The Internet Archive (IA), partnering with New York Art Resources Consortium (NYARC), which consists of the Frick Art Reference Library of The Frick Collection, and the libraries and archives of The Museum of Modern Art and the Brooklyn Museum is requesting $245,553 for a two-year project grant in the National Digital Infrastructure and Initiatives category of the National Leadership Grants for Libraries Program. This funding will build on the momentum of prior foundational work conducted in an IMLS Forum grant, related National Symposium held in 2019, and other community building and research work done as part of the "Advancing Art Libraries and Curated Web Archives" project.[1] The applicants seek funding to implement the next phase of this work, namely to create a cooperative, sustainable program that brings together art libraries and museums across the United States to scale capacity for coordinated curation and archiving of web-published primary source art and cultural heritage collections. The project will also expand discovery and use of these archives through the creation of a unified access portal of content from member collections.

Statement of National Need

Formerly print-only art records such as art gallery communications, exhibition catalogs, artist biographical material and other primary resources that humanities researchers rely on are now often solely published on the web. These web-based art materials are highly ephemeral, and their preservation and access poses technical, methodological, and resource challenges to individual heritage institutions. The migration of valuable primary source material from analog to born-digital formats, disseminated primarily via the web, has shifted methodologies around collection creation and preservation for cultural heritage organizations as well as research methodologies undertaken by art history scholars. This migration has also created opportunities for collaboration and shared curation services among cultural heritage organizations, along with unified discovery, access, and research support for the scholarly community.

This project addresses these needs and opportunities by creating a national network of art museums and libraries, pairing it with a large-scale, non-profit digital library and technology services provider, and launching a multi-institutional initiative to vastly expand the ability of art and cultural heritage organizations to collaboratively preserve and provide access to essential, ephemeral, born-digital art resources published on the web. This project will leverage applicants' foundational work and expertise to determine the appropriate structure and areas of focus for this cooperative action. The Internet Archive has nearly 25 years of web archiving, community development, and technology services experience and NYARC has over a decade of shared access, curatorial, and archiving collaborations. Prior IMLS-funded work, led by the same core team of Internet Archive and NYARC staff, resulted in several publications, surveys, workshops, a National Symposium, and follow-up stakeholders meetings to conduct financial, member, and sustainability modeling.[2] A core group of 10-15 art and museum libraries, including Getty Research Institute, Nelson-Atkins Museum of Art, Indianapolis Museum of Art at Newfields, The National Gallery of Art, Art Institute of Chicago, the Metropolitan Museum of Art, and SFMOMA have committed to this effort already, by participating in stakeholder meetings, building extended networks, and pursuing early-stage collaborative curation, Dozens more art

[1] https://archive-it.org/blog/learn-more/art-libraries/ & IMLS Awarded Grants, https://www.imls.gov/grants/awarded/lg-88-18-0069-18-0

[2] https://archive.org/details/aalcwanationalforumreport

and museum libraries have shown interest through participation in webinars, surveys, and the forum itself. This initiative has taken on an even greater importance with the current COVID-19 pandemic causing both new economic uncertainty in many institutions and increased ephemerality of community-based art and artists' web content.[3] The pandemic has made collaborative efforts and shared funding and service models more critical to libraries in pursuing their mission. This proposal will create shared digital infrastructure for art and museum libraries and enable the cost efficiencies of centralized administration, technology, and access; while benefiting from distributed, coordinated curation, discoverability, and researcher support.

The expertise of the cooperative members will be critical to identifying the appropriate collecting scope and coverage of the collections to and ensure the ongoing preservation of relevant content, and the success of the access and research outputs for the target communities that they serve. While key areas of focus have already been identified, including art galleries, artists' websites and local arts organizations, with a specific focus on artists of color, women, and marginalized or under-documented communities, new areas of focus may become necessary, for example websites of artists doing work around COVID-19 or gallery and arts organizations' online messaging of closures (whether temporary or permanent) due to local, state-level or national stay-at-home orders. Current statistics indicate that, for NYARC's New York City Galleries collection,[4] which has been collecting websites of art galleries in the city since 2014, 131 of the 398 galleries (just over 30%) represented in the collection have closed since being captured, meaning that the copy (or copies) of the gallery website in this collection may be the only record for future research use.

Web archive collections are often digital analogues to traditional physical holdings, but also provide the opportunity to incorporate new types of content and new collecting areas that are thematically relevant and inherently web-based. Selection criteria for inclusion in the collaboratively-created collections include:

- Community Need: Institutions participating in the network will identify community needs, based on their collecting goals and experience as well as outreach to both the subjects of the collections and anticipated end users. As noted above, participants have already done preliminary work to identify collecting areas of interest suited to collaborative, coordinated archiving.
- Coverage: Ensuring proper depth and breadth of coverage is important geographically as well as topically. The project's national scope is key and will include a range of museum types, with different geographical and topical coverage. The diversity of the network's initial member organizations is documented by the letters of support and commitment as well as the list of participants in events of the predecessor planning grant, and those who have expressed interest as part of continuing work.
- Uniqueness: While some primary art resources are published solely on the web, others may complement or extend other work. One key piece of this project will be linking its multi-institutional new web and born-digital collections to related existing collections or resources in member libraries and archives.

[3] "As Shutdown Crawls On, Artists And Nonprofits Fear For Their 'Fragile Industry,'" NPR, https://www.npr.org/2019/01/09/682925400/as-shutdown-crawls-on-artists-and-nonprofits-fear-for-their-fragile-industry
[4] https://archive-it.org/collections/4847

- Ephemerality: The project will prioritize the preservation and inclusion of web-published resources that are particularly at-risk, including galleries or arts organizations that are closing, pop-up or other temporary events-based content, and websites with frequent content change or deletion.

This project will formalize a collaboration among art libraries in the stewardship of historically valuable art-related materials published on the web. Some collaborating libraries are currently pursuing web archiving projects internally or for their own institution's web properties, while other art libraries have yet to implement a web archiving program, due to budget, staffing, or other limitations. Thus, born-digital, web-based arts content is not currently being collected in a comprehensive or cohesive way across the art library and museum community, meaning that at-risk content is being lost and current or future researchers who are interested in studying this content may be stymied in finding and accessing critical documentary sources.

Archive-It is the web archiving service of the Internet Archive and the most widely used service by cultural heritage organizations in the United States. Per an analysis of Archive-It partner institutions as well as the 2017 NDSA web archiving survey[5], museum and art libraries are currently a tiny percentage (less than 5%) of the institutions archiving web-published materials, even as much of the material comprising their current print collections, and critical to their collecting mandates, are now only published on the web. Preserving and providing access, at scale, to web-published materials is a technically complex process and curating the sheer breadth of art resources on the web is labor intensive. Identifying and promoting scholarly use of web archives is an enterprise that thus greatly benefits from collaboration, resource sharing, and technical and administrative centralization. This proposal builds on extensive preliminary work that will establish this cooperative model. The proposal also benefits from the success of the Internet Archive in pursuing a similar initiative, Community Webs, that is building a national network of public libraries documenting local history and marginalized communities via web archiving, a program that has received multiple awards from IMLS and multiple philanthropic foundations.[6]

Creating a multi-institutional cooperative for the aggregation of comprehensive, web-based arts collecting, will ensure national coverage in a way that no single institution could accomplish on its own. Centralizing infrastructure and technical and program management at the Internet Archive will allow this work to be housed alongside existing expertise and capacity for web archiving. These affordances will allow art and museum libraries to focus on the curatorial parts of the workflow that they are best suited to administer. The collaborative model will allow for ongoing coordinated collection, will provide education around outreach, metadata standards, and other topics of interest to the member organizations, and will allow for a diversity of archived resources, with local expertise and curation feeding a shared, national collection. This inclusive approach also aims to increase representation of under-documented and marginalized communities that are often missing from the historical record.

Project Design

[5] National Digital Stewardship Alliance (NDSA). "Web Archiving in the United States: a 2017 Survey." https://osf.io/3qh6n/
[6] See https://communitywebs.archive-it.org/ & http://blog.archive.org/?s=community+webs.

Goals and Objectives

The project design, work areas, and goals of this project are based on priorities identified by key museum and art library stakeholders during collaborative roadmapping and needs-gathering activities that occurred as part of the prior IMLS-funded work. Over the course of the two-year project proposed here, the project team will establish a cooperative initiative with a grant-funded Program Manager housed at the Internet Archive, enable at least two dozen participating libraries to coordinate curation of national, thematic web archive collections of born-digital art resources, and an Internet Archive software engineer will create a unified access and research portal, and will ensure all technical systems are interoperable with participant discovery and public access catalog systems. An in-person datathon will be held to facilitate research use of these collections and to provide iterative end-user feedback on the collections and access methods. This work will occur over four phases, each designed to contribute to the central goals of the project, which are:

1. Collaboration and Interoperability: The project will formalize a cooperative entity to facilitate collaborative and integrated access, and research support for born-digital, web-published art resources. This entity will include 15-20 members at launch, including NYARC, The Art Institute of Chicago, the National Gallery of Art, and the Nelson-Atkins Museum of Art, as well as other museums and art libraries, all working on comprehensive curation, preservation, and access. The cooperative will have established protocols for governance, participation, expansion, sustainability, costs, and resource sharing. The diversity reflected in the size, geographic location, and communities served by this group will ensure the comprehensiveness of the collections created and that the initiative provides value to institutions of any size or resource level.
2. Technical Capacity Building: The project will centralize the software and technical services, as well as operational and administrative staffing, at the Internet Archive, utilizing its Archive-It service to enable curation and access to a diverse set of thematic art-related web archive collections encompassing a national scope. Some of the launch partners are already using Archive-It for preserving their own web presences. The project will orient preliminary collaborative curation activities that emphasize the inclusion of content and commentary from underrepresented communities. These collections will, at least initially, focus on art gallery and artist websites in order to preserve, and make accessible for research, some of the most at-risk art primary source web content. Staffing will be located at the Internet Archive, but accountable to the cooperative, and will benefit from Internet Archives cost-sharing of portions of the human and technical resources in order to ensure the long-term sustainability of this effort.
3. Research Facilitation and Access: The project will create, support, and promote access and research opportunities to these vital arts-related web collections for a variety of end users through a unified search, discovery, and access portal. This effort will also include creation of derivative datasets for computational analysis, as well as virtual and in-person instructional events, user guides, and datathons to facilitate scholarly use.
4. Sustainability and Growth: The project will ensure the sustainability and growth of the initiative by developing a declining subsidy economic model for original participating institutions and a new member growth and a cost model that will allow continued collaboration, permanent digital preservation and access of collections, and long-term participation of member organizations beyond the original grant timeline and funding.

Due to the evolving nature of web content, the collections will continue to expand in size as the arts organizations and content areas included in the collections grow and change, so sustainability of the collaborative collecting models is central to its work and success.

Work Plan

The project will begin September 1, 2021 and work will occur over four phases, including formation, collection development and community engagement, building research and access methods, and sustainability, promotion and evaluation. These phases will inform one another and effectively build towards the goals for the project as a whole. A to-be-hired Program Manager (a job description is in the proposal's supporting documents) will work full time on this project and will be responsible for key functions and deliverables, including member and cooperative entity management, technical support for collection building and metadata, and oversight of the development of the access portal and research services. Project outreach and user engagement work will be shared between IA and NYARC. After the grant is awarded but prior to the official launch in September 2021, the project team will finalize commitments from the 15-20 initial member organizations and begin recruitment for the Program Manager role.

Phase 1: Formation *September 2021 - December 2021*

This phase will involve hiring staff and member recruitment and onboarding for establishing a cooperative entity. This entity will include 15-20 members at launch, all working on collaborative curation to ensure the collections' comprehensiveness and to avoid duplication of effort. Members will develop protocols for governance, participation, resource sharing, and planning for future growth and sustainability beyond the two-year grant-funded project.

Phase 1 Activities:

- The Project Director with consultation from NYARC staff, will hire the Internet Archive-hosted Program Manager role.
- Project staff will publicize the project, and opportunities for engagement, via a project website, social media, arts email lists, other outreach and marketing methods, and an open webinar(s), co-hosted by IA and NYARC staff.
- The project team will begin the onboarding of initial cooperative members, including a virtual kickoff meeting, finalizing initial collecting goals and benchmarks, and formalizing respective responsibilities of members and staff.

Phase 2: Collection Development & Community Engagement *January 2022 - July 2022*

This phase will utilize the centralized, non-profit technology platform of IA's Archive-It web archiving service, as well as the work of the Program Manager, to build a diverse set of thematic art-related web archive collections encompassing a national scope and emphasizing the inclusion of at-risk content from traditionally underrepresented communities. Staffing will assist with collection management and will benefit from IA's cost-sharing of staff and technology. This phase will also provide written and live trainings on collection development and establish governance, organizational, and operational practices, such as best practices around copyright, intellectual property, access, metadata standards, and other program development areas.

Phase 2 Activities:

- Based on the assessment in Phase 1, the art library project participants and Program

Manager will facilitate the creation and enhancement of arts collections by consolidating existing resources, adding metadata, and filling any gaps in coverage using Archive-It.

- The Program Manager will recruit additional practitioners for continued cooperative entity growth, in order to ensure sustainability, diversity, and adequate coverage of the collections, aiming for 25-30 members by the end of the grant period.
- The Program Manager will develop and deliver documentation and live training webinars to cohort members on topics including community outreach and engagement, collaborative selection and scoping of web content, metadata standards for web archiving, and access and research methods.
- With help from participating libraries including NYARC, the Program Manager will recruit stakeholders for end-user focus groups to discuss access/research methods, in preparation for Phase 3.

Phase 3: Building Research and Access Platforms *August 2022 - April 2023*
In this phase, the Program Manager, with assistance from NYARC, will create, support, and promote opportunities for curatorial, access, and research interactions with the project's vital library collections for a variety of end users and researchers. This will include community nomination and co-creation activities, instructional sessions on web archive research methods, (virtual) events, creation of datasets for computational analysis, a search, discovery, and access portal, and other services iteratively identified and produced in conjunction with end users.

Phase 3 Activities:

- The project's software engineer, a fractional role housed at the Internet Archive, will build the access and research portal, with input gained from the Program manager, working iteratively with end users and researchers.
- The software engineer will generate datasets for further research and analysis and documentation and example use cases for how these datasets can support research. The engineer will also produce metadata transformation scripts to ensure integration of web collection metadata into aggregated collections, such as DPLA, as well as member's own library discovery systems.
- The Program Manager will lead publicizing access/research use via listservs, blog posts, and social media and, alson with Internet Archive and NYARC staff, will hold at least 2 instructional webinars on access and data use, 2 virtual or in-person datathons, and develop additional interactive methods for iterative and collaborative use and refinement of the collections.

Phase 4: Sustainability, Promotion, and Evaluation *May 2023 - August 2023*
While the success of this project will be evaluated throughout the course of the grant period, through surveys and outreach to members, peers, and end users, the final phase will incorporate the cumulative results of these evaluations to ensure the long-term sustainability of the cooperative entity and the discoverability, expansion, and use of its collections.

Phase 4 Activities:

- The Program Manager will coordinate continuity and sustainability planning with the cooperative entity's members.
- The Program Manager will design and conduct a project evaluation for members to

inform medium and long-term priorities for this cooperative program.

- Program Manager will publish all educational materials, including workshop recordings and written guides, as well as research use/access case studies from datathons via the ARLIS/NA Commons.

Outcomes and Success Metrics

The success and key outcomes of this project will be measured based on the size, breadth, and diversity of the arts collections created, the cultivation of users and use cases for the access portal and datasets, and the sustainability and growth of thematic collecting amongst art and museum libraries.

Collection building success would be evidenced by the creation of at least 4-6 topical collections encompassing hundreds of websites with a comprehensive national scope. By the end of the grant period, the collections will likely include over 20TB of art resources, constituting tens of millions of individual art resources including websites, born-digital publications, auction catalogs, audio/video, social media, and other materials. The inclusion of diverse communities in the collection development process through outreach, focus groups and other engagement will help ensure the resulting collections are inclusive and provide thorough coverage of vital art history topics. The success of the cooperative model will be measured by having 30 or more art and museum library members by the end of the grant period, with sustainable funding models for ongoing membership and growth in collection building.

Upon creation of an access portal and datasets for these collections, the project team will hold at least 4 webinars, workshops, and/or datathons reaching at least 50 distinct end users including humanities researchers, reference librarians, data scientists, students and others to cultivate and support a variety of use cases. Ongoing use of the outputs of these sessions, including recordings, written documentation for the access portal and datasets, and at least 5 blog posts on curatorial and use-related aspects of this project.

Additional qualitative and quantitative analysis of outcomes will include focus groups and datathons for end users where use cases are discussed, provisioned and demonstrated, as well as surveys of the member organizations, relevant practitioner and end user communities, and usage statistics for collections, recorded trainings, and published outreach materials and attendance numbers for live webinars. All publications and outputs will be made publicly available under an open license for further sharing and reuse.

Diversity Plan

The diversity of organizations included in the cooperative entity, all curating collections in conversation with local and historically marginalized communities, provides heretofore untapped digital and cultural resources ripe for discovery and research. Institutional diversity will be achieved via recruitment of cohort member organizations diverse in geography, institution size, budget, urban vs rural, and area of curatorial specialization. With a model built on distributed, but coordinated curation and the cost-efficiencies of centralized technical services, infrastructure, and collection maintenance staff, the project allows for the inclusion of art libraries financially or technically unable to otherwise archive ephemeral web materials. While the areas of arts expertise of the member organizations will inform collecting goals, care will be taken to prioritize and solicit feedback from galleries, artists, and arts organizations identified for preservation that represent racial, socioeconomic and cultural diversity so that the resulting collections will represent historically excluded or underrepresented content and voices.

One key piece of ensuring diversity of the content in the resulting collections is encouraging the libraries participating in this project to solicit and integrate feedback from diverse communities of content creators, curators, and end users. Participants in the collaborative collecting cooperative will receive training, in the form of both written documentation and guides as well as live workshops, to explore best practices for this type of outreach, both from a community engagement and cultural competency standpoint as well as ethics and inclusion scholarship from the web archives field.[7] Gaining buy-in from the organizations and communities being archived will increase the potential impact of the project. By engaging in an active conversation with creators of different perspectives, experiences, and knowledge to inform ongoing selection, curators will have increased access to timely content, community collaborations, and/or emerging partnerships they would otherwise have missed.

National Impact

This project will have an immediate and lasting impact on a key domain of U.S. cultural history through the creation of a diverse national network of art and museum libraries dedicated to cooperatively building publicly-accessible thematic web collections. Creating shared digital, technical and infrastructure resources will remove barriers thus allowing participation of smaller and less-resourced organizations to participate in the stewardship of vital born-digital art history collections, an activity they may not otherwise be able to pursue on their own. A cooperative member framework for collection building, community engagement, and resource sharing will enable art and museum libraries to accomplish their mission and serve their users and community in new ways. The cooperative governance and membership models developed over the course of the project will be scalable, with the potential to include hundreds of art libraries and museums nationwide, for a truly national impact and access to archiving and digital library services will allow for collection and stewardship of an evolving and proliferating variety of digital cultural records. Collaborations of the type and design in this program -- shared technical services, distributed curation, localized collecting powering unified access -- along with a focus on an inclusive approach of co-creation with underrepresented communities to inform innovative scholarly uses of arts-related materials, are key to overcoming institutional or content silos and guaranteeing vital national digital infrastructure for libraries.

The cohort of organizations involved in this project and the web archive collections they create will be national in scope, ensuring breadth of access and cumulative impact, but the model is guided by local, community-centered expertise in selecting materials essential to future scholarship and learning. Centralizing the technical infrastructure removes cost and staffing burdens, allowing collaboration driven by knowledge sharing, not by institution size, type, or budget. Collaborative collections resulting from this project are identified by prospective cohort members and associated scholars as critical to the art historical record and particularly at risk for

[7] See IMLS funded Ethics and Archiving the Web National Forum, at which project staff spoke and participated, https://eaw.rhizome.org/.

loss. The impact of preserving this content, which might otherwise be lost or inaccessible to future scholars, will only continue to grow over time as web content changes or disappears.

The provision of public access and research datasets for the resulting collections, is particularly critical to the overall outcomes. The costs and complexity of web archiving mean that relatively few art and museum libraries are building these collections, resulting in a huge gap of resources documenting our contemporary times and cultures that will be available to learners and researchers. An enormous swath of heritage materials critical to studying art, art history, cultural creation, museum studies, and similar disciplines is already being lost. The impact on the viability of future scholarship is significant, as is the impact on the future legitimacy of libraries as repositories of memory and primary sources. A collective action of art and museum libraries working together is the only model that can ensure a comprehensive portion of this heritage is collected, preserved, and accessible for scholarly use.

Leveraging the Internet Archive's vast experience supporting computational research at Petabyte, "big data," scale, and its work with hundreds of mission-aligned libraries and archives on preservation and access, will give others the ability to support new forms of analysis and other novel research methods for digital collections. This proposal includes innovative work on collections aggregation for improved cross-institution discovery, making collection information machine-accessible for further interoperability, the creation of openly-available datasets for computational research use, and various related support and training services on methodological approaches. This lays the groundwork for any participating library to support any type of research use, regardless of local technical capacity -- an affordance that could make a true impact on how libraries support learners and cultural scholars, from the most basic to the most data savvy. The unified end-user discovery and access point will also institutional and community buy-in to the cooperative model and thus the sustainability and richness of the ongoing collecting activities.

The access portal, research datasets, and any instructional documentation or videos will be made openly available to any web users, thereby broadly expanding the impact on students, scholars, researchers, and the general public. These outputs, as well as written materials, including research outputs, open access scholarly literature, datasets, presentations, and blog posts will be publicized, promoted, and disseminated by a variety of means. These include practitioner outreach, as well as more targeted outreach to specific end-user groups such as art historians and researchers via groups such as the College Art Association[8], digital humanities and art history listservs, the Digital Art History list, and public calls for research participants at datathons. Additional discovery opportunities will be made possible by metadata inclusion in arts-focused linked open data or content aggregation systems such as LD4P[9] and Linked Art[10] which will allow metadata for this reference resource to appear alongside other relevant digital and analog materials.

The project team will ensure awareness and usability of its outputs via a strong dissemination plan. The ARLIS/NA Commons[11] and ARLIS/NA email lists will be a key part of outreach to

[8] https://www.collegeart.org/
[9] https://wiki.lyrasis.org/pages/viewpage.action?pageId=104568167
[10] https://linked.art/index.html
[11] https://www.arlisna.org/professional-resources/arlis-na-learning-portal

practitioners, as more than 1000 art and museum librarians in the United States (and North America) are members of this professional organization, and the Commons (formerly Learning Portal) is a public resource, available to ARLIS members and non-members alike. The team will publicize live informational webinars via the email lists, and share recordings after the fact, as well as providing long-term access to recorded workshops, white papers, policy documentation, instructional guides, and other publications via the Commons. Another key piece of outreach to the practitioner community will be presentations and working groups at conferences such as ARLIS/NA Annual Meeting , the International Internet Preservation Consortium Web Archiving Conference, Society of American Archivists Annual Meeting, MuseWeb all of which are either geared towards art and museum libraries and/or those collecting web content or have significant sub-groups that focus on those areas. The leadership of the ARLIS/NA Web Archiving Special Interest Group is involved in this proposal and has been part of prior work that has been instrumental in outreach and gathering support. Outreach to the wider public will also be part of the cooperative work. Following models for local community outreach developed via collaborative programs such as Community Webs as well as educational outreach to K-12 and college students to encourage interactions with web archives as primary sources. Both of these paths will be useful for broader contributions to and dissemination and use of these art historical collections.

Building a sustainable financial model that supports centralized technology and staffing, with varying price points of entry and ability for additive services, will ensure and amplify the long term impact of this project. Establishing, documenting, and promoting a collaborative model that gradually gains financial independence beyond grant funding through membership and added services revenue is a model that can inform similar collaborations in other types of organizations (historical societies, for instance), additional types of collection areas (performing arts, perhaps), and other areas of library technical services (such as citizen-oriented digitization).

The project will have national impact in myriad ways. It will establish a successful, implemented model of centralized technology, distributed but coordinated curation, and unified access and research tools, that allow any type of art library -- and in the future, any type of library -- to participate in digital stewardship activities. Empowering cooperative collecting on a national scale will preserve critical contemporary primary sources that may otherwise disappear, thereby giving researchers, academics and the general public the potential to better understand our times and reward patrons' trust in libraries as reliable storehouses of the national heritage and memory. Project elements supporting new research uses will push the boundaries of art historical methods, further advancing libraries as hubs of innovation. Outreach and engagement activities centralize a professional and user community-driven approach. Finally, the project's focus on prioritizing collaborative collections representing creators and constituencies otherwise missing from art and museum collection allows libraries to avoid what can be a wrenching appraisal and selection process influenced by costs and budget constraints, thus establishing a more inclusive and diverse national collection of born-digital archives better representing the plurality of our communities and country.

Schedule of Completion

	2021				2022											
Activities	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Program Manager hiring	■	■														
Publicize the project		■	■													
Onboarding initial members			■	■												
Create collections				■	■	■	■	■	■	■	■	■				
Ongoing recruitment of cohort members						■	■	■	■	■	■	■	■	■	■	■
Cohort trainings				■	■	■	■	■								
End user focus groups									■	■	■	■				
Build Access Portal													■	■		
Generate datasets, transform metadata															■	
Publicize research use															■	■

	2023							
Activities	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
End user trainings/datathons	▨	▨	▨	▨				
Sustainability planning					▨	▨		
Conduct project evaluation						▨	▨	
Publish outcomes and deliverables							▨	▨

DIGITAL PRODUCT FORM

INTRODUCTION

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to digital products that are created using federal funds. This includes (1) digitized and born-digital content, resources, or assets; (2) software; and (3) research data (see below for more specific examples). Excluded are preliminary analyses, drafts of papers, plans for future research, peer-review assessments, and communications with colleagues.

The digital products you create with IMLS funding require effective stewardship to protect and enhance their value, and they should be freely and readily available for use and reuse by libraries, archives, museums, and the public. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

INSTRUCTIONS

If you propose to create digital products in the course of your IMLS-funded project, you must first provide answers to the questions in **SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS.** Then consider which of the following types of digital products you will create in your project, and complete each section of the form that is applicable.

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS

Complete this section if your project will create digital content, resources, or assets. These include both digitized and born-digital products created by individuals, project teams, or through community gatherings during your project. Examples include, but are not limited to, still images, audio files, moving images, microfilm, object inventories, object catalogs, artworks, books, posters, curricula, field books, maps, notebooks, scientific labels, metadata schema, charts, tables, drawings, workflows, and teacher toolkits. Your project may involve making these materials available through public or access-controlled websites, kiosks, or live or recorded programs.

SECTION III: SOFTWARE

Complete this section if your project will create software, including any source code, algorithms, applications, and digital tools plus the accompanying documentation created by you during your project.

SECTION IV: RESEARCH DATA

Complete this section if your project will create research data, including recorded factual information and supporting documentation, commonly accepted as relevant to validating research findings and to supporting scholarly publications.

SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS

A.1 We expect applicants seeking federal funds for developing or creating digital products to release these files under open-source licenses to maximize access and promote reuse. What will be the intellectual property status of the digital products (i.e., digital content, resources, or assets; software; research data) you intend to create? What ownership rights will your organization assert over the files you intend to create, and what conditions will you impose on their access and use? Who will hold the copyright(s)? Explain and justify your licensing selections. Identify and explain the license under which you will release the files (e.g., a non-restrictive license such as BSD, GNU, MIT, Creative Commons licenses; RightsStatements.org statements). Explain and justify any prohibitive terms or conditions of use or access, and detail how you will notify potential users about relevant terms and conditions.

The project's primary products will be a suite of open educational resources and training materials, such as training videos, case studies, and associated educational materials will be released as CC0 -- fully public domain. The participating librarians will retain individual rights to the material they produce by participating in the project, such as blog posts, conference presentations, and articles, but will otherwise be released CC-BY.

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

The Internet Archive will assert no ownership rights to the new digital products created in the project and will impose no conditions on access and reuse.

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

The web archive collections will include content that is publicly available on the web at the time it is captured. The cooperative entity will work collaboratively with content creators when possible and will agree on any notification or permissions process as well as a takedown policy if there are objections to content that is captured.

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS

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

The project will create both open educational resources and web archive collections. For the OERs, materials will be organized around collection development, community engagement and web archive research topics and will include video/webinar, study guides, bibliography, case studies, and curriculum. Materials will include the videos, powerpoints, and digital publications. For the web archive collections, participating libraries will contribute to one or more topical web archive collection. This will amount to 10s of Terabytes of data and 10s of millions of web-published documents, like html/text, PDFs, audiovisual, and other. Web archive data will be preserved in the ISO-standard WARC format, and derivative datasets will be released in json-based formats.

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

Zoom or Camtasia will be used to produce the videos and webinars. All publications, videos, etc will be made available in the ARLIS/NA Commons and the Archive-It ZenDesk help center will provide an additional forum for access to OERs. The Archive-It web archiving service will be used to build the web archive collections and it includes a variety of software tools, including crawling, indexing, metadata, and access tools.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG, OBJ, DOC, PDF) you plan to use. If digitizing content, describe the quality standards (e.g., resolution, sampling rate, pixel dimensions) you will use for the files you will create.

Training videos will be in .mp4 and digital print materials will be dissemminated in .pdf. The web archive collections will be stored in the WARC format.

Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

The Archive-It web archiving service has a variety of custom quality control tools built into its services, including out of scope reporting, Wayback QA, proxy viewing tools, and others. The grant-funded Program Manager will review all OERs and publications before they are disseminated.

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

All materials created in the course of the project, including the training materials and the web archive collections, will be preserved and publicly accessible in perpetuity by the Internet Archive, with multiple copies stored in geographically distributed datacenters run by IA. Redundant storage copies are validated regularly for file fixity.

Metadata

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

The web archive collections include the ability to add metadata in Dublin Core. Once published both an OAI-PMH feed and a JSON API can be used for open publication of the metadata, and metadata mapping or transformations will occur as needed for integration with external access and aggregation services. Web content metadata is harvested as part of the archiving process and is indexed in CDX indicies and in the preservation WARC files.

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

As the metadata listed above is required for storage and replay of the web archived items, it is preserved as part of the WARC files. See above for WARC preservation plans.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

Existing OAI-PMH and RESTful JSON APIs allow programmatic access to descriptive, technical, and content metadata for the web archives. The work plan and budget includes engineering time for metadata mapping or transformations if needed for ingest into external access and aggregation services.

Access and Use

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

Both the training materials and the web collections will be openly available online. The web archives will be at both archive-it.org and archive.org and viewable via the Wayback software, and a specialized access portal will be built to facilitate access and discovery for this content. Additionally datasets will be generated and made available for download and local analysis. The training materials will be publicly available at archive.org in a special collection and will also be available through on the ARLIS/NA Commons, as expressed in the letter of support for this project.

D.2. Provide the name(s) and URL(s) (Universal Resource Locator), DOI (Digital Object Identifier), or other persistent identifier for any examples of previous digital content, resources, or assets your organization has created.

Archive-It: https://archive.org/ and user forum at https://support.archive-it.org/hc/en-us
Internet Archive: https://archive.org/

SECTION III: SOFTWARE

General Information

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

While software generation and distribution is not the core activity of this grant, we intend to create scripts for metadata transformation and an Internet Archive hosted access portal to facilitate browsing, full-text search and research use of the web archive collections. The portal will list all websites archived as part of the collaboratively-created collections, along with any descriptive metadata generated, and will utilize existing open source software for browsing (Wayback) and full text search indexing (ElasticSearch). The primary audiences served by this portal will include humanities researchers, reference librarians, data scientists, students and the general public.

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

While the archive-it.org public site allows for similar browsing and full text search capabilities, the access portal will also include research datasets generated specifically for this project and will allow for unique features and cooperative entity branding that will help enhance end-user experience and recognition of the project. This will simultaneously improve the discoverability and usability of these cohesive collections by end users and also increase buy-in, publicity and sustainability from member institutions.

Technical Information

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

The programming languages used will primarily be Python, and Unix (bash) scripting. The access portal will use the python-based Django framework, existing Archive-It RESTful data APIs, python-based Wayback software for content replay, and ElasticSearch for full-text search. These software choices include existing open source tools whenever possible and were chosen to integrate with Internet Archive's Archive-It service, which will be utilized for collection building, capture, and metadata addition, thus maximizing the reuse of system-generated data to populate the access portal while also utilizing frameworks that are familiar to IA staff, thus minimizing both technical and staffing costs of long-term hosting and maintenance of this software.

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

Metadata transformation scripts will allow member-generated metadata to interoperate with external aggregation and access systems. As mentioned in B.1, the access portal will use existing Archive-It RESTful data APIs, python-based Wayback software for content replay, and ElasticSearch for full-text search. These software choices include existing open source tools whenever possible and were chosen to integrate with Internet Archive's Archive-It service, which will be utilized for collection building, capture, and metadata addition, thus maximizing the reuse of system-generated data to populate the access portal while also utilizing frameworks that are familiar to IA staff, thus minimizing both technical and staffing costs of long-term hosting and maintenance of this software.

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

No dependencies

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

Development documentation will make use of the Jira issue and project tracking software. Documentation will include functional/technical requirements research, systems architecture/mappings, basic user guides (for developers and non-developers), and we be further articulated and documented in training materials.

B.5 Provide the name(s), URL(s), and/or code repository locations for examples of any previous software your organization has created.

IA has created dozens of software tools and systems including widely-adopted open-source tools for web archiving and digital libraries. Code is available at: https://github.com/internetarchive

Access and Use

C.1 Describe how you will make the software and source code available to the public and/or its intended users.

IA will release all project software and code with an open-source license, such as Apache 2.0, GNU GPL, or BSD licenses and all code will be published, updated, and available on Github.

C.2 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

currently unnamed subproject

URL:

https://github.com/internetarchive/

SECTION IV: RESEARCH DATA

As part of the federal government's commitment to increase access to federally funded research data, Section IV represents the Data Management Plan (DMP) for research proposals and should reflect data management, dissemination, and preservation best practices in the applicant's area of research appropriate to the data that the project will generate.

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

This project will generate derivative datasets based on the web archive (WARC) files captured as part of the collecting process. The purpose of these files is to facilitate more straightforward analysis of the contents of the archive, but will not collect or generate any additional data that was not included in the original archives. Dataset generation will occur in the second year of the program and will happen 1-2 times per year thereafter as the collections continue to grow and additional archived content is available.

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

N/A

A.3 Will you collect any sensitive information? This may include personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information. If so, detail the specific steps you will take to protect the information while you prepare it for public release (e.g., anonymizing individual identifiers, data aggregation). If the data will not be released publicly, explain why the data cannot be shared due to the protection of privacy, confidentiality, security, intellectual property, and other rights or requirements.

N/A

A.4 What technical (hardware and/or software) requirements or dependencies would be necessary for understanding retrieving, displaying, processing, or otherwise reusing the data?

Datasets will be downloadable and can be processed using command line tools and notebooks described here: https://github.com/vinaygoel/ars-workshop

A.5 What documentation (e.g., consent agreements, data documentation, codebooks, metadata, and analytical and procedural information) will you capture or create along with the data? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the data it describes to enable future reuse?

N/A

A.6 What is your plan for managing, disseminating, and preserving data after the completion of the award-funded project?

Datasets will be preserved on archive.org and made available via the project's access portal.

A.7 Identify where you will deposit the data:

Name of repository:

Datasets will be made available via the access portal, a dedicated collection on archive.org and the ARLIS/NA Commons

URL:

forthcoming

A.8 When and how frequently will you review this data management plan? How will the implementation be monitored?

N/A