve the broader web archiving community in defining the needs and use cases that will inform project research outcomes. Meeting notes and summary reports will be published from both activities and will be distributed via the partners' websites, blogs, and social media channels. The project team will also coordinate with less formalized communities, including the self-organized Archive-It Regional User Groups, SAA Web Archiving Roundtable, Web Archives for Historians group, and others.

Openness and transparency of project activities, outcomes, and technical work is not just a key ethos of this specific project but a core value to building collaborative communities and partnerships across institutions and user groups. The grant team will use open and neutral online platforms such as Google Groups to coordinate community discussion. An online wiki space will be created for open access to both formal and informal project documentation, including technical papers, meeting and call notes, and research work. Partners will also make use of institutional and third-party (NDIIPP, IIPC, AHA) blogs, newsletters, social media accounts (e.g., Twitter, Facebook). The group will hold quarterly

open conference calls or webinars to report the project's progress. The engagement of target communities will be measured by the number of participants on open calls, Google Group and listserv membership, applicants and attendees at the national symposium (and a follow-up survey), and acceptance and publication of conference presentations and blog posts. All software and technical products created by the project will be released under open-source licenses and published on Github, SourceForge, or other code-sharing platforms. The project will publish technical documentation on APIs, systems interoperability, and data modeling; a final white paper on community and economic models for technical collaboration in web archiving; a summary report from the symposium; and meeting minutes. All publications will be released under Creative Commons BY-NC-SA license. Jefferson Bailey and Lori Donovan (Internet Archive) and Nicholas Taylor and Tom Cramer (SUL) are responsible for coordinating communication, outreach and dissemination. Other project staff will contribute to authorship and presentation of talks, blogs, and publications.

6. Sustainability

The aim of the project is to create not just a community model for sustainable collaborative web archiving technology development, but also a community that begins to resemble it. The practical, objective-oriented collaboration between the project partners will naturally form the seed of this community and is being undertaken with the belief that it will have broad utility across the community, leading to sustained efforts towards further collaboration and interoperability. For all of the documented needs presented above and the potential upsides, there is good reason to think that other organizations will perceive extensive value to being involved in project work both while it is being undertaken and well after the conclusion of its grant-funded activity. In modeling the success of other communities and collaborations for wider adoption, the project will sow the seeds for broader sharing and coordination between institutions for all aspects of the web archiving lifecycle, from appraisal and acquisition, to management, preservation, and use.

The project participants' status as leaders in the field of web archiving, digital libraries, and research use of web collections ensures the broad visibility and impact of project outcomes. Project work will help define an agenda around collaborative development, APIs, and research access that will guide the ongoing work of professional organizations, other archiving institutions, and affiliated systems and services around preservation and collection management. Research outcomes and tested APIs and tools from the project will set the stage for these outcomes to catalyze a sustained community focus on scaling web archiving. Sustainability, in many ways, depends on same goals and outcomes this project aims to achieve: expanded use of shared and interoperable systems, increased efficiencies around infrastructure and technical development, improved access, and an overall advancement in the ability of cultural heritage institutions to archive the web.

Schedule of Completion
January 1, 2016 - December 31, 2017

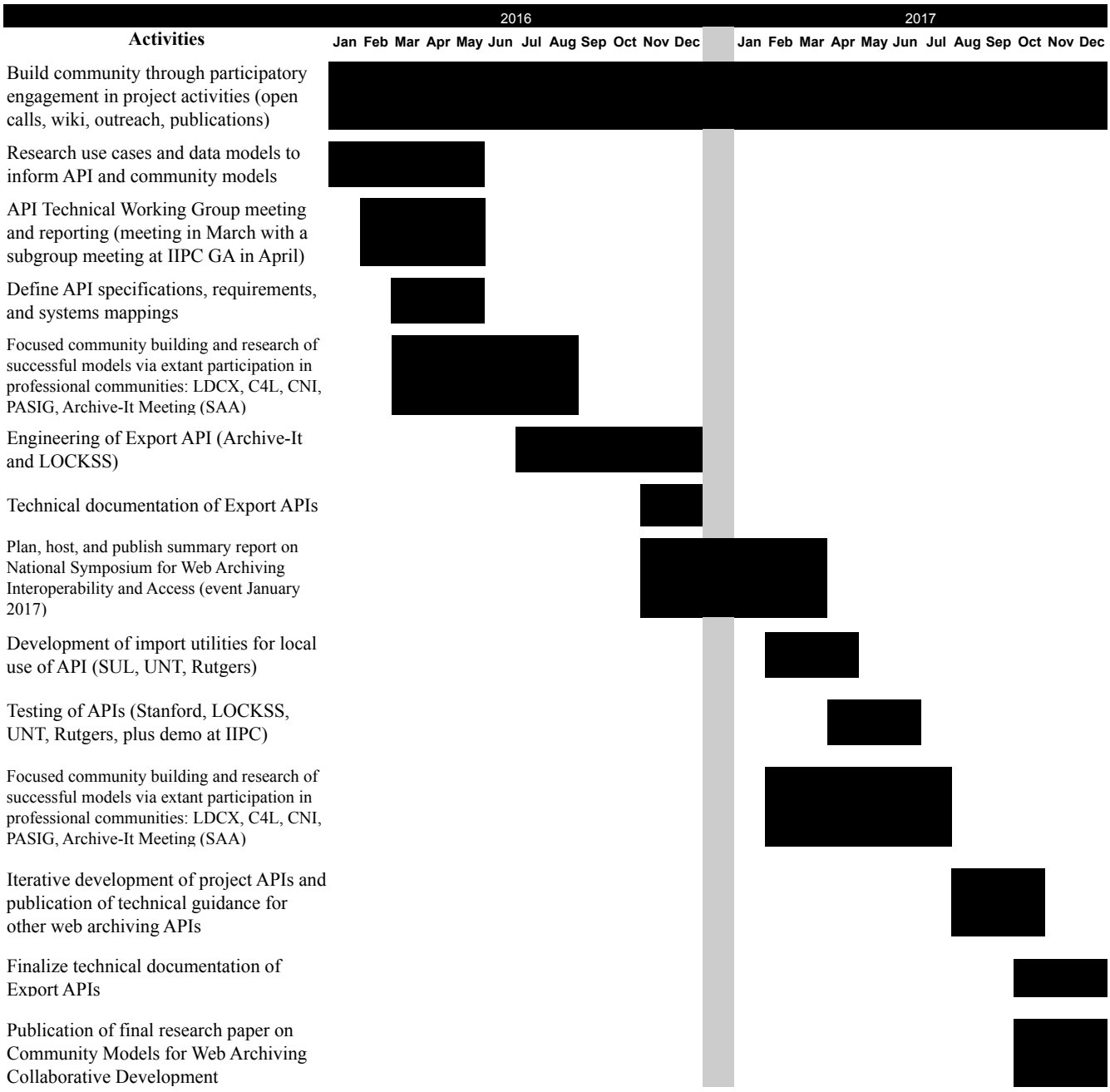
The project will be organized around four six-month phases. Some of the project activities will happen concurrently and some, like community building and outreach, will place across the timeline of the project. Each phase will also have a distinct set of goals and focused areas of work.

Phase one of the project will focus on information-gathering through interaction with the community and drafting the specification for an Export API. The project team will capitalize on the succession of relevant meetings (e.g., IIPC, LDCX, code4lib, CNI, PASIG) to perform outreach, canvas use cases, and enlist prospective contributors interested in the larger initiative. This community building work will complement the research also occurring during this phase on community models and API requirements. The Technical Working Group will meet at least twice, at a dedicated meeting in March and a breakout meeting at the IIPC General Assembly, working toward an initial specification of an Export API.

The community building effort continues in ***phase two***, most conspicuously with the SAA Annual Meeting (including the Web Archiving Roundtable) and the Archive-It Partner Meeting, but also with ongoing open communications about the project status, solicitation of feedback, and engendering of interest. The project team will also complete planning for the National Symposium for Web Archiving Interoperability and Access, to take place at the start of phase three. Meanwhile, with the initial specification of an Export API having been completed in phase one, Internet Archive and SUL will implement it for Archive-It and LOCKSS, respectively. The project team will also create technical documentation for the API.

In ***phase three***, the project team will follow on the implementation of the Export API prototypes with development and testing of compatible import utilities by SUL, UNT, and Rutgers University, allowing for transfer of web archive data from Archive-It and LOCKSS into local systems. The project team will demo the operation of the Export API at the IIPC General Assembly and, as in phase one, take advantage of conferences in the first half of the year to publicize progress and rally the community. Toward this end, phase three will also feature a dedicated National Symposium for Web Archiving Interoperability and Access, planned by the project team and funded by the grant.

Phase four will allow for further iterative improvement of the Export API, based on the most recently-received feedback and in consideration of a more broadly-envisioned web archiving API ecosystem. After finalizing documentation for the Export API, the project team will sketch out high-level designs for other potential web archiving APIs either suggested by the work on the Export API or identified by the community. Critically, they will also publish a final research paper on community models for web archiving collaborative technology development.


2016
2017
Activities
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Build community through participatory engagement in project activities (open calls, wiki, outreach, publications)
Research use cases and data models to inform API and community models
API Technical Working Group meeting and reporting (meeting in March with a subgroup meeting at IIPC GA in April)
Define API specifications, requirements, and systems mappings
Focused community building and research of successful models via extant participation in professional communities: LDCX, C4L, CNI, PASIG, Archive-It Meeting (SAA)
Engineering of Export API (Archive-It and LOCKSS)
Technical documentation of Export APIs
Plan, host, and publish summary report on National Symposium for Web Archiving Interoperability and Access (event January 2017)
Development of import utilities for local use of API (SUL, UNT, Rutgers)
Testing of APIs (Stanford, LOCKSS, UNT, Rutgers, plus demo at IIPC)
Focused community building and research of successful models via extant participation in professional communities: LDCX, C4L, CNI, PASIG, Archive-It Meeting (SAA)
Iterative development of project APIs and publication of technical guidance for other web archiving APIs
Finalize technical documentation of Export APIs
Publication of final research paper on Community Models for Web Archiving Collaborative Development

DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction:
IMLS is committed to expanding public access to IMLS-funded research, data and other digital products: the assets you create with IMLS funding require careful stewardship to protect and enhance their value. They should be freely and readily available for use and re-use by libraries, archives, museums and the public. Applying these principles to the development of digital products is not straightforward; because technology is dynamic and because we do not want to inhibit innovation, IMLS does not want to prescribe set standards and best practices that would certainly become quickly outdated. Instead, IMLS defines the outcomes your projects should achieve in a series of questions; your answers are used by IMLS staff and by expert peer reviewers to evaluate your proposal; and they will play a critical role in determining whether your grant will be funded. Together, your answers will comprise the basis for a work plan for your project, as they will address all the major components of the development process.

Instructions:
If you propose to create any type of digital product as part of your proposal, you must complete this form. IMLS defines digital products very broadly. If you are developing anything through the use of information technology – e.g., digital collections, web resources, metadata, software, data– you should assume that you need to 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
☑	New software tools or applications	Part III
☐	A digital research dataset	Part IV

PART I.

A. Copyright and Intellectual Property Rights

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 copyright or intellectual property status of the content you intend to create? Will you assign a Creative Commons license to the content? If so, which license will it be? http://us.creativecommons.org/

The multiple research reports and technical documentation produced as part of this project will released under the Creating Commons Attribution 4.0 (CC-BY-NC-SA) license.

All software and code developed as part of the project will be released under open source licenses, including GNU-GPL, Apache 2.0, and other open-source licenses.

A.2 What ownership rights will your organization assert over the new digital content, 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 of the digital resources.

The project will not be creating new digital content, only software, documentation, and research reports as listed in A.1. We will notify users of these digital resources via a number of methods, including publications (institutional blogs, a project wiki, journal articles), presentations at multiple conference (IIPC, SAA, LDCX, CNI, others), listservs, and newsletters.

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.

This project will not create any content or products which involve privacy concerns, require obtaining permissions or rights or raise any cultural sensitivities.

Part II: Projects Creating Digital Content

A. Creating New Digital Content

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

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

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, pixel dimensions).

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).

B.2 Describe your plan for preserving and maintaining digital assets during and after the grant period (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: Storage and publication after the end of the grant period may be an allowable cost.

C. Metadata

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, preservation). 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).

C.2 Explain your strategy for preserving and maintaining metadata created and/or collected during your project and after the grant period.

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

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).

D.2 Provide URL(s) for any examples of previous digital collections or content your organization has created.

Part III. Projects Creating New Software Tools or Applications

A. General Information

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

Though this is a research project proposal, the project's research includes the development and testing of a number of new tools. These include:

1) Application Programming Interfaces (APIs): Both Internet Archive (Archive-It) and LOCKSS will create an Export API to test the project's research into an API-based model of systems interoperability for web archiving. These APIs will facilitate the programmatic delivery of web archive data between distributed systems.
2) Local utility tools: For the project partners testing these Export APIs, they will create small utilities to facility the ingest into their local systems of data served via the APIs. These local utility will be done in multiple programming languages (Python, Ruby) to contribute to API testing.

The major function of the APIs is to allow for programmatic transfer between applications of web archive data in order to facilitate improved use and preservation of web archives and to test a model for API-based interoperability for web archiving. The intended audience for these APIs and ingest utilities include any institution currently building web archives and researchers and developers seeking more programmatic access to web archives for scholarship or tool building.

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

There is currently no API for WARC files (the ISO preservation format for web archive digital objects) or for programmatic access to this type of web archive content. Internet Archive has a CDX Server API (https://github.com/internetarchive/wayback/blob/master/wayback-cdx-server/README.md) which provides access to the index of the Wayback Machine. However, this API provides only basic information about URL captures and no access to the files themselves. Archive-It enables manual download of preservation files via a browser or bulk download, but no way to query or provide programmatic access to content. The APIs developed and tested as part of this project will allow programmatic access to web archive preservation files for institutions that have web archives content harvested using Archive-It and LOCKSS and will provide specifications and functional/technical requirements for others looking to create APIs for web archiving systems.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your new digital content.

The programming languages used will be Python, Ruby, Java, Perl, and Unix shell (bash) scripting.

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

The Export APIs are designed for interoperability and will extend both Archive-It and LOCKSS but will also provide a framework for adoption by other web archiving systems. APIs will be designed to interoperate more broadly with existing preservation/respository systems such as Fedora, DPN, Rosetta, and others by offering a standard, documented endpoint for access to web archive digital objects.

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

There will be no software or system dependencies necessary to use the APIs.

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

Development documentation will make use of the Jira issue and project tracking software and the project will maintain both a private and public wikispace (Confluence), hosted by Internet Archive for sharing development work across projects. A Google Group will be created to enable broader community involvement in project activities.

Documentation will include functional/technical requirements research, systems architecture/mappings, and research and technical working group meeting notes. Technical documentation and user manuals will be created for the APIs and will be freely available via the public wiki spaces and on the Github code-sharing platform.

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

Project partners have created dozens of software tools and systems including widely-adopted open-source tools for web archiving and digital libraries.

Internet Archive: https://github.com/internetarchive
Stanford: https://github.com/sul-dlss
LOCKSS: http://sourceforge.net/projects/lockss/
UNT: https://github.com/unt-libraries

C. Access and Use

C.1 We expect applicants seeking federal funds for software or system development to develop and release these products as open source software. What ownership rights will your organization assert over the new software or system, and what conditions will you impose on the access and use of this product? Explain any terms of access and conditions of use, why these terms or conditions are justifiable, and how you will notify potential users of the software or system.

Partners will release all project software and code with an open-source license, such as Apache 2.0, GNU GPL, or BSD licenses.

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

All software and tools developed will be publicly available to users via Github, SourceForge, Google Code, or a similar public code-sharing platform.

Part IV. Projects Creating Research Data

1. Summarize the intended purpose of the research, 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 research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity already been approved? If not, what is your plan for securing approval?

3. Will you collect any personally identifiable information (PII) about individuals or proprietary information about organizations? If so, detail the specific steps you will take to protect such information while you prepare the research data files for public release (e.g. data anonymization, suppression of personally identifiable information, 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 will you capture or create along with the dataset(s)? What standards or schema will you use? 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 research activity?

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?

References

Agata, T., Miyata, Y., Ishita, E., Ikeuchi, A., & Ueda, S. (2014, September). Life span of web pages: A survey of 10 million pages collected in 2001. In Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on (pp. 463-464). IEEE.

Bennett, W. L. (2005). Social movements beyond borders: understanding two eras of transnational activism. Transnational protest and global activism, 203-226.

Dagger, D., O'Connor, A., Lawless, S., Walsh, E., & Wade, V. P. (2007). Service-oriented e-learning platforms: From monolithic systems to flexible services. Internet Computing, IEEE, 11(3), 28-35.

Deuze, M. (2005). Towards professional participatory storytelling in journalism and advertising. First Monday, 10(7).

Dougherty, M., Meyer, E.T., Madsen, C., van den Heuvel, C., Thomas, A., Wyatt, S. (2010). Researcher Engagement with Web Archives: State of the Art. London: JISC. 2010.

Honey, C., & Herring, S. C. (2009, January). Beyond microblogging: Conversation and collaboration via Twitter. In System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on (pp. 1-10). IEEE.
Chicago

Tien, J. M. (2013). Big data: Unleashing information. Journal of Systems Science and Systems Engineering, 22(2), 127-151.

Weber, M. S. (2012). Newspapers and the Long-Term Implications of Hyperlinking. Journal of Computer-Mediated Communication, 17(2), 187-201.

Relevant Online Material

2013 National Digital Stewardship Alliance (NDSA) Web Archiving Survey Report
http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_USWebArchivingSurvey_2013.pdf

Comparison of Archive-It Subscriptions from 2014 to 2015
See: https://web.archive.org/web/20140521161952/https://www.archive-it.org/explore
versus https://web.archive.org/web/20150518082657/https://www.archive-it.org/explore

Discussion of NetarchiveSuite and Web Curator Tool upgrades related to Heretix:
See https://netarkivet.statsbiblioteket.dk/suite/AssignmentHarvester1 and
http://sourceforge.net/p/webcurator/enhancements/110/

A Snapshot of the U.S. Web Archiving Landscape through the 2013 NDSA Survey Report
http://www.slideshare.net/nullhandle/a-snapshot-of-the-us-web-archiving-landscape-through-the-2013-ndsa-survey-report/

Archive-It Storage and Preservation Policy
https://webarchive.jira.com/wiki/display/ARIH/Archive-It+Storage+and+Preservation+Policy

Announcement of California Digital Library and Archive-It Partnership
http://www.cdlib.org/cdlinfo/2015/01/14/announcing-a-new-partnership-california-digital-library-uc-libraries-and-internet-archives-archive-it-service/

Rutgers University ArchiveHub Initiative
http://archivehub.rutgers.edu/

WIRE Workshop Documentation
https://wp.comminfo.rutgers.edu/nsfia/

Internet Preservation Consortium (IIPC)
http://www.netpreserve.org

Society of American Archivists (SAA) Web Archiving Roundtable
http://www2.archivists.org

Archive-It Partners
http://www.archive-it.org/explore/?show=Organizations

Original Preliminary Proposal

Systems Interoperability and Collaborative Development for Web Archiving

Project Director: Internet Archive, Archive-It (Jefferson Bailey, Kristine Hanna) **Project Partners**: Stanford University Libraries, DLSS (Tom Cramer, Nicholas Taylor), Stanford University Libraries, LOCKSS (Victoria Reich, David S.H. Rosenthal), Rutgers University (Matthew Weber), University of North Texas (Mark Phillips) **Advisory Board**: Wolfgang Nejdl, (L3S), Ed Fox (Virginia Tech), Kent Norsworthy (University of Texas), Jimmy Lin (University of Maryland), additional pending.
Estimated budget: $350,000 for a 2-year project

Abstract: The project team seeks funding to expand national web archiving capacity through collaborative development of improved systems interconnections between Archive-It and project partners. These interfaces will be prototyped with input from the full web archiving community and have broad community benefit. As the Web becomes the medium of record for all kinds of information, web archiving is an increasingly vital function of memory institutions. Web archiving, however, poses severe programmatic and technical challenges given its considerable infrastructure and technology requirements; the dynamic, evolving nature of the Web; and the sheer volume and complexity of born-digital materials being archived. This project addresses those challenges through researching and testing collaborative technology development models and community building.

Current Landscape: These high barriers to web archiving help explain why Archive-It is the overwhelming web archiving solution of choice, both for new programs and for existing programs transitioning from fully in-house operations. Archive-It has been used by over 350 institutions in 49 states to build 2700+ collections totaling over 9 billion documents. In the 2013 NDSA Web Archiving Survey, 70% of respondents were using the service. A recent partnership with CDL's WAS, transferring WAS users to Archive-It, will mean 86% of NDSA respondents are Archive-It users. This state of affairs represents both a preservation risk and a magnifying opportunity. The ease of entry into web archiving via Archive-It has not so far spurred investment in robust local preservation; and collaborative development of interfaces for interoperating with a core provider like Archive-It could better leverage distributed resources and provide standards for other platforms, such as LOCKSS.

Research Focus: Given this landscape, the assembled project partners would pursue designed-based research and development to prototype a more robust, API-based interconnection among nodes in a network spanning the lifecycle of web archiving operations and to research and test economic and community models to support future collaborative technical development. In the process of this work, and with community input, the project will answer a number of key questions, primarily: **What are the functional and technical requirements of API-based system interoperability for web archiving, and what economic and community models can inform and support such collaborative technology development?** Also, what areas of the lifecycle benefit from technical integration of a core infrastructure with local systems and needs? How can such integration scaffold newer or non-technical web archiving programs? This project aims to answer these questions by focusing on defining and piloting an export API that matches the community-driven success of Archive-It as a national-scale digital platform with key institutions with extensive web archiving experience and technical capacity. Project outcomes will better connect a shared core platform with demonstrated need and capacity to do local development, enhance researcher access, build community, and allow systematic propagation of web archives to national aggregators such as DPLA and DPN.

Projected Goals, Work, and Outcomes:
Goal: Research and develop an open API to facilitate transfer of web archive data between distributed systems and inform future work on API-based systems integration.
Work: Identify candidate API(s) by reviewing preservation risks, researcher use cases, applicable standards, institutional and community needs, and the access policy landscape. Specify functional and technical requirements, prototype, implement, and iteratively improve one or more working APIs at partner institutions. This builds on Archive-It's work with LOCKSS, a Mellon-funded project with NYU Libraries on API integration, NSF-funded work with Rutgers' on research use of web data and Stanford and UNT's work in digital libraries development, preservation systems, and web archiving.
Outcomes: Tested and documented API(s) for import/export of web archive data for enhanced local preservation, researcher access, and potential ingest into platforms like DPLA and DPN. Tutorial and workshop for API utilization. Published summary paper to guide future efforts.

Goal: Research and test economic and community models to support sustainable collaborative technology development for web archiving and greater community input on technology planning.
Work*: 1)* Evaluate successful models for collaborative development of open-source digital library software (especially Hydra and Fedora) to inform a more robust framework for shared technical contributions to web archiving tools and services. *2)* Via a national summit event, open working group calls, and affiliated groups, canvas the needs and available resources of U.S. web archiving organizations, with a focus on future technical development contributors and local beneficiaries.
Outcomes*: 1)* A publication with tested guidelines for collaborative models to support web archiving development *2)* A summit event and summary paper documenting community input for strategic planning around new tools and services; a durable collaboration framework that will scaffold smaller web archiving operations and facilitate follow-on investment in new APIs and tools.

Partner Roles: Internet Archive (Archive-It) will lead research and development on API, economic/community models and handle project management. Stanford will contribute to researching economic/community models. Stanford, Rutgers, UNT will contribute to researching API requirements, some development and code review, and testing local implementation of project API.
Evaluation: The Advisory Board will evaluate project work twice per year. We will additionally present ongoing work to and solicit feedback from the national and international community through project staff's role on the IIPC Steering Committee, NDSA, and other organizations. Post-summit surveys and working group calls will ensure community input and researchers will also be surveyed.
Relevance to funding priorities: The proposal supports the funding priority of a national digital platform by increasing access to the shared services and infrastructure of an existing platform while freeing local capacity to focus on local services and sustainability. The project addresses access at scale through research and development work for enhanced access, local preservation, and improved discoverability via integration with aggregators. The project will research and initiate new economic and community models. **Potential impact:** Archive-It's status as a national web archiving infrastructure ensures broad impact and project work will lay the groundwork for future collaborative development efforts and create efficiencies of scale via improved integration. The API will benefit researchers and users by providing improved access and potential integration of web archives into other national platforms.