Internet Archive (IA) and a diverse group of partners requests $230,000 for a two-year National Leadership Grant for Libraries (NLG) Implementation project that will give memory organizations of any size the ability to support the use of their collections as data by scholars, students, and the public and will build digital infrastructure promoting computational access to museum and library collections.[1] The project will advance NLG Goal 3 to "improve the ability of libraries and archives to provide broad access to and use of information and collections" (Objectives 3.1 and 3.2 specifically), while also advancing Objective 5.2 by addressing a shared problem via collaboration on collections access for online users.

Project Justification: We propose to expand upon the Archives Research Compute Hub (ARCH), an open-source platform that centralizes infrastructure and data processing services to make library web archive collections available as data. *Expanding ARCH* will scale ARCH infrastructure and services to also work with digitized text, image, and audio-video collections. Additionally, to support data-driven research and educational uses that span multiple collections, *Expanding ARCH* will develop tools that allow libraries to upload locally-stored digital collections into ARCH, resulting in an aggregation of diverse collections and formats from many memory organizations. The ARCH platform was originally created with support from the Mellon Foundation to merge the technologies of the Archives Unleashed project and IA Archives Research Services to facilitate data mining, dataset creation, data visualization, and dataset publication for the tens of thousands of web collections build by hundreds of memory organizations using IA's web archiving services. *Expanding ARCH* will scale this platform to allow similar computational research on non-web digital collections and will enable small, medium, and large memory organizations to make their digital collections readily accessible for a broad range of computational analysis, from digital humanities, to data science and machine learning. *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. As the number and size of digital collections at memory institutions grows, researchers, students, and the broader public are increasingly interested in using collections in compute-intensive forms of scholarship. However, these methods often require custom technologies, digital infrastructure, and specialized knowledge that few libraries have or can afford. *Expanding ARCH* will bridge this gap via equitable, inclusive access to a platform supporting innovatives uses of their collections.

A number of efforts have sought to enable computational use of digital collections.[2] [3] *Collections as Data: Part to Whole* developed exploratory frameworks for local collections as data projects. Hathitrust Research Center and JSOR's Constellate support computational use of specific collection sets like monographs or journals. Commercial services like Proquest's TDM Studio, offer data mining as an add-on to licensed products. These efforts are bounded by their focus on guidance, a specific collection set, or for-profit business goals. Also, none of these services allow institutions to add their own collections to these platforms. Yet few research questions are delimited by a single institution, collection, or license, typically using many types of collections from multiple institutions. *Expanding ARCH* meets this need for a platform that facilitates computational research across a multi-institutional set of collections and formats.

The last few decades have shown the importance of community archiving by smaller institutions and how this type of collecting is essential in diversifying the historical record. Yet these same smaller institutions are in danger of being left behind in the broader movement of offering collections as data. What was once a lack of diversity in collecting now threatens to be a lack of diversity in collection access services. By working with a mix of small, medium, and larger partners, *Expanding ARCH* will prioritize its work to lower the technical barrier for any size memory organization to make their digital material available for innovative forms of research, thereby greatly broadening access to our national digital collection and centering libraries in an emerging field of knowledge production and scholarship.

[1] Partners include commitments or expressions of interest from Indianapolis Museum of Art, UNC-Chapel Hill, Cleveland Museum of Art, University of Denver, with four more libraries and museums still considering, but likely to join if a full proposal is invited.
[2] BigLAM (Libraries, Archives and Museums). 2022. https://github.com/bigscience-workshop/lam.
[3] Computing Cultural Heritage in the Cloud. Library of Congress. https://labs.loc.gov/work/experiments/cchc/.

Internet Archive has extensive experience working with scholars, memory organizations, and industry to support large scale text and data mining projects. Staff at IA have also led the global "Collections as Data" movement and IA's Community Programs initiative, funded by IMLS, NEH, and Mellon Foundation, includes over 250 smaller memory organizations documenting underrepresented communities via digital archiving -- collections that can be readily made available for computational use. *Expanding ARCH* will work with a diverse set of memory organization partners that will inform ARCH development to most effectively support research services. ARCH will also engage a cohort of ten scholars in additional testing and feedback. These cross-cutting groups will ensure the needs of both libraries and research user communities guide platform development.

Project Work Plan: The project design is structured around three areas of work: Institutional Workflows, Platform Development, and Researcher Engagement. These three areas of work will be conducted in four overlapping, but sequenced, phases. Phase 1: Research & Design (6 months): IA and Institutional Cohort will conduct research, interviews, and user profiling to understand the institutional workflows, tools, and designs for ingesting local digital collections into the ARCH system. IA will create a software development plan informed by this research. Phase 2: Iterative Technology Development & Testing (6 months): 1. IA will commence software development actualizing Phase 1 findings. Institutional Partners will ingest local digital collections into ARCH. IA will commence platform development. Scholars Cohort and IA will begin preliminary work to identify additional datasets, features, and products that fulfill the needs of computational researchers using ARCH and Institutional Cohort collections. Iterative software development continues. Phase 3: Expanded ARCH platform release (6 months): IA will release the prototype expanded ARCH platform, conduct Institutional and Scholar cohort testing, and do training, outreach, and further partnership development. Iterative software development will continue post beta release. Phase 4: Adoption & Community Cultivation (6 months) - Project team will solicit further feedback from Scholars Cohort and support specific research projects using the newly expanded ARCH platform. Further use and adoption by scholars will be sought. Documentation, workshops, and community cultivation will take place. Project team will publish open educational resources, papers, and guides supporting librarian skills development in supporting computational research by using ARCH.

Diversity Plan: *Expanding ARCH* will center diversity in its work across three vectors. First, a diverse mix of institutional partners will be involved, prioritizing the needs of under-resourced libraries and museums in platform development and workshop and training materials creation. Second, building on IA's Community Archiving programs, we will foreground digital collections that document underrepresented groups for inclusion in the ARCH platform and promote these collections to researchers for computational use. Third, we will assemble a diverse mix of scholars to provide input on feature development and ensure the platform centers a values-first and ethically motivated approach to supporting digital scholarship.

Project Results: *Expanding ARCH* will advance the capacity of memory organizations to allow computational use of their digital collections by scholars, students, and the public. Over 1,500 organizations have over 10,000 publicly-available digitized and born-digital archives available through Internet Archive. This project will enable text and data mining datasets on any of these collections, making hundreds of millions of historical records available in new ways. At least six memory organizations will partner on the project, making select digital collections not already hosted by IA available via ARCH, thus proving the extensibility of the platform. Also, project partners will recruit ten computational researchers to help inform software development, workflows, and for using ARCH in applied research. The project team will also conduct at least two workshops for 50+ librarians to teach them the skills and competencies for supporting data mining of their collections. *Expanding ARCH* will advance digital inclusion by scaling open infrastructure that gives memory organizations the ability to support computational use of collections that advance knowledge and the public good.

Budget Summary: Funds requested: $230,000. Internet Archive (Lead): $170,000 ($145,000 for salaries, $20,000 for materials and equipment, $5,000 for travel) for engineering, project management, infrastructure, workshops, outreach. Institutional Partner stipends: $50,000 ($10,000 each for 4-6 partners) for contributing digital collections, meetings, workflow development. Scholars Cohort stipends: $10,000 ($1,000 each to10 scholars) for researcher engagement, platform testing and feedback, meetings, and advisory services.