Building Archival Digital Discovery and Access Systems with Arclight

Introduction

The University at Albany, SUNY (UAlbany) and the Empire State Library Network (ESLN) request $249,996 for a two-year implementation grant to further develop Arclight into a single platform for digital archives and special collections materials. This project will enable Arclight to provide discovery and access to digital materials, metadata, and full-text content using International Image Interoperability Framework (IIIF) manifests alongside archival description in finding aids. This project will lower the barriers for implementing systems for archival discovery and delivery and enable under-resourced institutions to provide access to digital objects on-demand or by using description they have already created. This work is well-aligned with the NLG-L Program Goal 3 and also builds on the knowledge and accomplishments of three IMLS-funded projects, the Cross-Search and Context research project, the Lighting the Way National Forum, and the National Finding Aid Network project which found that users struggled to access digital content described by finding aids.

Project Justification

Archival Approaches to Digital Materials

Archival methods and best practices are designed specifically to allow repositories to efficiently manage tremendous volumes of materials. Repositories can easily contain tens of thousands of cubic feet of physical materials where everything should be discoverable and retrievable. Archivists understand that describing every object in an archival repository is both impossible and an inefficient use of our limited descriptive resources. Instead, archivists describe materials hierarchically at progressive levels of detail based on value and use. This means that materials that are only rarely consulted get minimal description and/or are described in aggregate for basic intellectual control, while commonly used items get detailed description and potentially digitization for easy access and visibility.[1] In practice, managing digital materials in a traditional DAMS or digital repository hinders archivists from using these very effective methods.

Archivists are now confronted with born-digital acquisitions which often include extremely large volumes of items and are ripe for automated description approaches that both provide potential to reduce the time requirements to describe materials, but also pose important risks as this work becomes more distanced from humans. Archivists also face increasing demands to digitize more of their collections for remote use, where time-intensive detailed metadata records are often the biggest barrier to making more materials available online.[2]

Existing archival description methods and best practices are a perfect fit for addressing these challenges. Archivists must describe digital materials hierarchically by prioritizing aggregate groups of materials first, and further describing materials more granularly based on user needs and available resources. This is the only way archivists can grapple with ever-larger volumes of materials they need to make discoverable. Archival approaches can also readily incorporate new methods to automate description from digital records. This can be anything from having a digitization vendor list newspaper dates, volume, and issue numbers in file names and using that to extract titles to emerging computer vision methods of text extraction from images as well as entity extraction and text summarization from large language models (LLMs) such as ChatGPT. Automated metadata creation is necessary to manage large volumes of digital records, but archivists must also account for the risks in these methods of

[1] Daniel A. Santamaria has best described this approach as “extensible processing.” Daniel A. Santamaria, Extensible Processing for Archives and Special Collections: Reducing Processing Backlogs (Chicago: Neal-Schuman, an imprint of the American Library Association, 2015). The Describing Archives: A Content Standard (DACS) Statement of Principles also calls for archivists to describe all materials at a baseline level and then add more detailed description based on user needs. (https://saa-ts-dacs.github.io/dacs/04_statement_of_principles.html#10-archivists-must-have-a-user-driven-reason-to-enhance-existing-archival-description)

[2] The Digitization Cost Calculator data shows that detailed metadata creation is the most time-intensive part of digitization. https://dashboard.diglib.org/.

perpetuating biases and generating inaccurate or incomplete information that will always reduce the quality of automated metadata. Hierarchical description allows archivists to manage these risks by supplementing automated description at lower levels with higher quality upper level description directly created by professional archivists.

Additionally, combining archival description and digital objects into a single access system would facilitate more digitization and make it more feasible to develop workflows for digitization initiated by user requests. Many repositories already maintain minimal file-level description for paper collections. By itself, a folder title such as "Minutes, 1988" is not a sufficient metadata record for a DAMS-style digital repository. However, this title is much more useful in an archival system like Arclight, where that title is linked to a series or collection description that shows the relevant organization and the purpose of the meeting. There is also some intriguing potential for indexing that inherited description along with the digital object at a lower weight. In many cases, this existing description could be sufficient for discovery without the need to create a more detailed metadata record for a digital repository. This would often make digitization viable where it otherwise would not be feasible. This should include digitization requests made by users, allowing repositories to slowly digitize their most requested items over time. Arclight would also allow the creation of more detailed metadata records, when archivists have user-driven reasons to increase discoverability – consistent with archival description best practices as stated in the DACS Statement of Principles.

Statement of Need

Archivists have not been able to use the extensible description methods that work so well for physical materials for digital objects because the access systems available to us were not designed to use archival methods. These systems, including Digital Asset Management Systems (DAMS), Institutional Repositories (IR), or Digital Repositories generally, all assume that every item must have a metadata record in a schema such as Dublin Core or MODS. Efforts have been made to incorporate hierarchical description into these digital repositories that generally have flat data structures, but these attempts have generally both failed to meet the requirements of archivists while introducing a substantial amount of data model complexity into these systems, making them even more challenging to implement and maintain.

This project is focused on developing Arclight into a single discovery environment for both archival description and digital objects. While repositories nationwide face this same problem, UAlbany has partnered with ESLN and Connecticut's Archives Online (CAO) to center the challenges of smaller institutions in this work and ensure that the deliverables both benefit, and are feasible for, the small and underfunded institutions that make of up the majority of archival repositories. We also partnered with some medium-sized academic repositories to demonstrate scalability over a diverse set of collections and description. In addition to implementing digital object functionality in Arclight itself, we will also create common patterns and specifications that would be replicable in tools and contexts beyond Arclight or any specific implementation.

Two Separate Systems

Archival repositories currently use two separate access systems to manage their collections: one system to provide access using archival description in the form of finding aids, and a second system to manage online digital content. UAlbany currently uses Arclight to provide access to archival description and Hyrax (a Samvera-based repository) for its digital content. For many repositories, this is the ArchivesSpace Public User Interface (PUI) and an IR or DAMS. The University of Connecticut (UConn) uses the ArchivesSpace PUI as well as the statewide Islandora instance managed by the Connecticut Digital Archive. Rensselaer Polytechnic Institute (RPI) and Union College both use the ArchivesSpace PUI as well as the Archipelago Commons digital repository. The Litchfield Historical Society uses the ArchivesSpace PUI coupled with whatever online hosting for digital objects they can find, with some items in the Connecticut Digital Archive, some in an aging local system, some hosted in the Internet Archive, and some not accessible online at all and held on local file shares – a situation which is not uncommon for many small repositories. This situation of two separate access systems is essentially true for every archival repository throughout the United States and beyond.

State-wide aggregators face the same situation and have implemented two separate systems that have siloed closely related materials. In New York, ESLN hosts EmpireADC for archival finding aids and New York Heritage for digital objects. Connecticut has the Connecticut Digital Archive for its digital objects and Connecticut's Archives Online for aggregating the state's archival finding aids. Texas has Texas Archival Resources Online (TARO) for finding aids and the Texas Digital Library hosts local DSpace instances for members to manage digital objects. California has the Online Archive of California for finding aids and Calisphere for digital objects.

This poses obvious usability problems as users must navigate multiple systems, often for items that might be right next to each other in a physical or digital folder. For example, while archival systems such as Arclight or ArchivesSpace often contain links to objects hosted in a digital repository, archival systems lack the functionality to search full-text content of documents, books, or A/V materials found in a digital repository interface, so a user may not discover relevant online materials. However, if a user searches a digital repository interface, they usually can search the full-text of an object and item-level metadata, but not relevant archival description.

Major research libraries have made several efforts to develop integrations to bridge the divide between archival systems and digital repositories, however these connections are challenging to develop and maintain and still have major usability limitations. UAlbany's Hyrax digital repository displays links to relevant archival collections and series description, and displays applicable scope and content notes next to digital content. These integrations remain challenging for users to navigate, as we at UAlbany have learned from the experience of teaching these interfaces to incoming freshmen. Additionally, the custom development work required for these systems make these integrations are beyond the resources of the vast majority of archival repositories.

Under-resourced repositories have limited choices and must try to integrate their archival finding aids with whatever digital repository that is available to them to meet even their users' most basic needs. ESLN offers statewide instances of Arclight (EmpireADC) and CONTENTdm (New York Heritage) that primarily support small and under-resourced repositories. However, these services are not connected, and archivists must duplicate their description to display in both systems. For some repositories it is only feasible to do this work manually. Even if repositories undertake this unnecessary labor, users who discover materials in New York Heritage cannot easily find related materials and users searching in EmpireADC cannot query full-text OCR text of the same content that is available in New York Heritage. Since the visibility of and demand for digital objects is so great, some small repositories have found it infeasible to maintain multiple description systems and have abandoned archival finding aids and describe objects one-by-one in New York Heritage. Others describe physical materials in finding aids but have limited or even no capacity to provide access to digital materials.

Simplifying Technology

Arclight is the best fit to provide a single access point to both archival description and online digital objects since it already provides access to archival finding aids and can be readily adapted for digital objects since it is based on and uses Blacklight. Blacklight is a faceted search and discovery interface that is used in many digital repositories and is a core component of Samvera-based repositories. Because of this, if digital object records are added to an underlying Arclight index, the application can be configured to display digital objects in search results. This includes discovery using full-text OCR and transcription content. Minimal additional customization is needed to create templates to allow navigation to and display of digital objects. Stanford University's Taube Archive of the International Military Tribunal at Nuremberg demonstrates how a customized Arclight implementation can provide discovery to both archival description and full text digital objects. UAlbany has also created a proof-of-concept demonstration instance of Arclight that similarly searches full-text digital objects alongside archival description of physical collections. These cases require both custom development that is impossible for most repositories and involves major data model and structure choices that may be incompatible with many local approaches.

For Arclight to provide access to digital objects out-of-the-box, it requires minor template adaptations as well as the development of common data models and harvesting pipelines that work for as many different archival repositories as possible. Since archival repositories use a wide variety of digital repositories to host digital objects, the emergence and proliferation of IIIF now provides a standardized method to deliver digital objects between disparate repository software systems and across institutional boundaries. IIIF manifests provide structure for digital objects and metadata and can also provide links to text content from OCR or transcriptions. IIIF is now supported by many repository platforms, supports audio and video files in addition to images, and has gained acceptance in the community as a fundamental feature. If a IIIF manifest is linked as a digital object within a finding aid, an Arclight harvesting pipeline can query this JSON file and index full text OCR and transcription content, as well as digital object-level metadata and include that information in the Arclight index alongside archival description in a single discovery interface.

An underlying goal for this project is to simplify archival systems so they become more feasible for under-resourced institutions to implement. While Arclight has proven to be a reliable and relatively user-friendly discovery system for archival description, most implementations have been limited to major research libraries, as repositories already supporting ArchivesSpace and a digital repository struggle to adopt *yet another* system. Archival repositories nationwide struggle to implement, support, and maintain systems to manage and provide access to their collections. With many repositories at or even beyond their sustainable resource limits, systems advancements must simplify and reduce overall resource and maintenance needs. Of these systems, digital repositories are by far the most challenging to implement and maintain. By using Arclight and IIIF, this project will develop and implement an approach that not only works with a variety of digital repositories, it also provides an alternative path for archival repositories, as UAlbany's implementation will use Arclight as a single search platform for both digital materials and finding aids without requiring a traditional digital repository.

Project Work Plan

To meet this major national need, UAlbany and ESLN will facilitate a collaborative community process to develop a Conceptual Model and open specifications for combining archival description and digital objects metadata and text content. These documents will be the foundation of two tools, an ArchivesSpace plugin and a description_indexer command line tool, which will act as a data harvesting pipeline for adding digital objects in Arclight. Then UAlbany, ESLN, CAO will complete three separate implementations working with seven different pilot partners, using this pipeline to index digital objects alongside finding aids in Arclight. Finally, we will undergo an iterative usability testing process using examples and tasks contributed by the seven pilot partners and Harvard Library to experiment with weighing different fields for search relevancy rankings.

Project Team

- Gregory Wiedeman, University Archivist, University at Albany, SUNY
- Katherine Mules, Applications Administrator, University at Albany, SUNY
- Jennifer Palmentiero, Digital Services Manager, Southeastern NY Library Resources Council
- Mark Wolfe, Curator for Digital Collections, University at Albany, SUNY
- Zachary Spalding, Systems Manager, Southeastern NY Library Resources Council

Advisory Board

- Sean Aery, Digital Projects Developer, Duke University
- Maureen Cresci Callahan, Head of Archives & Special Collections, University of Connecticut
- Alexander Duryee, Service Manager, Archives & Special Collections Systems, Harvard University
- Bonnie Gordon, Special Collections Analyst, Columbia University
- Regine Heberlein, Library IT Data Analyst, Princeton University and Co-Chair, TS-DACS
- Cory Lown, Software Developer, Stanford University

The Project Team will convene an advisory board of practitioners with leading expertise in working with archival data with representation from both archivists and developers. This group also includes many active members in the Arclight community, managing instances at major research libraries, and will serve as conduits to those existing institutional implementations. Overall, seven of the nine known existing or planned Arclight implementations are represented in this project. This group plans to meet bimonthly over Zoom during the course of the project to provide feedback and outside perspectives.

Pilot Partners

- Historic Huguenot Street
- Hudson Area Library
- Rensselaer Polytechnic Institute (RPI)
- Litchfield Historical Society
- Union College
- University of Connecticut
- Western Connecticut State University

We selected pilot partners based on their interest, diversity of repository size, the state of their existing data, and their access to an existing Arclight implementation. To make participation viable, partners must have existing archival description with links to digital objects in a consistent system such as ArchivesSpace. Partners also needed access to an existing Arclight implementation to index their data into. Because of these practical considerations, all pilot partners are from either New York or Connecticut as they all participate in either EmpireADC or Connecticut's Archives Online which are the only state-level Arclight instances currently in production.

In addition to the seven pilot partners, Harvard Library has undertaken a multiyear effort to assess and potentially improve the usability of Arclight and has expressed interest in participating in the usability testing phase of the project. Harvard Library will provide archival description and digital object data and develop user testing tasks for usability testing that will demonstrate the scalability of the Arclight harvesting pipeline as well as contribute their substantial experience and expertise in user experience testing for archival systems.

Phase One: Community developed Conceptual Model and Specifications

While adding digital object records into an Arclight index is very feasible technically, it opens a can of worms as the archival community lacks a common understanding of how digital objects and text content fit both conceptually and practically within archival description. This is an area that is currently undertheorized with substantial variations in practice. Thus, the first focus of this project will be to develop a collaborative community effort to better understand this challenge and develop both common system-agnostic frameworks and actionable specifications to sort out the best path forward. This takes up the recommendations of the Lighting the Way National Forum, which called for collaborative strategic planning for archival discovery and delivery using generative and care-focused facilitation methods.

The project will solicit participants via a public call to the Code4Lib, DLF, NDSA, and SAA announcements listservs aimed at practitioners with direct experience working with archival data. The project team will prioritize and make targeted outreach to ensure participation from members of marginalized and/or underrepresented groups, such as the BIPOC and LGBTQ+ communities, as well as practitioners with experience at institutions of various sizes. We expect participation from both major research libraries and one-person repositories. The project team will develop templates that will prompt participants to identify and contribute both common and outlier real-world data examples and desired use cases from their local context. This will be a hybrid process where we meet both remotely to include as many participants as possible, and a smaller group gets together in-person in a single room to easily collaborate and make rapid progress on deliverables.

A Large Cohort of 20 will meet remotely over the course of a week in October 2024 to participate in some facilitated exercises developed by the project team using anti-oppressive facilitation techniques. Exercises drawn or adapted from Liberating Structures and similar resources will be designed to foster participation from underrepresented voices, identify our fundamental purposes and the context we are working in, and define common approaches and paths forward. The goal for these sessions will be to identify a set of principles and general data structures. Local examples from participants will be tested against these outputs. Outputs will also be published on the project's webpage.

In December 2024, a subset of 8 members of the Large Cohort will meet in-person at UAlbany as the Small Cohort over the course of three days to draft the documentary deliverables that will be used for the project's implementations. This will include a generalized and system-agnostic Conceptual Model that will build on the outputs of the Large Cohort. The Conceptual Model will define a common approach for incorporating content and metadata from digital objects from archival description, include both theoretical concepts and real-world approaches for defining when digital objects are and are not representative of an archival component, and delineate the relationships between archival description and digital object metadata schemas. The Small Cohort will also draft a detailed Arclight Solr Index Specification that clearly defines how digital objects, full-text content, and digital object metadata will be managed in the Arclight index. To inform the Small Cohort, the project team will provide examples and options of how data is stored in Arclight for participants unfamiliar with Arclight and Blacklight technical patterns. Modified versions of these examples will be included in the Arclight Solr Index Specification.

In February 2025, the Large Cohort will convene again remotely to undertake a detailed review of the Conceptual Model and Arclight Solr Index Specification. The Small Cohort will incorporate changes from this feedback. These documents will then be made available on the project's website for public comment directed at the Code4Lib, DLF, NDSA, and SAA announcements listservs. Commenters will complete a Google Form and be offered opportunities for more detailed engagement. The project team will then incorporate changes and publish final versions of these documents on the project website in May 2025 that will form the basis for later development work. Announcements of publication will be made to the same listservs. Once finalized, we will work with the Technical Subcommittee on Describing Archives: A Content Standard (TS-DACS) to see if endorsement of the Conceptual Model by TS-DACS is useful and mutually desirable.

Concurrent to this process, the project team will research current approaches to find the best fit for exposing full-text content in IIIF manifests. This will include a detailed examination of leading implementations that already do this such as the Archipelago Commons repository and the From the Page crowdsourcing platform. The project team will also review the work CONTENTdm is currently undertaking to include transcriptions within IIIF manifests. The project team will consider these options and create an Arclight IIIF Specification that will define specific requirements for how Arclight will expect and handle full-text OCR and transcription content and digital object metadata in IIIF manifests for harvesting into the Arclight index. The specification will be published on the project website and shared with the IIIF Slack channel for public comment and feedback.

Phase Two: Plugin and Harvester Development

From March-November 2025, UAlbany will develop two tools for harvesting data into Arclight that will ease adoption and will parse linked IIIF manifests to incorporate digital object metadata and transcriptions in the Arclight index.

A vendor will be contracted to develop an ArchivesSpace plugin that will harvest data directly into Arclight as archivists edit and save resources and archival objects. The plugin will also parse IIIF manifests included in the description according to the Arclight IIIF Specification and include full-text content and digital object metadata in the Arclight index. The vendor will be charged to take future ArchivesSpace updates into account when designing the plugin and take approaches that limit the potential for future conflicts and reduce maintenance needs over time. Since we are a public university, this will be a public Request for Proposal (RFP) and vendor bid process managed by our university's Research Purchasing office, which will begin after the grant is awarded.

Concurrently to the plugin development, UAlbany will build out the existing description_indexer command line tool that was used for the demo Arclight instance into a full release. This will harvest description from the ArchivesSpace API, parse linked IIIF manifests for text content and metadata, and add that data to an Arclight index using

ArchivesSnake and pySolr. This work will build upon UAlbany's successful development work undertaken during the Mailbag project, funded as part of the Email Archives: Building Capacity and Community program. During that project, we developed a collaborative development process that acknowledged the reality that all contributors are progressing on their own stage in the development learning curve. The project team will build on this experience and utilize Github project boards, issue templates, and code reviews to ensure quality documentation, code readability, and test coverage.

The description_indexer tool will be developed separately from the ArchivesSpace plugin, as it uses tools and programming languages that are very distinct from ArchivesSpace and do not have an established community of vendors. Developing this tool separately also has added benefits, as detailed working knowledge of the codebase will easily allow us to make very granular changes to how data is indexed in response to feedback from usability testing in Phase Four of the project. While both the ArchivesSpace plugin and the description_indexer tool will have the same minimal requirements, they offer different affordances for different members of the community, as institutional implementations will likely favor the plugin, while aggregators should prefer the description_indexer. UAlbany anticipates implementing the plugin, but expects to also use description_indexer as a utility for managing and maintaining data in the Arclight index.

Near the conclusion of Phase Two, UAlbany will develop publicly available documentation and make minor modifications to Arclight templates to display digital objects and metadata in an Arclight instance. Since Blacklight already has extensible design patterns for metadata fields, this may only involve additions to the default Arclight Catalog Controller, as well as added test coverage and an additional template or helper method to display digital objects if a IIIF manifest is included in the index. The project team will work with the active Arclight community to add these minor changes to the core Arclight code base, building on work undertaken during the Fall 2023 Arclight community work cycle. For options that will be configurable, such as the installation of a IIIF viewer, the project team will add detailed instructions for this to the publicly available Arclight documentation.

Phase Three: Implementations

In 2025, UAlbany, ESLN, and CAO will each implement Arclight as a single discovery system for both archival description and digital objects. UAlbany's effort will begin in Fall 2024, as we work to migrate our over 163,000 digital objects from Hyrax and store them on a network file share according to a detailed specification. Each digital object package will consist of a well-defined package of files, a IIIF manifest as a JSON file, metadata in a YAML file, and OCR text or transcription content in standard formats such as hOCR, SRT, or VTT. These packages will be in the spirit of a Dissemination Information Package (DIP) in the Open Archival Information System (OAIS) model. Unfortunately, this makes these packages not a good fit to use existing specifications, such as the Oxford Common File Layout (OFCL) as they are more focused on preservation functionality. Since we do not currently use Hyrax as a preservation system, we have existing preservation storage that will be unaffected by this migration. The local specification we develop will be documented and published on the project website.

UAlbany will link the IIIF manifests created during this migration within our archival description in ArchivesSpace and configure a standalone IIIF image server to serve the files with the text content linked in the IIIF manifests according to the Arclight IIIF Specification developed in Phase One. Once the development work in Phase Two is complete, both the ArchivesSpace plugin and the description_indexer tool developed in Phase Two will be able to find linked IIIF manifests and add these digital objects to an Arclight index. UAlbany will then implement the front-end Arclight changes also developed in Phase Two and install a client-side IIIF viewer such as Mirador or Universal Viewer which will display digital objects.

ESLN's implementation will focus on the four New York Pilot Partners, representing both small and medium sized institutions. Hudson Area Library and Historic Huguenot Street are small repositories that already use EmpireADC for their archival description and have links to digital objects they host in New York Heritage. Union College and RPI are medium-sized academic archives and similarly have consistent archival description with links to digital objects, though they host their digital objects in local instances of Archipelago Commons: ARCHES and Digital Assets. RPI currently participates in and has description indexed in EmpireADC and Union College is in the process of joining and will have description in EmpireADC by the start of this project.

ESLN will implement an instance of Arclight that will incorporate the development work completed in Phase Two along with a client-side IIIF viewer. Once this is complete, ESLN will use the description_indexer tool to add digital objects into the Arclight index so they will show in search results and directly display. For Hudson Area Library and Historic Huguenot Street, the tool will harvest digital object data from New York Heritage and for RPI and Union College it will harvest this data from their local Archipelago commons digital repository instances. Digital objects from all these sources will then display in EmpireADC alongside archival description, demonstrating that this approach is feasible for a wide variety of institution sizes and repositories.

CAO will also implement an instance of Arclight that incorporates the development work completed in Phase Two along with a client-side IIIF viewer. Their effort will focus on three Connecticut Pilot Partners, the University of Connecticut (UConn), the Litchfield Historical Society, and Western Connecticut State University (WCSU). All these repositories also have consistent description in ArchivesSpace with existing links to digital objects and currently participate in CAO, a statewide Arclight instance hosted at WCSU. Both UConn and the Litchfield Historical Society host digital objects in the Connecticut Digital Archive (CTDA), a statewide Islandora instance hosted at UConn. CTDA recently completed an Islandora update that utilizes IIIF. The Litchfield Historical Society also has some objects hosted by the Internet Archive that also supports IIIF. CAO will use the description_indexer tool developed in Phase Two to index digital objects from each institution including digital objects from these multiple sources into this Arclight instance.

In total, this phase will complete three separate implementations across eight institutions of various sizes working with five different digital repositories or methods of hosting digital objects, demonstrating the feasibility of this approach in a variety of different local contexts.

Phase Four: Iterative Usability Testing and Experimentation

Joining both finding aids and digital object metadata and transcriptions in a single system will provide both usability challenges and new opportunities to experiment with weighting and inheriting description according to archival theory and standards. This includes both the weight of content fields containing OCR or transcription text and metadata so that digital objects are visible in results but do not overwhelm titles and notes. There is also interesting potential for indexing file, series, and collection level metadata with digital object records at lower weights. The project will examine this by undertaking iterative user experience testing.

Once implementations are complete by spring 2026, the project team will provision additional Arclight test instances of data from both UAlbany and the pilot partners that includes digital objects. These test instances will allow us to use the description_indexer tool to change and experiment with the weights of how fields are indexed in Solr. The Project Team will develop tasks for usability testers collaboratively with the pilot partners who will have a more thorough understanding of their description and how their collections are typically used. Harvard Library will also provide examples and develop user tasks from their collections. These will also be indexed into a test Arclight index which will not only help us demonstrate scalability, but their participation in user testing will help us build on their experience with usability testing for both Arclight and ArchivesSpace. Additionally, we will invite participation from the Society of American Archivists User Experience Section to see if there is any interest from archivists in designing tasks and viewing users navigating archival data. We hope that this both invites existing outside expertise and provides a pathway to directly share this usability testing experience and lessons learned with the professional community to have a border impact beyond this project.

The Project Team will solicit user testing participants from UAlbany’s diverse undergraduate student body, over 40 percent of which are from historically underrepresented communities, including a higher percentage of BIPOC students than the New York state population. We will purposely avoid weighting participation towards users with familiarity with archives or expertise in current finding aid systems, such as history faculty who may favor current access paradigms. If we design systems that will be usable for uninitiated users, all users will benefit. We anticipate three rounds of testing with a pool of 10 students each round for a total of 30 participants, who will each be provided $50 gift cards to incentivize their participation.

Participants will undergo both mediated and unmediated usability testing by following tasks using the Arclight test instances. For mediated testing, participants will be asked to schedule a test time using Calendly and be observed over Zoom. For unmediated testing, participants will be provided a Zoom room that will be set to automatically record participants when they enter. All participants will be asked to complete brief surveys after testing, which will gather consistent data in addition to qualitative observations. Once clear conclusions can be drawn from a round of testing, the Project Team will re-index description in the test instances at different weights and undertake another round of testing with different participants. We will publish a final report of our findings in an open access repository. The Project Team may make minor user interface changes over the course of testing, but major navigation issues outside of search results that we are not able to address will be thoroughly documented as issues in the Arclight Github to inform future community development sprints. Final configuration for the Arclight Solr index will be derived from this work and added to the Arclight core code base.

Diversity Plan

The project team acknowledges that its white, cisgender backgrounds are heavily overrepresented in the archives profession, particularly among archivists that work with technology. In selecting cohort participants, the project will prioritize and make targeted outreach to members of marginalized and/or underrepresented communities to both ensure a diversity of voices and help professionals from a broader variety of backgrounds gain confidence and experience in this space. In selecting usability testing participants we will draw from a diverse population, ask volunteers to self-identify during a selection survey, and weight participation towards identities and experiences that are marginalized within the professional archives community, such as the BIPOC and LGBTQ+ communities. Overall, this project is committed to care-focused and anti-oppressive facilitation methods such as "taking stack," or exercises from Liberating Structures. These structures attempt to ensure that conversations are balanced, fostering a welcoming space that encourages contributors from diverse backgrounds and experiences.

Project Results

This project will build Arclight into a single access platform for both archival description and digital content, allowing institutions to preserve both archival context and their valuable descriptive labor while making it feasible to make a larger volume of materials discoverable and accessible online in an environment that provides greater discovery and a better user experience. The Project Team will present this work at major national conferences, including SAA, DLF, and Code4Lib, as the UAlbany, EmpireADC, and CAO implementations will provide clear examples of a new paradigm for how archival repositories can conceptualize access systems, and demonstrate the viability of this approach for a diverse set of institutions and digital repositories. Beyond any specific implementation, this project will also establish broader patterns for access systems and develop system-agnostic specifications and open documentation that will enable archivists to undertake this work in a variety of different contexts and systems.

This project will work to reduce implementation barriers and resource costs for systems in two ways. First, the project will develop an ArchivesSpace plugin that will lower the barriers to implementing and maintaining Arclight alongside a backend ArchivesSpace instance. The plugin will be a direct integration that will index ArchivesSpace resources, archival objects, and digital objects in Arclight as records are saved in ArchivesSpace, avoiding the need to add and maintain a separate workflow to export and index EAD. This will also eliminate substantial data noise between the two systems due to the legacy of EAD and make it easier for medium and large archival repositories to implement Arclight beyond the current group of major research libraries.

More importantly, this project will make it possible for smaller repositories to directly utilize the usability advantages of Arclight through participation in state and regional consortia. Large consortia, such as ESLN who hosts two separate systems with over 400 participating institutions could eliminate these silos and provide the option for participants to have a single point of access for archival description and digital objects. Small repositories will be able to describe and provide access to digital materials using archival methods and best practices that are currently unavailable to them due to system constraints.

Additionally, as Arclight takes on the role of public access to digital objects, this work will also reduce the responsibilities for digital repositories and create new possibilities. In addition to working with a variety of IIIF-compliant digital repositories, Arclight will make it feasible for archival repositories to provide access to digital materials without the need for a traditional digital repository at all. UAlbany's local implementation will manage digital objects and provide IIIF manifests from network file shares according to a local specification, an approach that may be easier to maintain for under-resourced institutions than complex digital repositories. For this approach, a IIIF image server is helpful to provide zoom, rotation, etc., but is not necessarily required. Using this approach, small repositories could provide access to digital materials with only access to an Arclight instance and a web server serving content from a network file share.

Using Arclight as a single system for both archival description and digital materials will enable repositories to apply the same archival methods and best practices that work well for large volumes of physical materials to digital materials. Instead of requiring detailed metadata records for every digital object, this will empower practitioners to use their professional judgment to match their limited time and resources to best meet user needs like they do for physical materials. This model also makes it possible to provide integrated access to born-digital materials, which often include large volumes of files that must be described in aggregate. Emerging methods of automated description are necessary to manage this volume, and archival hierarchy helps to mitigate the risks of these methods by connecting automated description with professional quality metadata records created by humans.

Using Arclight as a common platform for description and digital objects also allows repositories to maintain the important connections between these records. Due to current system limitations, many repositories nationwide currently digitize items and separate them from existing description to place them in incompatible digital repositories which risks losing the valuable archival context that preserves how these materials were created and used. Using Arclight for both description and digital objects allows repositories to maintain these important relationships.

A single discovery point for both archival description and digital objects also enables archivists to make a larger volume of digitized materials available online using existing description instead of requiring resource-intensive detailed metadata records. This will reduce costs for the many types of materials that would only require minimal description, such as UAlbany's student newspaper, which only requires dates, volumes, and issue numbers to be sufficiently discoverable by users if that metadata can be returned in search results alongside full-text content. This also enables digitization by user request, as many items can be rapidly digitized using overhead or even sheet-feed scanners and made publicly available online without extraordinary resources, as UAlbany's productive program can attest. Current digital repositories often make these services unviable due to metadata costs and items digitized for users are deleted or sit on in-house file shares in repositories throughout the United States.

The challenge of using archival methods for digital objects applies to all repositories, large and small. For small repositories Arclight can make this feasible where previously their only options were to either use finding aids and provide limited or no access to digital objects, or break up each digital object, create a detailed metadata record for each and provide standalone access in a resource-intensive digital repository. For major research libraries, this approach will make their description resources go further, allows them to make a greater volume of materials available online, and enables them to provide access to born-digital materials at scale. For repositories of all sizes, Arclight will enable them to digitize on user request, maximize the benefits of archival description, and more efficiently steward their limited descriptive resources to better meet user needs.

Year 1 (August 2024 - July 2025)												
Activities Year 1	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**
Phase 1: Community developed Conceptual Model and Specifications												
Develop and release public call for participation	■											
Review and select applicants	■	■										
Research IIIF approaches		■	■	■								
Develop Arclight IIIF Specification				■	■	■	■					
Large Cohort First Meeting			■									
Small Cohort Meeting and draft documents					■							
Large Cohort Review Meeting							■					
Public call for comment on documents and incorporate feedback								■	■			
Release Final versions of documents										■		
Phase 2: Plugin and Harvester Development												
Develop and release RFP with Research Purchasing	■	■										
Vendor Q&A period			■									
Vendor Bid Due and Evaluation				■								
Bid awarded					■							
Vendor plugin development								■	■	■	■	■
Development of description_indexer tool								■	■	■	■	■
Phase 3: Implementations												
UAlbany digital object migration planning	■	■										
UAlbany digital object specification design		■	■									
UAlbany digital object migration				■	■	■	■	■	■	■	■	■

Year 2 (August 2025 - July 2026)

Activities Year 2	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul
Phase 2: Plugin and Harvester Development												
Vendor plugin development	■	■	■	■	■							
Development of description_indexer tool	■	■	■	■								
Arclight template/controller development			■	■								
Documentation writing for implementors				■								
Phase 3: Implementations												
UAlbany digital object migration	■											
UAlbany data indexing		■	■	■								
EmpireADC data indexing Pilot Partner data		■	■	■								
CAO data indexing Pilot Partner data		■	■	■								
UAlbany implement Arclight changes					■							
EmpireADC implement Arclight changes					■							
CAO implement Arclight changes					■							
Phase 4: Iterative Usability Testing and Experimentation												
Provision test instances						■						
Task design						■	■					
Development of participant solicitation and guidance							■	■				
Round 1 testing									■			
Round 2 testing										■		
Round 3 testing											■	
Implement indexing weights from testing												■
Document results for publication											■	■

Digital Products Plan

Type

This project will develop several digital products including formal and informal documents, code, and video recordings.

Documents

- Conceptual Model
  - A document outlining the conceptual relationships between components of archival description and digital objects. This will be written in Markdown.
- Arclight Solr Index Specification
  - A detailed specification for how digital object records should be stored in an Arclight Solr index. This will be written in Markdown and follow the format of a Request for Comments (RFC) proposals.
- Arclight IIIF Specification
  - A detailed specification for how to link full-text OCR and transcription content within IIIF manifests, so that they can be harvested into an Arclight index. This will be written in Markdown and follow the format of a Request for Comments (RFC) proposals.
- Implementation documentation
  - Detailed instructions for implementing Arclight with digital object, such as how to install a client-side IIIF viewer. This will be written in markdown.
- Publication
  - The Project Team will write and publish an article on the results of the Phase 4 usability testing. This will be written in a word processor and likely be published as a PDF.

Source Code/Software

- description_indexer
  - A Python command line utility for harvesting archival description in EAD and ArchivesSpace for indexing in Arclight.
- ArchivesSpace plugin
  - A JRuby plugin that directly indexes description from ArchivesSpace into Arclight.
- Arclight template work
  - Additions and alterations to existing Arclight Ruby on Rails templates to display available digital objects within International Image Interoperability Framework (IIIF) viewers.

Video Recordings

  - 30 usability testing sessions recorded using Zoom as MP4s and M4As.

Availability

All documents and code will be made available online. The Conceptual Model, Arclight Solr Index Specification, Arclight IIIF Specification, and implementation documentation will all be written in Markdown and stored in a public Github repository and be linked from the project website. The project team will aim to publish the usability testing publication in a "gold" open access journal. If that is not realized, the document will be made available in Scholars Archive, UAlbany's open access institutional repository.

All source code will also be openly available. The description_indexer and ArchivesSpace plugins will be available in a public Github repository. The Arclight template work will be added to the existing Arclight Github repository.

The video recordings of the usability testing sessions will not be made openly available and will only be used during the course of the project to protect the privacy of participants. These recordings will be only shared with Pilot Partners and other project collaborators and be disposed of at the conclusion of the project.

Access

All digital products will be made available under permissible licenses. The Conceptual Model, Arclight Solr Index Specification, Arclight IIIF Specification, and implementation documentation will all be made available under a Creative Commons Attribution license (CC-BY).

All code developed for the project will be made available for use and reuse under permissible open source licenses. The description_indexer and ArchivesSpace plugin will be made available under the MIT License. The Arclight project uses the Apache 2 license and all changes to Arclight templates will be made available under those same terms.

Sustainability

The focus on strong documentation in this project is designed to assist the sustainability of the outcomes as documents are typically easier to sustain than software. The Conceptual Model and two Arclight specifications will not only assist in software design, but they are also designed to ensure that the progress made by this project will outlive the software. In a long enough time horizon, all software will become obsolete and need to be replaced over time. Detailed design documents help explain our incremental progress so that future maintainers can build on previous work instead of reinventing the wheel.

Additionally, our experience with the Mailbag Specification has shown than if you undertake an open design process that incorporates community needs and feedback, these types of specifications can be relatively stable, as we have not received any issues or requests for changes in over two years at the time of writing. The Bagit specification is very widely used in the community, and similarly has not had any content changes in over 10 years since its 1.0 release. If the Arclight specifications become broadly implemented as we anticipate, we will seek an institutional home for these documents beyond UAlbany, such as a professional organization or the IIIF Consortium, as appropriate.

That said, the sustaining description_indexer and the ArchivesSpace Plugin is also essential for ensuring that implementors can maintain and sustain use of Arclight as we are proposing. While this is much more challenging, UAlbany is committed to maintaining the description_indexer tool for at least the medium term, as this will be a key part of our digital repository infrastructure. We hope that with broader adoption, the tool will also receive some long-term support from the professional community. The ArchivesSpace community is robust, and the Arclight community is burgeoning, with regular community calls starting in Fall 2022. The community recently completed a successful community sprint in December 2023. While this group has mostly been limited to major research libraries, this project should make it more feasible for a broader and more inclusive group of implementors, which would greatly strengthen this growing community.