University of Utah, J. Willard Marriott Library - Historical Newspapers in Hydra

January 12, 2017

Abstract

The University of Utah (UU)'s J. Willard Marriott Library and the Boston Public Library (BPL) are seeking $249,999 to fund a software development project to enhance support for historical newspaper content in digital cultural heritage repositories. The goals of the project will be to create: (1) a shareable, system- and programming-language-agnostic RDF-based data model addressing structural and descriptive metadata features unique to digitized newspapers; (2) a set of modular, open-source plugins for the popular Hydra/Fedora digital repository framework for ingesting, describing, discovering, displaying, and disseminating digitized newspaper content; and (3) a community of practitioners -- including developers, librarians, content specialists, and managers -- dedicated to addressing challenges and collaborating on best practices associated with managing digitized newspapers.

The proposal describes a two-year schedule for design, software development, and community engagement to take place from July 1, 2017 through June 30, 2019. The domain model will be built using RDF (the lingua franca of Linked Data and the Semantic Web), covering all common forms and file types of digitized newspaper content, for both physical (microfilm reels, pages, digital files) and intellectual (issues, articles) objects, as well as descriptive metadata specific to newspaper content. The software deliverables, providing both administrative and display components, will be Ruby gems, which are modular by design and easily integrated into Hydra-based digital asset management applications (including the IMLS-funded "Hydra in a Box"). The open-source code will be freely available on GitHub.com and RubyGems.org. By enabling this functionality as a plugin (rather than yet another stand-alone system), this project will provide standardized tools and workflows for institutions to enable management, discovery, and aggregation of newspaper content that can be integrated into existing cyber-infrastructure.

The project addresses an acute need among institutions providing digital access to cultural heritage materials. Newspaper content poses unique challenges for management and dissemination, especially in terms of article-level access. Finding relevant materials in online digital collections can often be a difficult task, a problem that will only increase as more content is digitized and projects such as the Digital Public Library of America (DPLA) aggregate metadata records from thousands of institutions into a single discovery environment. Our vision is to extend support for handling digitized newspapers in the Hydra/Fedora platform by enabling more granular levels of categorization, better discoverability, and more relevant refinement of results. With modeling and encoding focused around article-level objects, it will be possible to develop additional search and visualization features to more efficiently present large amounts of information and offer an improved end-user experience.

BPL and UU will coordinate the modeling and software development process, working closely with community partners and utilizing agile and test-driven development methodologies. The initial phase (4 months) will be devoted to determining requirements and hiring a developer. Development work will be iterative, loosely organized into three phases (6 months each), focusing on administrative functionality, interactive functionality, and implementation. Each phase will include significant time for community review, testing, evaluation, and bug fixes, which will inform the upcoming development cycles. The final months of the project will be devoted to assessment, outreach and promotion.

The primary audience will be libraries managing digitized newspaper collections, but we believe this project will have a substantial impact on both the management *and usage* of digitized newspapers. First, this project will create a standard way to handle newspapers in the rapidly-growing Hydra framework. Second, this project will establish a set of conventions and protocols for disseminating newspaper content in large-scale aggregated contexts such as DPLA, serving as a model across the DPLA network. Third, it will open up access to newspaper content at a much more granular level for information-seekers, connecting them to relevant information more effectively. And fourth, it will create pathways for machine processing techniques such as text mining, automated classification, and the automated conversion of page-level newspaper content into article-level content.

Historical Newspapers in Hydra: Building a Platform to Restore Access to Cultural Treasures

1. Statement of National Need

"The newspapers are making morning after morning the rough draft of history." --The State, December 5, 1905

Historical newspapers represent an invaluable resource for scholars, historians, students, genealogists, and the public, connecting us with the record of our ancestors, and bringing voices from past eras to life again. Libraries have in turn made the digitization of newspapers a priority over the past few decades, resulting in an overflowing abundance of digital images and extracted text that are now available through digital repositories and online search tools. However, the zeal with which these digitization efforts have been carried out has not been matched by improvements in the systems used to manage and disseminate historic newspaper content. Too often digitized newspaper collections are difficult to search, segregated from other relevant digitized materials, and incompatible with next-generation web-based protocols for decentralized discovery, sharing, and reuse of resources across networks. Whitelaw (2015; for references please see **Supporting Document 1: Works Cited**) describes this issue very succinctly: "Decades of digitisation have made a wealth of digital cultural material available online. . . . Yet in response to this abundance, collection interfaces wheel out miserly lists, one page at a time. . . . As an interface, search fails to match the ample abundance of our digital collections and the generous ethos of the institutions that hold them."

Newspaper content poses unique challenges for management and dissemination, especially in terms of article-level encoding, modeling, and access. As Lorang and Soh (2013) have noted, "Locating relevant materials in digital collections is already often a difficult endeavor and will become increasingly so as more content is digitized and as projects such as the Digital Public Library of America (DPLA) aggregate content from dozens of institutions into a single environment." There is an acute need in the library community for open-source digital asset management (DAM) systems that address these issues and are responsive to the evolving landscape of how digital collections are aggregated and accessed, not only by users but by automated computational processes vital to the emerging fields of digital humanities and civic data initiatives.

The lack of support for article-level content management and access in current DAMs is a significant problem; articles are the basic intellectual unit of information in newspapers, but are not always well-served by levels of digitization and access that are structured around physical manifestations such as issues and pages. Articles often span non-sequential pages within a single issue, and researchers may want to parse articles spread across several issues to get the data they want (Tranouez, 2015). News articles often report an evolving news story; later articles expand upon earlier ones, and a user viewing any of these stories may want to refer to other relevant articles (Allen, et al., 2007). The information seeking behavior of users points to the importance of putting articles at the center of any digitized newspaper collection. Current trends in the newspaper digitization point to a growing tendency to open up newspaper collections at the article level (Klijn, 2008). Given this centrality of articles in newspaper collections, it is imperative to build support for article-level modeling for digitized newspapers in evolving DAMs.

Most repository systems also lack support for mechanisms to enable sharing and aggregation using emerging protocols and standards. In order to provide this functionality, updated standardized domain models and interchange formats must be developed. While some XML-based standards exist (e.g. METS/ALTO), there is a pressing need for a framework for modeling and describing digitized newspaper content using RDF, which is the lingua franca of Linked Data and the Semantic Web. The International Image Interoperability Framework (IIIF[1]; see **Supporting Document 2: Web links for technical terms and products**) has developed a set of specifications for sharing and reuse that have been widely adopted by museums, libraries, and archives. These standards show tremendous promise for increasing the accessibility and usage of digitized cultural materials and provide a stronger platform for web-based aggregation, computational analysis, and data mining. The significance of digitized newspaper content demands systems that natively support these functions.

There are many repository systems (both commercial and open source) that offer varying degrees of support for digitized newspaper content, but all of them invariably fall into one of two categories: (1) standalone DAMs built specifically to manage digitized newspaper content, or (2) generic DAMs that support variety of materials but lack advanced features for any one type of content. The former category poses a problem when institutions want to consolidate their infrastructure and integrate support for different content types into one system. Chronam[2], open-oni[3], and Veridian[4] are all examples of systems that fall under this category. While these systems do provide more newspaper-centric features (such article-level access and/or IIIF API endpoints), they are extremely limited in terms of the data formats and file types that can be ingested, and institutions must maintain multiple DAMs to host digitized newspaper content and other types of collections. This "silo-ing" of newspaper content also limits its usefulness, since connections to an institution's other digitized materials (photographs, manuscripts, etc.) are not evident to the user. The latter category includes systems such as DSpace[5], CONTENTdm[6], and BePress Digital Commons[7]. While these products may allow institutions to ingest multiple content types, their generic architecture and design prevents more granular usage, lacks scalability for large collections, and extending support or developing plugins for additional features is often difficult or impossible.

As organizations whose missions are to develop and foster learning environments, share historical cultural treasures, and empower generations of learners, we believe that the best way to further research and scholarship related to newspaper content is to build feature-rich, user-focused systems using open-source, scalable, community-supported platforms. Over the past 5 years, Hydra[8] and Fedora Commons[9] have emerged as the predominant frameworks for cultural heritage institutions seeking to develop DAMs that support robust storage, management, and dissemination of vital historical materials. Built on a foundation of widely-used, community-driven open-source technologies with an established track record of success in enterprise-level implementations, Hydra and Fedora represent an ideal platform because of their scalability, flexibility, modularity, and preservation capabilities. With continued financial and development investments from dozens of public and private partners, as well as funding agencies such as IMLS, Mellon Foundation, and JISC, the Hydra community has grown exponentially since its debut in 2008 (currently featuring contributions from over 130 developers from over 60 institutions), and adoption is set to increase even more dramatically with the release of the "Hydra in a Box" turn-key system[10], which will include a hosted service provided by DuraSpace, expanding the ability of institutions of all sizes to implement this technology.

It is clear that Hydra and Fedora provide the necessary technical underpinnings and the vibrant community to support a 21st-century digitized newspaper management system. Using these technologies as the platform will ensure the long-term sustainability of newspaper collections and provide substantial, lasting benefits to the cultural heritage community. However, Hydra does not currently offer support specifically for managing newspaper content, a glaring gap that needs to be addressed as soon as possible. We are proposing to build pluggable, modular software to provide robust features for managing digitized newspaper content in Hydra, including an RDF-based domain model and support for a wide variety of formats and file types, enabling more granular levels of access and increased accessibility and dissemination for these important materials. With modeling focused around article-level objects, it will be possible to develop tailored search and visualization features to more efficiently present large amounts of information and offer an improved end-user experience.

Our approach differs significantly from many projects in that we are proposing building modular, pluggable pieces of code rather than a separate, stand-alone application. An institution would not be forced to run a limited-focus piece of software for their digital newspapers, thereby increasing adoption, lowering management costs, and desegregating newspaper content from other digital collections. The code will be designed to be integrated into any system using the Hydra/Fedora framework (including IMLS funded projects like "Hydra in a box"), and will be available to the broader Hydra community as a plug-in or add-on (rather than a replacement) to their existing repository infrastructure, providing a standardized way of managing digital newspapers alongside other digitized content.

Both Utah's Marriott Library and BPL have substantial experience developing open-source software, and are well-poised to execute this project in order to achieve a lasting national impact and further IMLS goals to improve the management, access, and use of content and collections. Marriott Library has been managing the Utah Digital

Newspapers (UDN) program[11] since 2002 and is a recognized leader in newspaper digitization, partnering with universities, colleges, public libraries, and other municipal agencies to provide access to a vast wealth of historical newspaper collections (over 22.1 million articles and ~1.784 million pages). The BPL is a contributing Hydra partner and maintains Digital Commonwealth,[12] a statewide digital repository for Massachusetts and DPLA content hub with over 190 member institutions. BPL is currently involved in a massive regional newspaper digitization effort, representing a rich variety of unique local newspaper content.

This project will directly address National Digital Platform goals in a number of significant ways. By adding crucial missing functionality to Hydra for newspaper content management and dissemination, we will develop and improve open-source digital library tools. Interoperability of digital assets will be enhanced by the implementation of a platform-agnostic Linked Data-aware RDF-based domain model for digitized newspapers content, as well as standardized APIs to support aggregation, sharing, and reuse. This interoperability will in turn allow for more efficient machine-based access to image and text content, opening up pathways for the intersection of digital libraries and cutting edge work in the fields of digital humanities, civic data engagement, distributed search, computational analysis, and the Semantic Web. With web-scale cultural heritage content aggregation growing in importance, standards for data modeling and interchange maturing, and open-source community collaboration flourishing, the time is right for a project to improve access to the rough drafts of history, connect scholars to the materials they need, and expose the hidden treasures in historical newspapers.

2. Project Design

The proposed project has the following goals:

- To facilitate the creation and promotion of a shareable, RDF-based data model for digitized newspaper content using the Portland Common Data Model[13] framework (PCDM), including modeling for both physical (microfilm reels, pages, digital files) and intellectual (issues, articles) objects, as well as descriptive metadata specific to newspaper content.
- To develop and release a set of open-source software components that provide functionality for ingesting, managing, and disseminating digitized newspaper content using the Hydra and Fedora application framework.
- To foster a collaborative community of cultural heritage institutions and software developers (from the academic, public, and private sectors) in order to produce consensus on best practices regarding modeling, management, and dissemination of digitized newspaper content.
- To establish a set of best practices for sharing and aggregation of digitized newspaper content across digital library systems.
- To increase the discoverability (SEO[14]) and usage of digitized newspaper content by scholars, researchers, genealogists, journalists, students, and the general public.
- To increase the availability of digitized newspaper content for web-based machine processing techniques such as text mining, automated classification, and feature extraction.

The successful completion of this project will have the following outcomes:

- A comprehensive data model for all common forms, structures, and file types of digitized newspaper content, expressed as RDF, addressing structural and descriptive metadata features unique to digitized newspapers.
- A set of freely-available, open-source software "plugins," written in the Ruby language (known as "gems"[15]), that can be easily installed and deployed in Hydra-based digital asset management systems, which provide functionality for ingesting, managing, and disseminating digitized newspaper content.
- A self-sustaining, engaged community of developers and content managers from a variety of organizations that serves as a forum for collaboration and discussion of best practices for working with digitized newspaper content.
- An increase in the amount of digitized newspaper content being shared with digital library aggregations such as DPLA.

- An increase in the accessibility and usage of digitized newspaper content by a wide variety of users, facilitated through digital library systems that provide access to newspaper content using protocols and practices developed and promoted during this project, such as RDF for Linked Data-friendly metadata and IIIF APIs for interfacing with images and OCR text.

These goals directly address the national need for digitized newspaper content management within digital library systems built on the increasingly-widespread Hydra framework. By expanding the functionality of these systems to provide more granular access to newspaper content and increased interoperability, these objectives will also serve the stated IMLS National Digital Platform objectives of "improving interoperability, usability, or user community involvement" with digital library tools.

Assumptions: Since we are proposing to create a set of software components that can be readily integrated into existing (or future) Hydra-based systems -- rather than yet another stand-alone application that exclusively focuses on a particular content type -- this project is built on the assumption that Hydra and Fedora will continue to be a strong, sustainable platform for digital asset management system development. There is substantial evidence to support this; both Hydra and Fedora have a proven track record of engaged community involvement, along with a steadily-increasing number of implementations in libraries, museums, and archives of many different types across the globe. Hydra and Fedora also have fiscal sponsorship from DuraSpace, an independent nonprofit supporting open technology projects that facilitate the management of digital cultural heritage materials, which provides another layer of stability. The rapid growth of the Hydra community over the past 8 years (from 3 partner institutions in 2008, to 34 current partners), as well as its diversity (including prominent research universities, major public libraries, and public broadcasting networks) provides more evidence of the viability of this software platform.

This project also assumes the existence of an active development community that is willing to work together to provide feedback on proposed models and design decisions, as well as test iterations of the software as they are released. This idea is founded on the vibrancy and participatory spirit of the broader Hydra and Fedora communities, which collectively have hundreds of contributors to their codebases. In these communities, communication between developers, managers, and other technologists takes place over a variety of channels, including Google Groups, email listservs, conference calls, interest group meetings, and in the online project management tool Slack. The demonstrated interest in the recently-formed Hydra Newspapers Interest Group[16] (chaired by co-Principal Investigator Eben English), which at the time of this writing features representatives from 14 institutions, also supports this assumption. (See **Supporting Document 3: Hydra Newspapers Interest Group membership**.)

Risks: According to our project design, the overwhelming majority of the actual development work will be executed by a software engineer to be hired using grant funds. To counter the risk of being unable to find a qualified developer, we will conduct a nationwide search for the position and allow the developer to work remotely, which will significantly increase the pool of available talent. The project will budget ample time (3 months) for the developer search and hiring process, and the project schedule is front-loaded with activities that will principally be carried out by representatives from BPL and Utah, such as gathering requirements, modeling, and design. Therefore, an extended search for a qualified developer would not greatly affect the project schedule.

The most significant risk to achieving the project's objectives would be a lack of adoption from the broader Hydra and Fedora communities. This would most likely be caused by the production of software deliverables that do not meet the needs of community stakeholders, leading to lack of interest and use. To counter this risk, several strategies will be employed: community engagement, iterative development, and frequent releases of the software. This project will engage with community stakeholders at a number of levels. We will convene an Advisory Committee composed of institutions with significant experience managing digitized newspapers and who are highly vested in the proposed deliverables, which will meet regularly to review project progress and provide feedback on system requirements, architecture, and design. We have commitments from 12 academic and industry leaders to participate on the Advisory Committee so far, including representatives from Lyrasis, Newspapers.com, DPLA, Hydra-in-a-box, CRL, Yale, Cornell, Princeton, etc. (See **Supporting Document 4: Advisory Committee**).

Members of the core project team will also discuss requirements, modeling, and code at meetings of the Hydra Newspapers Interest Group, a monthly forum for exploring management of digitized newspapers within Hydra- and Fedora-based systems. Lastly, we will provide frequent updates via established software development community communication channels, detailing project progress and soliciting testing and feedback, which will increase the visibility and awareness of the project.

The iterative development methodology employed by the project will also help safeguard against creating deliverables that either do not meet community requirements or that cannot easily be implemented in current Hydra-based repository systems. Releasing frequent, incremental versions of the software will allow for testing and feedback from stakeholders throughout the project timeline, and will bring to light issues such as missing features, difficulty with installation, incompatibility with versions of software dependencies, or other errors, and will reveal new requirements that may not have been illuminated during the Design phase. These issues will be recorded by the project team, integrated into upcoming development cycles, and re-tested after the next release. The project development schedule also includes a "Launchpad" phase, which will specifically focus on implementing the software in existing Hydra-based systems, automated installation scripts, documentation, and migration of newspaper content from other systems. Together, these measures will greatly reduce the risk of delivering products that do not address community needs or that cannot be readily integrated into Hydra installations already in use.

Project Activities: The project activities will begin with an initial Design phase, during which the project team will work closely with community stakeholders to formalize the functional requirements the deliverables need to support. The subsequent development work will be loosely organized into three phases, focusing on administrative functionality ("Admin"), interactive functionality ("Interactive"), and system implementation ("Launchpad").

In the Design phase, requirements will be elicited using a variety of methods, including surveys sent out to the Hydra and Fedora communities, analysis of existing newspaper content discovery systems, and meetings of interested stakeholders conducted in-person and via conference call. The project team will then synthesize the requirements into a set of user stories, which are descriptions of the functionality written from the perspective of a user of the system (either an end user or an administrative user). These user stories will then be used to organize the actual development work.

Another main activity during the Design phase will be the creation of a comprehensive data model for all common forms, structures, and file types of digitized newspaper content, expressed as RDF, which addresses the unique characteristics of this material. Project staff will create a structural metadata profile using the Portland Common Data Model (PCDM) framework -- a flexible, extensible domain model that is intended to support a wide array of repository applications and services -- to express the relationships between intellectual entities (such as volumes, issues, and articles) and digitized file-level content (images and OCR text). The data model will also establish a baseline set of metadata to express descriptive information particular to newspapers that is not well-supported by RDF serializations of common schemas such as Dublin Core or MODS. The model and recommendations produced will be platform-independent and therefore not limited to the Hydra community, being applicable to any digital library system capable of supporting Linked Data RDF, either as a data storage model or a platform for sharing and aggregation.

The products of the Design phase will be submitted for review and discussed with both the project's Advisory Committee as well as the Hydra Newspapers Interest Group, and will be revised based on feedback provided. This process of discovery, review, and revision will be repeated periodically throughout the project life cycle to ensure that the development activities and deliverables remain responsive to user needs, which are constantly evolving.

While each development phase will have a primary focus on a particular set of requirements, and multiple software modules providing different functions will be created during the project, our broader approach will be to view the components as an integrated product being developed iteratively. Each phase will include significant time devoted to community review, testing, evaluation, and bug fixes. The project requirements and user stories will be revised and updated at the beginning of each phase to reflect the results of the previous stage, so that new or evolving user

needs are addressed as they arise. (For example, so that administrative features can still be addressed throughout the life cycle of the project, or that interactive-focused features are not invalidated by design decisions during the administrative phase.)

The Admin phase will focus on creating a Ruby gem implementing the domain model created during the Design phase and fulfilling the requirements of administrative users of the system. Such requirements may include ingest workflow, batch import, metadata creation and editing, file management, creation of image derivatives, indexing for search, and other needs. This phase will also involve the creation of an API conforming to the IIIF Presentation API specification, which is a protocol for describing the structure of a multi-image object in order to support a feature-rich online viewing environment. This API will not only support the functionality of the display components of this project, but will provide a standard interface to newspaper content that can be accessed by any application that supports the IIIF API, opening up pathways for machine-based processing and computational analysis. We strongly believe that providing support for the IIIF standard will greatly increase the impact of this project and lead to wider dissemination of newspaper content.

The Interactive phase will focus on creating a Ruby gem fulfilling the requirements of "public" users of the system, such as sequential viewing of images, deep zooming, searching within OCR text, browsing within volumes and issues, metadata display, keyword search result highlighting, image download, and other requirements identified in the Design phase. The code architecture will seek to use standard, programming-language-independent APIs such as the IIIF Image and Presentation APIs, with the goal of making this gem as modular as possible, so that this software could be implemented independently of the administrative gem.

The Launchpad phase will focus on maximizing interoperability with existing Hydra projects (such as "Hydra in a Box"), supporting protocols for sharing of content with aggregators (such as DPLA), working to automate the installation process wherever possible, thoroughly documenting the code, and developing scripts for batch import of content from proprietary systems (such as CONTENTdm) into the open-source Hydra/Fedora repository framework. This work will overlap with an Integration phase devoted to advising and consulting with institutional stakeholders as they attempt to integrate the software into their existing repository applications and workflows. The feedback from these implementation projects will be continually integrated into the work plans for the Launchpad phase. (The Integration phase will also involve promotion and outreach, see below for more information.)

All deliverables will be hosted on GitHub.com and made freely available via an Apache 2.0 software license. The structural metadata model will be expressed as a PCDM profile document and hosted in the PCDM wiki, which is managed by DuraSpace. The gems will each have their own GitHub repository underneath the Project Hydra GitHub organization, and will also be downloadable from Rubygems.org.

Project Methodology & Management: The development process will be managed using processes informed by Agile development frameworks such as Scrum, which emphasize adaptive planning driven by user needs, iterative development, frequent delivery of releases, continuous refinement, and flexible adaptation to changing requirements. This process helps guarantee that user feedback is consistently integrated as the project moves forward, rather than waiting until most of the development has been finished (at which point requirements or the system landscape may have significantly changed from when the project began). Development work will be conducted using test-driven development methodologies (TDD), which focus on iterative changes, user interaction with the system, and automated testing that allows errors to be discovered more quickly as the codebase and dependencies grow in complexity.

The Project Director will oversee the hiring process for the developer, manage the overall project timeline and budget, and ensure that all necessary funding and technology resources are available to the project team. This project will use two Product Managers (one each for the administrative and interactive gems) who will work closely together and will be responsible for coordinating requirements gathering, user stories creation, managing development activities, user testing, acceptance testing, and all other activities needed so that deliverables meet the project goals. The Developer (to be hired) will be responsible for creating the software, and, as they become more

familiar with the Hydra community, will become involved with outreach, user testing, and representing the project at appropriate conferences. The Community Manager will facilitate meetings of the Advisory Committee, represent the project at meetings of the Hydra Newspapers Interest Group, and be charged with leading all outreach activities and coordinating feedback from stakeholders testing the software. (Please see the List of Key Project Staff and Consultants for more information.)

BPL and Utah together have decades of experience in developing software tools that help curate, manage, ingest, and disseminate digital objects, from developing in-house management tools like SIMP (Neatrour, et al., 2014) to actively contributing code to community-supported projects within the Hydra community. For this project, staff from both institutions will work as a single project team, using a variety of online communication and project management tools to coordinate efforts. Project progress will be tracked through several measures, including a weekly meeting (conducted via conference call or video chat) with all project members as well as an online tracking system (such as JIRA or GitHub Issues) to organize user stories and report issues. Progress reports will be provided monthly to Advisory Committee, Hydra Newspapers Interest Group, and Hydra Partners listserv, and detailed updates will be sent quarterly to all Hydra and Fedora community channels.

Project Evaluation: The success of these efforts will be evaluated from two interrelated perspectives: the success of the software products in meeting the stated requirements, and the broader success of the project itself in increasing the standardization, interoperability, availability, and dissemination of digitized newspaper content management. While the latter will be more difficult to quantify, there are a number of metrics to address the viability of the software itself. The deliverables will be frequently tested by stakeholders with each release, and the feedback from these test implementations will be used to gauge how well the requirements have been fulfilled. Usability testing with public users of the system will also be conducted to evaluate the software. The number of institutions that integrate the deliverables (the data model and/or the software) into their repository ecosystem will serve as a useful metric of both product and project success. Other metrics of evaluating the project's impact will include the number of metadata records for newspaper content submitted to digital library aggregators such as DPLA via the software, the amount of newspaper content migrated from proprietary software systems into open-source Hydra- and Fedora-based systems, the increase in the amount of newspaper content that can be accessed using standard APIs such as the IIIF Presentation API, and the number of contributors to the codebase outside of the project team. (Please see the Statement of National Impact for more information.)

Audience & Community Participation: The primary audience for the project will be cultural heritage institutions seeking to ingest, manage, disseminate, and share digitized newspaper collections using an open-source software solution built on a community-driven, extensible data model that supports discovery and access at a variety of levels (including volumes, issues, pages, and articles) and facilitates sharing of metadata and image content via standard APIs. The project plan is designed to create a self-sustaining collaboration between like-minded institutions -- leveraging the rich diversity of experience and use cases that exist within the Hydra and Fedora communities -- and will solicit participation at every stage of the process, from design to testing to code contribution.

Interested institutions may participate at a variety of different levels. An Advisory Committee representing key stakeholders will be convened quarterly to provide guidance on project planning, vision, and sustainability, as well as provide feedback on products of the design, development, and integration phases. Another point of community interaction with be the monthly meetings of the Hydra Newspapers Interest Group; plans for an initial investigation of data modeling and requirements are already part of this group's activities, and will serve as another layer of review for project activities and deliverables. Lastly, all progress reports and other communications to digital library development listservs and Google Groups will include a call for participation and feedback.

Resources: Resources needed to undertake this project can be divided into the following categories:

- *Personnel*: Staff from BPL and Utah will assume the roles of Project Director, Product Manager, and Community Manager. IMLS funding is being requested to hire a developer, who will implement the results of the Design phase and create the software. Both partner institutions are committing significant staff time as an in-kind contribution, please see the Budget Justification for full details.

- *Technology Infrastructure*: A test environment consisting of virtual servers and dependency applications will be necessary during the development process to continuously deploy and test the code within the context of the Hydra technology stack. Utah will be responsible for creating and supporting this technology infrastructure.
- *Test corpus of materials*: A range of digitized newspaper content types and formats will be needed in order to test the functionality of the models and software against different use cases and support migration from other systems. BPL and Utah will contribute a significant amount of previously-digitized material for this purpose, and solicit examples of other content types from members of the Advisory Committee and Hydra Newspapers Interest Group.
- *Time*: The project will take place from July 1, 2017 through June 30, 2019, with 3 months devoted to the initial design phase, 20 months dedicated to development, and 6 months devoted to integration into stakeholder repository systems, community outreach, and promotion. (These phases will have some overlap -- please see the Schedule of Completion for details.) Given the specific scope of the project (which aims to support newspaper content within an existing repository infrastructure, rather than creating a stand-alone application from the ground up), and based on the extensive software development experience of the partner institutions, we firmly believe this will be an adequate period of time to accomplish the stated goals.
- *Financial*: The project budget will be managed by Utah, and participation by staff from BPL and UU will be cost-shared. IMLS funding is being requested to cover two areas of need:
    - *Staffing*: To cover salary and benefits for a software engineer for a 20-month period, to execute the work proposed.
    - *Travel*: To cover travel expenses (airfare, lodging, meals, conference registration) incurred needed to attend in-person meetings for all project staff, and attendance at relevant conferences and workshops for the developer.

Communication & Promotion: This project will seek to rely on community input in all phases, and the feedback received from stakeholders will be crucial to success and future sustainability. The communication plan will reflect our belief that broader involvement will be an essential factor in how widely the software is adopted, as well as its potential for increasing the availability, dissemination, and usage of digitized newspaper content. In addition to the quarterly meetings of the Advisory Committee and monthly meetings of the Hydra Newspapers Interest Group, project updates will be posted quarterly to relevant listservs and Google Groups such as Hydra-Tech[17] (>700 members), Blacklight Development[18] (>500 members), IIIF Discuss[19] (>600 members), and Fedora Community[20] (>1000 members), and links to the updates will be posted in Slack channels popular with library technologists. The updates will include lists of features added, results of user testing, links to video demos or screencasts of the software in action, links to code hosted on GitHub.com, and calls for feedback and participation. Content from these updates will additionally be posted to a publicly-accessible wiki that will be hosted as part of the project's GitHub.com repository. We will also seek to collaborate with other relevant projects such as DPLA and "Hydra in a Box" to take advantage of their communication networks, which will provide an even wider audience.

The project will be promoted at in-person and virtual conferences focusing on the intersection of libraries and technology. Project staff will submit proposals for invited talks or papers discussing the features and benefits of the data model and software, as well as workshops where demonstrations on how to integrate the software or migrate newspaper content from proprietary systems into Hydra- and Fedora-based systems can be provided. (Many library technology conferences include a day of workshops given by community members.) Key conferences that will be targeted include: Code4Lib (including national and regional meetings), Hydra Connect, Open Repositories, Digital Library Federation Forum, Coalition for Networked Information, DPLAfest, and LITA Forum, among others. Partner institutions will also promote the project on their respective social media outlets, using the richness of historical newspaper content to draw attention from libraries, news media, faculty, students, researchers, and the public. These communication and promotion efforts will collectively build widespread awareness of the project, bringing new institutions to the table, and establishing a solid base of stakeholders and implementations that will serve to help meet the project goals and sustain success beyond the life of the grant.

Sustainability: The sustainability of the project is inextricably bound to the extent to which the data model and software deliverables address the use cases and requirements of cultural heritage institutions seeking to manage and provide access to digitized newspapers. The need for this solution has been well articulated by the digital library community; if the product works, it will be adopted, and the more institutions that adopt it, the more vital it will be to the usability of each institution's digital collections. Therefore, the first priority for sustaining the project's long-term viability will be to successfully execute the project plan, forging meaningful partnerships and channels of communication with the stakeholder community each step of the way.

In terms of maintaining the code, both BPL and Utah are prepared to commit staff resources to this task beyond the grant funding period to ensure that the deliverables remain compatible with the continually evolving Hydra technology stack. We will also work with Hydra-based projects that are poised to gain a large user base (e.g. "Hydra in a Box"), making sure that the deliverables can be integrated with these systems.

With greater adoption comes a wider pool of developers available to contribute code as new requirements arise, to the point where the software is no longer sustained by any one developer or institution alone. The Hydra and Fedora development landscape is rife with such success stories, featuring numerous applications and plugins that have become integral to the mission of wide array of libraries and are now maintained by a self-sustaining community of developers and technologists, with parent institutions making open-source software development a regular part of daily operations. By following this well-established model, we hope to duplicate their success.

3. Statement of National Impact

The successful completion of this project will have a significant impact on the availability and usage of digitized newspaper collections by greatly enhancing the ability of cultural heritage institutions to efficiently manage and provide more granular access to these essential historical materials. Hydra and Fedora represent the fastest growing open-source DAM solution, and by providing robust support for newspaper content within this framework, this project will help ensure that historic newspapers continue to play a vital role in the national digital landscape.

Tangible results and products produced by this project will include: (1) a shareable, system- and programming-language-agnostic RDF-based data model addressing structural and descriptive metadata features unique to digitized newspapers; (2) a set of modular, open-source plugins in the form of Ruby gems for ingesting, describing, discovering, displaying, and disseminating digitized newspaper content that can be easily integrated into existing Hydra-based repository systems or used to build a stand-alone newspaper-centric system; and (3) a community of practitioners -- including developers, librarians, content specialists, and managers -- dedicated to addressing challenges and collaborating on best practices associated with managing digitized newspapers.

These results will lead to greater integration of newspaper content with existing repository systems, resulting in better discoverability of newspaper content, since these materials will be combined with extant digitized materials (such as photographs, manuscripts, and monographs) as opposed to being confined to a separate "silo" application. Increased access by both human users and automated computational processes will be facilitated by the accessibility of newspaper content via modern APIs such as those defined by IIIF, as well as Linked Data-centric RDF data serializations that natively support linking and re-use. This project will also facilitate improved sharing of newspaper content between repositories and with aggregators such as DPLA, by providing a community-supported data model and a standardized protocol for making content harvestable.

The modular, plugin-based architecture used to create the software deliverables will address the acute need for institutions to have the ability to manage digitized newspaper content via a solution that is readily adoptable within existing digital repository infrastructure, rather than being forced to implement and maintain yet another system limited to a single content type. The focus of this project on article-level modeling, indexing, and access will address the need for information systems that more efficiently connect the user directly to relevant historical content, especially in the context of large aggregated collections, saving the time of the user and exposing connections to

other relevant materials more directly. The crucial need for robust support for newspaper content within the increasingly popular Hydra and Fedora framework will also be satisfied by this project. This project will also address the need for digitized newspaper content to be accessible via protocols that allow sharing of resources beyond the confines of the host repository and further the goals of the Semantic Web.

Performance Goals: The results of this project will be measured against several of the stated performance goals from the "Content and Collections" IMLS agency-level objectives. By increasing the accessibility and discoverability of digitized newspaper content, this project will further the goal to "Broaden access and expand use of the Nation's content and collections." And by providing a path for robust management of newspapers within the open-source, collaboratively-maintained Hydra and Fedora framework, which provides content managers with a wide range of useful features and a solid platform for collection-building, this project will improve "management of the Nation's content and collections."

The success of the project in meeting these goals will be evaluated using the following metrics:

- Number of downloads of the software.
- Number of implementations of the software in Hydra systems.
- Number of newspaper issues and articles made available by adopting institutions.
- Number of newspaper issues and articles harvested from adopting institutions by DPLA.

(The targets involving DPLA harvesting are obviously subject to the activities and priorities of that organization, and are included here strictly as tentative projections. While DPLA does not currently harvest newspaper content, they do intend to aggregate these materials in the near future. Please see DPLA's letter of support for more information.)

The timeframe for measuring project success will begin one year after the release of 1.0 version of the software, which will represent a "minimum viable product" meeting the core requirements identified by the project team during the Design phase. Evaluation data will be collected by surveys and direct emails sent out to implementing institutions, as well as an analysis of DPLA's collections.

The targets for each metric will be, respectively:

- 10 implementations in full production in Hydra systems utilizing the software deliverables.
- 1,000 downloads of the software from RubyGems.org. (This goal is not as audacious as it seems: due to the nature of Ruby on Rails applications, each time the software is installed constitutes a unique download.)
- 250,000 issues and/or 1,000,000 articles made available in Hydra systems utilizing the software deliverables.
- Metadata records for 100,000 issues and/or 250,000 articles harvested by DPLA from Hydra systems utilizing the software deliverables. (This is smaller than the previous target due to the fact that not all institutions implementing the software are currently harvested by DPLA.)

Sustaining Benefits: As discussed in the Project Design, sustaining the benefits of this project will be dependent on the extent to which the data model and software deliverables satisfy the requirements of institutions seeking to manage and provide access to digitized newspapers. The long-term viability of the deliverables will ensure the sustainability of the project benefits, including increased usage of and access to vital historical information, both for human users and machine analysis. Through user-centered, iterative design and continual engagement with the stakeholder community, we will ensure that the deliverables meet the demonstrated needs and create the partnerships, community, and communication pathways necessary to sustain the project beyond the grant funding period. Above all, by delivering a product that works, within Hydra's flourishing application framework, maintained by a committed community of developers, supported by institutional commitments to open-source software solutions, we strongly believe that this project will be poised for long-term success, thereby increasing usage and improving management of our nation's content and collections.

University of Utah, J. Willard Marriott Library - Historical Newspapers in Hydra January 12, 2017

Schedule of Completion

Year 1: July 2017 - June 2018

	Jul '17	Aug	Sep	Oct	Nov	Dec	Jan 18	Feb	Mar	Apr	May	Jun
Advisory group meeting												
Developer search												
Requirements & needs assessment												
Architecture & modeling (Design)												
User stories creation (Design)												
Hiring developer												
Kickoff meeting (project team)												
Developer training/onboarding												
Advisory group meeting												
Development environment setup												
Development (Phase 1: Admin)												
Advisory group meeting												
Testing & revision (Phase 1: Admin)												
Advisory group meeting												
User stories & requirements review												
Development (Phase 2: Interaction)												

University of Utah, J. Willard Marriott Library - Historical Newspapers in Hydra January 12, 2017

Year 2: July 2018 - June 2019

	Jul '18	Aug	Sep	Oct	Nov	Dec	Jan 19	Feb	Mar	Apr	May	Jun
Development (Phase 2: Interaction) con't												
Advisory group meeting												
Testing & revision (Phase 2: Interaction)												
Project team meeting (in-person)												
Presentations at conferences												
Advisory group meeting												
User stories & requirements review												
Development (Phase 3: Launchpad)												
Advisory group meeting												
Community outreach (Integration)												
Implementation by partners (Integration)												
Testing & revision (Phase 3: Launchpad)												
Advisory group meeting												

DIGITAL PRODUCT FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products can be challenging. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions
You must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

PART I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

All documentation and training materials will be released under the Creative Commons Attribution 4.0 (CC-BY) and all software created will be released under the Apache Software Foundation 2.0 License. The University of Utah and the Boston Public Library will retain copyright ownership.

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

We will assert no additional ownership rights beyond what is listed in A1.

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

The development of the software will not involve any privacy concerns or need to obtain permissions or rights, nor will it raise any cultural sensitivities.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

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

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

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

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

C. Metadata

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

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

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

D. Access and Use

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

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

Part III. Projects Developing Software

A. General Information

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

The software will consist of an administrative work flow management tools kit and a display toolkit that will provide the

necessary functionality for institutions to curate and publish historical newspapers on any Hydra platform.

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

open-oni, ChronAm, and Veridian. Our software will allow Hydra users the functionality to extend their platforms to support newspaper content. We will also open source our data model for newspaper model based on PCDM which can be used for other platforms.

B. Technical Information

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

The software will be written in Ruby utilizing the Rails framework. We choose to use Ruby and the Rails framework because they are the primary language and framework used by the Hydra community and they provide flexibility and scalability.

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

The developed software will be unique in that we will package the components as gems which will provide better interoperability with other Hydra heads. The model used by most systems is to create a unique and very specific system to handle content specific use cases, we however believe this model is out dated. Hence, we will create gems that will work with other Hydra heads which will reduce the need to create separate instances for each content type. This will allow organizations to extend functionality in their existing Hydra systems and will also provide more incentive to increase adoptions.

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

The software will require the following dependencies; Fedora Commons 4.0, Apache Solr, Blacklight, and the Hydra framework

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

The technical documentation will be created as the development progresses. All code will be properly documented and administrator and user documentation will be created as the software project nears completion. All documentation will be made available on GitHub.com

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

1. Digital Collections (DAM): https://collections.lib.utah.edu
2. SIMP Tool (Work flow Management): https://simp.lib.utah.edu
3. Digital Commonwealth: https://www.digitalcommonwealth.org/

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

All software developed for this project will be released under the Apache Software Foundation 2.0 license.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

All software will be available for download from GitHub.com and RubyGems.org.

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository: github.com

URL: https://github.com/projecthydra-labs

Part IV: Projects Creating Datasets

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

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

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

A.8 Identify where you will deposit the dataset(s):

Name of repository:

URL:

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?