Internet Archive, working with multiple library and museum partners, seeks a $244,000 IMLS National Leadership Grant for Libraries for *Expanding ARCH: Equitable Access to Text and Data Mining Services*. *Expanding ARCH* is an open source infrastructure implementation project that will improve the library, archive, and museum (LAM) community's ability to support computational use of their collections at scale by scholars, students, and the general public.[1] Over the course of two years, *Expanding ARCH* will advance IMLS Goal 3 by improving the ability of libraries and archives to provide broad access to and use of information and collections (Objectives 3.1 and 3.2), while also advancing Objective 5.2 by addressing a shared problem via collaboration on collections access for online users. The proposed work builds on the existing ARCH platform, which currently allows LAM professionals and disciplinary scholars to perform text and data mining on web archive collections. *Expanding ARCH* extends this platform to allow organizations to make their local digital collections of any type and format available for text and data mining in ARCH and creates new data analysis jobs, datasets, and features for additional types of computational research using LAM collections.

PROJECT JUSTIFICATION

Statement of Need

For more than a decade libraries, archives, and museums have worked to support computational use of collections.[2] In that time, demand for collections as data supporting computational use has only grown. No longer purely the province of STEM disciplines, universities are working to advance the ability to leverage computational methods across all disciplines. Representative examples in this vein can be seen in the Atlanta University Center Consoritum's *AUC Data Science Initiative* and Purdue University's *Integrated Data Science Initiative.* Beyond single university settings, multidisciplinary commitments to computational methods can be seen at the scholarly association level. The Association for Computers in the Humanities and the Text as Data Association leverage computational methods to better understand our past, present, and future.[3] In the professional sphere, journalists are critically assessing and potentially leveraging artificial intelligence tools like ChatGPT.[4] At the multinational level, the European Union is making a major multi-sector investment in the Digital Europe Programme, including the creation of Common Data Spaces.[5] A specific data space will be developed to support the computational use of European Union member state library, archive, and museum data. This is a pragmatic recognition of the central role that LAM collections have to play advancing computationally driven efforts. Simply put, curated and

[1] Funded partners include University of North Carolina, Indianapolis Museum of Art, University of Denver, and Williams College Museum of Art. Multiple additional volunteer partners are outlined in the narrative.
[2] We define computational use as algorithmic approaches to inquiry facilitated by text and data mining, machine learning, computer vision, data science, and artificial intelligence.
[3] Association for Computers in the Humanities, https://ach.org/about-ach/; Text as Data Association, https://textasdata.github.io/about/
[4] Israely, Jeff. "How Will Journalists Use ChatGPT? Clues from a Newsroom That's Been Using AI for Years." *Nieman Lab*. https://www.niemanlab.org/2022/12/chatgpt-and-the-future-of-trust/.
[5] Digital Europe Programme, https://digital-strategy.ec.europa.eu/en/activities/digital-programme.

representative data is required to derive insights at scale. The potential of LAM collections as data is abundant.

Current Approaches & Limitations

A diverse range of organizations work to develop and provide access to collections as data amenable to computational use. Collaboratives like the Hathitrust support computational use of a digital library through the Hathitrust Research Center workset builder and data capsules.[6] The national libraries of Scotland and the United States provide bulk access to a set of institutional collections geared for computational use via one-click downloads and API (application programming interface) via *The Data Foundry* and *LC for Robots*.[7] Federally and privately funded projects like *Always Already Computational: Collections as Data* and *Collections as Data: Part to Whole* have developed community, resources, and examples of how institutions large and small can provide access to and support the use of collections as data.[8] Proquest's TDM Studio offers commercial text and data mining services that work with their licensed digital collections.

These efforts have done a great deal to strengthen the capacity of LAM organizations to support computational work, but they are all bounded in ways that present significant challenges for faculty, students, and the general public. In the case of Hathitrust, the collection is largely a single type, monographs, however the scope of user-desired collection content types is typically much broader. For National Libraries, collections are often delimited by single institution holdings when user requests for collections as data tend to span multiple institutions. The *Collections as Data* projects provide guidance and use cases that await a robust non-profit community infrastructure solution. Commercial offerings, like TDM Studio and other efforts like it, are gated by costly licenses, presenting a significant cost-based challenge to user access. Taken collectively, existing efforts are bounded by their focus on guidance, a specific collection set, or for-profit business goals.

Despite the growth of LAM organizations providing access to collections as data there is no multi-institutional discovery mechanism for these collections. Users face an archipelagic environment without a way to find collections available for computational use beyond project sites and word of mouth. When users are able to find institutions that can support their requests, access mechanisms and policy are sufficiently variable to frustrate user efforts to develop datasets that can support their work on an institutional let alone multi-institutional basis. On the library research support services side of the equation, there is the challenge of attempting to respond to bespoke computational research requests that often outpace the diversity of available datasets, necessitating time intensive custom dataset creation work. Also, it continues to be the case that technical development and user access to collections as data remains a predominantly well-resourced institution endeavor. Finally, making digital collections available for computational use often requires significant digital infrastructure and technical expertise that the vast majority of LAM

[6] Worksets, https://analytics.hathitrust.org/staticworksets. Data Capsules, https://analytics.hathitrust.org/staticcapsules.
[7] "Data Foundry – Data Collections from the National Library of Scotland." https://data.nls.uk/. "LC for Robots". https://labs.loc.gov/lc-for-robots/.
[8] Collections as Data, https://collectionsasdata.github.io/.

organizations are unlikely to have the resources to build or staff. There is the community risk of computational research services being monopolized by for-profit, market-driven entities, similar to what happened with academic journals. In order to realize the full potential of the library, archive, and museum community to support computational work, future efforts must account for variation in infrastructure and staff resources at small and midsized organizations. Addressing the challenge of organizational inclusion helps ensure the formation of more representative collections and increases the possibility that a broader set of organizations can participate in the support of computational work locally and writ large.

Areas of Work

Expanding ARCH aims to meet the above needs through the collaborative development of an infrastructure that supports collection upload, discovery, and computational use of multi-institutional LAM collections. We propose to achieve this objective through the expansion of Archives Research Compute Hub (ARCH), an open-source platform that currently supports the computational use of web archive collections from Internet Archive's widely-used Archive-It service, where over 1,000 LAM organizations have created more than 20,000 web archive collections totaling over 5 petabytes.[9] The platform also supports text and data mining on custom collections from the 50 petabyte web archive of the Wayback Machine. The *Expanding ARCH* project entails two broad areas of work. First, the project will create user-informed workflows and conduct software development that enables a diverse set of partner libraries, archives, and museums to add collections to ARCH for computational use. Working with these partners will help ensure that ARCH can support the needs of small, medium, and large size organizations. Second, the project will work with librarians, scholars, and researchers to expand the number and types of data analysis jobs and resulting datasets and data visualizations that can be run using ARCH across multi-institutional digital collections.

Background

The Archiving & Data Services department of Internet Archive, which will direct this proposal's work, has long facilitated computational access to IA digital collections. Working directly with digital humanists, computer scientists, social scientists, and others, the team has provisioned large-scale collections access through custom dataset generation for researcher use. Archive-It Research Services, a more basic, pre-packaged web archive dataset service, was also developed in an effort to make datasets more accessible to users with a range of technical skills. Archive-It Research Services was an add-on feature that allowed institutions to request a number of pre-determined datasets that fulfilled frequent data mining use cases (text analysis, named entity recognition, network analysis, et cetera). Other efforts pursued similar work, notably the Archives Unleashed project, which created a broader range of datasets for web archive collections created using Archive-It. However, that effort required collections to be transferred to a computing cluster in Canada and used via a self-installed toolkit, rather than as a web service. Emerging from grant-funded research projects, the effort did not have a long-term business and sustainability plan.

[9] Archives Research Compute Hub (ARCH), https://webservices.archive.org/pages/arch.

In 2020, the Mellon Foundation awarded funds for Archive-It Research Services and Archives Unleased to combine their services and create a general web-based platform for text and data mining of web archiving collections that would be run on IA's self-owned, non-profit infrastructure and integrated with Archive-It's self-sustaining service model mixing free, subsidized, and paid digital archiving and computational research services.[10] Alongside the technical development work, the project funded two cohorts of research teams to use the platform and seeded a broader effort of community cultivation and scholarly practice in studying large-scale digital collections.[11] This work resulted in the ARCH platform.[12]

Current Status & Proposed Expansion

ARCH is currently in beta testing with a planned Spring 2023 launch. Currently, 35 organizations and 60 researchers and library professionals have used the platform in its beta version and ARCH staff have conducted 2 workshops for humanists and librarians, as part of its development.[13] ARCH consists of a web application that allows institutions to access their web archive collections (or provide access to researchers) and perform 14 different data analytical jobs, download the resulting datasets, and automatically create data visualizations for some datasets. Additional features include the ability to publish datasets openly on archive.org, to create Google Colab jupyter notebooks, to cite datasets, and to export data visualizations. *Expanding ARCH* will enable a greater number of small, medium, and large memory organizations to leverage ARCH by building tools that allow them to add any type of digital collections (not just web archive collections) to the platform, even if the data is added only temporarily in support of a single research project. With new types of digital collections and subcollections in ARCH, the project team will also create new types of data analysis jobs, datasets, and data visualizations to support a broader variety of research use cases. Overall, *Expanding ARCH* will prioritize the needs of small and mid-sized organizations that face challenges supporting computational use of collections due to gaps in staffing, expertise, and infrastructure.

Expanding ARCH will be achieved in concert with a set of strategically identified partner organizations, diversified according to organization size, organization type, collections, and target user communities served. The core, funded partner group includes a small research library (University of Denver), a large research library (University of North Carolina at Chapel Hill), a small college museum (Williams College Museum of Art), and a regional museum (Indianapolis Museum of Art). A number of other libraries and museums, such as Smithsonian Institution,

---

[10] Archive-It and Archives Unleashed Join Forces to Scale Research Use of Web Archives, http://blog.archive.org/2020/07/28/archive-it-and-archives-unleashed-join-forces-to-scale-research-use-of-web-archives/.

[11] Unlocking the research potential of web archives: An ARCH cohorts update, https://archive-it.org/blog/post/unlocking-research-potential-arch-update/.

[12] Archives Research Compute Hub (ARCH), https://webservices.archive.org/pages/arch.

[13] Internet Archive Welcomes Digital Humanists and Cultural Heritage Professionals to Humanities and the Web: Introduction to Web Archive Data Analysis https://archive-it.org/blog/post/internet-archive-welcomes-digital-humanists-and-cultural-heritage-professionals-to-humanities-and-the-web/.

Cleveland Museum of Art, and the Opioid Industry Documents Archive (University of California San Francisco and John Hopkins University) have agreed for their open collections to be used in the project. Additional libraries, archives, and museums, including some from the 35 institutions participating in the current ARCH beta pilot project, have verbally agreed to informally participate if the project is funded. *Expanding ARCH* will adopt an iterative design process in which partners provide input that informs ongoing technical development, workflows, and research use. Intentional diversity of partners will help ensure a robust set of ARCH enhancements and features that benefit a broad set of organizational users.

For institutional workflows, focused attention will be paid to ensure they map to a range of staffing, infrastructure, and collection realities. Workflows must be responsive to a small number of FTEs that can support computational work, limited infrastructure, and collections with varying levels of description. All partner organizations will successfully contribute at least one collection to ARCH. ARCH will also develop new dataset creation jobs that map to user-community needs to be designated by partner organizations and scholars. For example, Williams College Museum of Art would like to explore support for faculty and students invested in working with image collections as data. Accordingly, ARCH will explore datasets built from object recognition, color sampling, or compositional analysis. With feedback from partner users as well as researcher users, project staff will iteratively refine workflows and software development plans, UX wireframing, documentation and training, and outreach. All technologies and tools created in the project will be released with an open-source license and project staff and partners will socialize platform developments and final outcomes in blog posts, publications, workshops, and at relevant conferences.

PROJECT WORK PLAN

The project work plan is designed around two broad and intersecting areas of work. Work Area One includes supporting partner organizations adding digital collections to ARCH and researchers using these collections. This work area is referred to as "User Workflows," with "users" including both institutional users (for example, those adding digital collections to the ARCH platform) and research users (for example, scholars, students, librarians, and others) that will be analyzing the digital collections in ARCH. Work Area Two includes the project's technical work building the tools that allow external digital collections to be added to ARCH and creating new types of text and data mining capabilities, datasets, and data visualizations resulting from the digital collections added by partners. Work Area Two is referred to as "Platform Expansion."

The work plan will be overseen and managed by the project co-PIs and Product Manager and conducted with the primary project partners (those receiving subawards) and with the affiliated project partners participating in the project without funding or in a general user capacity. Project staff funded by the grant and focused on the work included in this proposal are a Product Manager, a Data Engineer, and the co-PIs. As ARCH is an existing platform, other product and engineering

staff will implicitly support this project, but are not budgeted or key personnel. The work plan consists of four six-month phases and costs are detailed in the proposal budget.

Phase 1: Research, Design, and User Studies (August 2023 to January 2024)

Internet Archive and partner organizations will conduct research, interviews, and user studies to inform development of institutional workflows for adding digital collections to the ARCH platform. Project staff will create a software development plan and implement a range of UX research processes. Necessary infrastructure will be provisioned and project staff will conduct promotion and outreach to publicize the project and recruit additional volunteer participants.

Work Area One: User Workflows

Partner Input & Workflow Design

- Project staff will conduct a literature review and environmental scan studying digital collections data transfers and data requirements of downstream data analysis tools.
- At least one blog post and one short open-access paper will be published about this work.
- Input from project partners will be solicited to inform technical development, responsible access policies (e.g., Santa Barbara Statement on Collections as Data, CARE Principles), and workflows for adding digital collections. Input will be solicited via an online survey, one virtual meeting of all project partners, and at least one virtual meeting with each project partner to better identify and understand unique institutional perspectives.[14]
- Workflow and policy designs will be shared with all partners for iterative improvement.

Community Engagement & Outreach

- Blog posts, announcements, listservs, newsletters, and email campaigns, leveraging IA communication channels and those of affiliated communities (SAA, DLF, DH, etc.), will promote the project and solicit additional participants.

Work Area Two: Platform Expansion

User Experience Design & Wireframing

- Project staff will create user personas, customer journey maps, functional requirements, and UX wireframes. These will be shared with project partners for feedback.
- Project staff will begin to create preliminary user documentation to be published online in the dedicated ARCH Help Center (built in Zendesk).

Infrastructure Provisioning & Software Development

- Dedicated hardware will be provisioned for expansion of the ARCH high-performance computing cluster, storage, and web application for the digital collection transfers.
- Software engineering plans and resources will be established, including Jira (for software project management), Gitlab repository, and other software management tools.

[14] Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner. "Santa Barbara Statement on Collections as Data --- Always Already Computational: Collections as Data," May 20, 2019. https://doi.org/10.5281/zenodo.3066209.Global Indigenous Data Alliance. "CARE Principles of Indigenous Data Governance.". https://www.gida-global.org/care.

- Software development will begin for adding external collections to ARCH.

Phase 2: ARCH Adding Collections Expansion: (February 2024 - July 2024)

Institutional Partners will begin to add sample local digital collections into ARCH. The project team will complete software development actualizing Phase 1 findings. Project staff, working with institutional and research partners, will begin preliminary work to identify and test the additional datasets informing the Phase 3 work expanding ARCH analytical tools.

Work Area One: User Workflows

Partner Input & Workflow Design

- All partner libraries, archives, and museums will add at least one (possibly more) digital collection to the ARCH platform.
- At least one online meeting will be held with each project partner, along with email inquiries and questionnaires, to understand institution experiences.
- Workflow designs and documentation will be iteratively improved per feedback.

Community Engagement & Outreach

- Project team will hold at least two group online meetings to solicit feedback from researchers and other users on data mining practices to inform Phase 3 work.
- Project staff will give at least one conference presentation.

Work Area Two: Platform Expansion

User Experience Design & Wireframing

- At least two webinars will be held in order to train institutions on how to use the newly built tools for adding digital collections into ARCH.
- Help documentation and UX designs will continue to be improved via partner input.

Software Development

- Software development work from Phase 1 will be completed for institutions to add digital collections to ARCH via web uploader, API, or web harvesting.
- An updated ARCH system architecture will be created that incorporates the project's work and published in the Github repository, https://github.com/internetarchive/arch.

Phase 3: ARCH Data Analysis Expansion (August 2024 - January 2025)

The new ARCH platform that enables adding external digital collections will be released and promoted. The engagement of computational researchers will be conducted to inform the addition of new datasets and data visualizations. Iterative software development will continue on prior features (adding collections) and commence on new features (new datasets).

Work Area One: User Workflows

Partner Input & Workflow Design

- There will be continued meetings with computational researchers about new datasets and scholarly use cases to inform the project's Phase 3 and 4 technical work.

- There will be a group meeting with funded partners to understand institutional perspectives on supporting data-driven research on their collections in ARCH.

Community Engagement & Outreach

- Two blog posts will be written summarizing the findings of meetings and research.
- Project staff will give at least one conference presentation.

Work Area Two: Platform Expansion

User Experience Design & Wireframing

- New web page templates, wireframes, and journey maps will be created for adding additional datasets, data visualizations, and related functionality to ARCH.

Software Development

- Iterative development will continue as needed on adding collections to ARCH.
- Software development work will commence on adding new data analysis jobs and resulting datasets and data visualizations to ARCH.

Phase 4: Promoting Adoption & Community (February 2025 - July 2025)

New data analysis jobs, datasets, and data visualizations will be added to ARCH. Project staff will work with partner libraries and individual researchers to support scholarly and custodial use cases. All code, documentation, and publications will be completed with a focus on promoting the project's accomplishments via conference presentations and community engagement.

Work Area One: User Workflows

Partner Input & Workflow Design

- Institutional partners and researchers will generate datasets and data visualizations, provide feedback guiding technical work, and inform documenting success stories.

Community Engagement & Outreach

- Two blog posts will be written summarizing the new ARCH features and highlighting specific use cases from the institutional partners and research users.
- Project staff will give at least two conference presentations.

Work Area Two: Platform Expansion

Software Development

- Software development work adding new data analysis jobs, datasets, and data visualizations will be completed early in Phase 4.
- All project technical documentation will be completed and all code will be released under an open-source license in the project's Github repository.
- The second half of Phase 4 will focus on operational maintenance, monitoring, and deployment, and tools for sustainable technical operations.

Digital Product and Measurement Plan

The Digital Product Plan includes releasing all the project's code, algorithms, and tools under open-source license in the project's Github repository, https://github.com/internetarchive/arch/ and all documentation, workflows, and publications under open license CC-BY-NC-SA. Collections added to ARCH, and any resulting datasets, will inherit the access policies of the custodial institutions. All public code and publications will be archived and accessible via Internet Archive in perpetuity. The Digital Measurement Plan will build on existing ARCH logistics. These include a weekly 15-minute Monday standup engineering meeting, a weekly 30-minute engineering and product meeting, and a weekly one-hour feature/roadmap meeting that alternates between engineering and product. Project staff and Project Directors will have a standing monthly meeting on grant deliverables and Project Directors will meet quarterly with the IA Finance department to track grant finances. Meetings between project staff and partners, and communication and dissemination plans are detailed in the Project Work Plan. The Digital Product and Performance Measurement Plans are detailed further in additional documents.

DIVERSITY PLAN

The project staff has intentionally included a variety of institution types as funded partners, including a mid-sized university, a mid-sized museum, a small museum, and a large university. This diversity will provide a range of institutional perspectives on collections, staff, infrastructure, and workflows. Funded partners will be encouraged to contribute collections to ARCH that document underrepresented communities. Partners will also be encouraged to advise on policies that support responsible use of collections as data in a DEI framework. Additional volunteer partners will represent further diversity, including soliciting participants from the 150 institutional partners in IA's Community Programs that represent small organizations documenting local history, including marginalized groups, tribal and indigenous archives, and community archives. In addition, the project will recruit participation from diversity-oriented scholarly communities, such as the U.S. Latino Digital Humanities Program. As documented in the Project Work Plan, project partners will be instrumental in guiding workflow and technical development through a variety of facilitation and engagement activities. The *Expanding ARCH* project will strengthen the field's commitment to diversity, equity, and inclusion by building a platform that can allow library and archive collections of underrepresented communities to be made available for contemporary forms of computational research and by involving diverse scholarly communities in making use of the platform.

PROJECT RESULTS

At the conclusion of this project, *Expanding ARCH* will have implemented an open source, collaboratively developed infrastructure that improves library, archive, and museum community ability to support computational use of their collections at scale. All partner organizations will have added at least one (likely more) digital collection to the ARCH platform. ARCH will provide the following features: low barrier institutional ability to add digital collections; centralized infrastructure that supports multi-institutional collections as data discovery and access; user-driven work with multi-institutional collections as data; and new content specific data processing jobs (for

example object recognition in images, speech to text). This feature-set maps directly to identified LAM institution challenges supporting computational use of collections. Enabling low barrier institutional ability to add collections to ARCH addresses an organizational inclusion challenge by ensuring that small and mid-sized organizations can more easily make their collections available for computational use. A focus on supporting multi-institutional, multi-format collection contributions to ARCH also helps foster the development of an aggregate collection that is not bound by a single institution's holdings or a primary collection content type. An aggregate collection is also a more representative collection (e.g., unification of previously distributed collections - artist works, photographic works, manuscript materials) affording the potential for greater research impact. Centralized collections as data discovery and access, saves time of the user and the LAM professional. The creation of new content specific ARCH data processing jobs provides more ready access to collections as data derivatives tuned for computational work, streamlining the path from collection access to collection use.

Iterative development work with a diverse set of partner organizations will ensure that *Expanding ARCH* is able to support small, medium, and large cultural organizations. The project team will actively solicit feedback on work in progress from the broader community through blog posts and conference presentations. Feedback will further inform and strengthen the ability of the project to serve a wide range of LAM organizations. As an open source project, all code, algorithms, and tools will be openly released on Github. All documentation, workflows, and publications will be released under open license. ARCH will be sustained through a combination of in-kind services, philanthropy, and non-profit based service cost recovery.

Methods of research are changing. And libraries, as custodians of the rich, diverse collections documenting our nation's heritage, memory, and accomplishments, must change how their collections are accessible and usable to meet these new forms of research. The Collections as Data movement has succeeded in shifting mindsets, promoting advocacy, and seeding nascent local experimentation in reorienting access. However a vast majority of LAM institutions will continue to face local funding, staffing, and infrastructure challenges as they work to facilitate computational use of their collections, given the custom computing and unique engineering expertise required. Decentralization of computational research services will be possible among large research universities, but the wealth of our vast national collections are held by small and mid-sized libraries, archives, and museums. Only mission-aligned, non-profit, community-driven, and infrastructure-rich partners will be able to provide a non-exploitive, non-commercial option for making their collections available for computational use. This is the need, the community, and the purpose that the ARCH platform, and this proposal, aims to serve.

Schedule of Completion

	Year 1 (August 2023 - July 2024)											
ACTIVITIES YEAR 1	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**
Phase 1: Research, Design, and User Studies (Aug 2023 to Jan 2024)												
User Workflows												
Research and literature review for digital collection transfers	■	■										
Outreach, promotion, and additional partner recruitment	■	■	■	■								
Partner meetings on digital collections transfers workflows and development		■	■	■	■							
Platform Expansion												
Provision digital infrastructure	■											
Software development planning	■	■										
User personas and UX design		■	■	■								
Software development for adding digital collections via online transfer			■	■	■	■						
Phase 2: ARCH Adding Collections Expansion: (Feb 2024 - Jul 2024)												
User Workflows												
Partners begin adding digital collections to ARCH							■	■	■	■	■	■
Continued partner meetings, surveys, feedback, and workflow design							■	■	■	■		
Recruitment and solicitation of input from researchers and scholars										■	■	■
Platform Expansion												
Webinars, workshops, and user documentation on adding collections							■	■	■			
Software development continues and is completed							■	■	■	■	■	■
Code published open-source in github and features pushed in production											■	■

	Year 2 (August 2024 - July 2025)											
ACTIVITIES YEAR 2	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**
Phase 3: ARCH Data Analysis Expansion (Aug 2024 - Jan 2025)												
User Workflows												
Meetings, surveys, and feedback from researchers on datasets and use cases	■	■	■									
Meetings, surveys, and feedback from institutions on research services	■	■	■									
Publication of findings and outreach and conference presentations				■	■	■						
Platform Expansion												
Wireframes, mockups, and journey maps for adding datasets and visualizations		■	■									
Post-release iterative development enhancements for adding collections	■	■										
Software development for adding new datasets and data visualizations		■	■	■	■	■						
Phase 4: Promoting Adoption & Community (Feb 2025 - Jul 2025)												
User Workflows												
Institutions and research test and use new datasets and visualizations							■	■	■	■		
Blogs, publications, and conference presentations on new dataset features								■	■	■	■	■
Platform Expansion												
Software development for adding new datasets and data visualizations							■	■	■	■		
All code and documentation published open-source in Github											■	■
Operational monitoring, maintenance, sustainability development work											■	■

Digital Products Plan

Type

The *Expanding Arch* proposal includes work to expand an existing digital platform. Beyond software development, which we describe below, the proposed work includes two areas. First, expanding the platform to allow partner institutions to add their external digital collections to ARCH in order to facilitate researcher data mining of these collections; and, second, work to create new datasets and corresponding data visualizations that are created from these collections. ARCH currently supports creating 14 different derivative datasets from web archive collections. The work proposed here aims to add additional derivative datasets specific to text, image, or complex (such as software or video) digital collections. The proposed work will determine the number and type of datasets that are valuable to research use, but we expect at least 6-8 new datasets. When possible these datasets will also have corresponding data visualization automatically generated in the browser, such as bar charts, pie charts, network graphs, et cetera. The datasets will generally be in common structured formats, such as .csv, .tsv, .json to facilitate their use in additional data analysis tools. The data visualizations will be created using the d3.js library or, when necessary, additional libraries for rendering structured data in charts and graphs.

The software development will cover three areas: the data processing necessary for creating the datasets from the digital collections, the applications, tools, and systems allowing partners to add digital collections via web uploader or API, and the extension of the ARCH web application that users interact with for adding collections and creating, managing, and sharing datasets. The data processing work is written in the Scala programming language, as it works with the parallelization and map-reduce job oriented large-scale data processing that is performed in our Hadoop-based high-performance computing cluster. The tools and applications supporting the digital collections transfers will be written in Python, the de facto programming language in the department performing this work, and the web application is based on the Django web framework, with front-end development done in python, javascript, typescript, and web components. Databases are generally Postgres, search is generally Elasticsearch, deployment is done in Ansible playbooks, Temporal handles workflow execution, Prometheus handles monitoring, and applications are made available in Docker containers for local development. All these languages, systems, frameworks, and tools are used in other digital products and services run by the department and have been tested in very high-use and data-intensive workflows covering hundreds of millions of transactions and tens of terabytes of data processed per day.

Availability

All code, applications, tools, scripts, etc created as part of this project will be publicly available and released under open-source license (MIT) on our Github page. As an extension of an existing platform, this repository already exists, https://github.com/internetarchive/arch. For the datasets generated from partner digital collections, the institution will determine whether these are to be

made publicly available (almost all certainly will choose so) and, if so, access options are for them to be published via Internet Archive on archive.org under open license (CC-BY) (the digital repository system underpinning archive.org is a linux/ubuntu and python based custom repository system), or as Google Colab notebooks, or the institution can download the datasets and make them available via their own website or repository under their own terms of use.

Access

All code, applications, tools, scripts, etc created as part of this project will be publicly available and released under open-source license (MIT) on our Github page. The digital collections added to ARCH by partners will fall under the access and rights policies of the contributing institution. Partners have been strongly encouraged to use openly-licensed digital collections as part of their participation in the project. Similarly, the access policies for the datasets and data visualizations created from partner digital collections will be determined by the custodial institution; however multiple workflows and tools are available for releasing these derivatives as open access data. All publications, training materials, and non-technical documentation will be published under open license (CC-BY or similar). No privacy concerns or cultural sensitivities are expected, as partner institutions are selecting what digital collections to use in the project. That said, the Project Work Plan describes that meetings and workflows will take into account ethical practices towards data sharing as part of the workflow and procedural development work of the project.

Sustainability

The department at Internet Archive that is proposing this grant runs a number of sustainable digital products and this business, product, engineering, and operational team will be able to apply similar models and practices to the ARCH platform. Like other digital products in the group, ARCH will be sustained beyond the period of performance through a mixture of in-kind services, philanthropic support, and non-profit based, fee-for-service cost recovery. Via a combination of subsidy, external funding, and earned income, ARCH will be maintained by Internet Archive into the foreseeable future beyond the grant period. Similarly, Internet Archive guarantees the perpetual preservation and access of all archived data collected or created using any of its institutional services. ARCH utilizes a number of different repository systems maintained by Internet Archive (both block and object storage) including high replication (generally 4 or more copies) held in multiple geographic locations in multiple countries. An extensive data and data center security and preservation plan is provided to any service users that includes migration planning and Internet Archive is committed to the preservation and accessibility of all archived data in perpetuity.

Organizational Profile

The Internet Archive is a 501(c)(3) non-profit digital library founded in 1996 with the mission to provide "universal access to all knowledge." The organization's purpose is to build an internet library with the aim of offering permanent access for historians, scholars, journalists, students, the blind and reading disabled, as well as the general public, to historical collections that exist in digital format including texts, audio, movies, images, software and web pages. The mission is stated in the organization's 2021 Form 990 Department of the Treasury filing and is approved by the Board of Directors. The digital library is available at https://archive.org. Located in San Francisco, California, the Internet Archive currently maintains a freely-accessible, online digital library and archive of more than 99 petabytes of data: 800 billion websites, 6 million films and videos, 14 million audio recordings, 41 million texts (including 7 million digital books), and over 900,000 pieces of software. Each day, over 2 million visitors use or contribute to the Internet Archive, making it one of the world's 200 most popular sites with over 200 page views per second and 33 billion total downloads. Internet Archive's web archive, the largest and oldest publicly-available web archive in existence, was launched in 1996 and currently archives over 750 million web pages every day, all available via the widely-used Wayback Machine (https://archive.org/web/), and totals over 50 petabytes of data. The Internet Archive actively works in partnership with national libraries and archives, universities, governments, research institutions, and other organizations across the world on digital preservation, open access, computational research services, digital library standards, and open source technology development. Internet Archive also has an established record of related research projects funded by the NSF, IMLS, NEH, Sloan Foundation, Knight Foundation, Mellon Foundation, and others. The Internet Archive was designated a public library by the State of California in 2006. The Archiving & Data Services department is the organizational unit responsible for carrying out the work in this proposal. The department runs a suite of free, subsidized, and paid digital services and products for web archiving, digital preservation, computational research support, and web and data services. Over 1,200 organizations in over 30 countries use these services including research, national, public, and special libraries of all sizes, federal agencies, governments at all levels, social-impact nonprofits, NGOs, and other mission-aligned heritage organizations. Alongside its free and paid digital products, the department has also been awarded over 15 grants from a variety of federal and foundation funding bodies.