Empowering Data-Driven Research Through Digital Archives with Internet Archive’s ARCH

Helge Holzmann
Internet Archive
300 Funston Ave
San Francisco, CA 94118, USA
helge@archive.org

Workshop @ RESAW 2025 — Siegen, Germany

Abstract

In this comprehensive 2-hour session, we will explore and discuss the latest advancements and innovations of the Internet Archive's ARCH platform.

ARCH (Archives Research Compute Hub) is a cutting-edge platform engineered to facilitate the building of research collections, enable computational analysis, and support the generation of datasets from terabytes and even petabytes of data. ARCH supports the open publication and preservation of user-generated datasets created from thousands of libraries, archives, and memory organizations worldwide, empowering researchers, students, and information professionals to study, analyze, and interpret digital collections in unprecedented ways.

Designed with a focus on curating research collections using primary digital sources such as web pages, texts, and images, ARCH enables users to effortlessly create over a dozen distinct datasets from these sources with a simple click. These datasets can be directly downloaded either through an in-browser interface or via an API, enhancing accessibility and user experience.

Moreover, ARCH facilitates the efficient utilization of these research-ready datasets by offering in-browser data previews and visualizations. More interactive analysis is encouraged and supported by enabling the integration of computational tools such as Jupyter Notebooks, Google CoLab, Gephi, and Voyant into the research process.

A significant feature of ARCH is its one-click publication mechanism on archive.org, allowing datasets to be easily accessed, shared, and preserved indefinitely. This feature not only promotes open access to information but also ensures the long-term preservation of valuable data.

To support and enhance user experience, ARCH provides comprehensive technical support, online training, and extensive help center documentation. These resources are designed to optimize the effective use of the platform, making sophisticated research processes more

accessible to users who may not have advanced coding or scripting skills.

ARCH benefits from the robust, non-profit infrastructure of the Internet Archive and utilizes open-source tools to streamline the computational handling of digital collections. This enables librarians, collection managers, and educators to offer sophisticated research tools to their communities, thereby democratizing access to advanced research methodologies.

Recently, ARCH has integrated AI-powered tools that enhance the platform's capabilities. These tools are readily accessible on our dedicated computing cluster, equipped with GPU support, making advanced computational tasks more feasible for our users.

ARCH is available for both institutional and individual use, offering flexible access options for a diverse range of professionals including researchers, librarians, archivists, museum staff, journalists, and more.

This format provides a comprehensive overview of ARCH’s features, but we will also delve deeper into the technical details and underlying technologies. It will feature a combination of presentations, brief demonstrations, and interactive live sessions. Participants will have the opportunity to engage with the tools interactively, ask questions, and view actual datasets, making this an informative experience that offers participants a clear view into how the ARCH platform can enhance their research capabilities.

Author Keywords: *Digital Archives and Preservation; Computational Analysis of Collections; Open Access Publication; Interactive Research Tools; AI-powered Data Analysis*