Historical Newspapers in Hydra: Building a platform to restore access to cultural treasures

The University of Utah (UU)'s J. Willard Marriott Library and the Boston Public Library (BPL) are seeking $249,999 to fund a two-year software development project (July 1, 2017 - June 30, 2019) to design, develop, and distribute a set of Ruby gems that will provide the Hydra community and the greater public with a Hydra-based solution to manage digital historical newspaper content. As the lead organization, the Marriott Library will develop software in coordination with BPL, and in consultation with an advisory group of representatives from numerous cultural heritage institutions. The goal of the grant is to build software to enable organizations to ingest, manage, and publish digitized newspapers on the popular open-source Hydra platform.

Statement of National Need: Newspaper content poses unique challenges for management and dissemination, especially in terms of article-level encoding, modeling, and access. Locating relevant materials in digital collections is already often a difficult endeavor, a problem that will continue to grow as more newspaper content is digitized and projects such as the DPLA aggregate immense quantities of content from thousands of institutions into a single discovery environment[1]. We believe there is a better way to connect scholars to the materials they need, save their valuable time, and expose the hidden treasures in historical newspapers.

As organizations whose missions are to develop and foster learning environments and share the historical and cultural treasures of the world, we believe that the best way to further research and scholarship related to newspaper content is to build systems with article-level support using open-source, scalable, community-supported platforms. 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. We chose the Hydra/Fedora platform because the platform will provide the necessary technical underpinnings and the needed community support to ensure the long term sustainability of newspapers collections. On the technical front, the Hydra/Fedora platform is the ideal architecture because of its scalability, flexibility, and preservation capabilities. With continued financial and development investments from dozens of public and private entities, it is clear that the Hydra and Fedora platforms provide the necessary foundation to build such a system, and using these technologies as the platform will provide substantial long-term benefits to the cultural heritage community.

This grant addresses the "national digital platform" priority in that both Digital Commonwealth[2] (a digital repository for Massachusetts libraries, developed and hosted by the BPL, and Service Hub for DPLA) and UU intend to share content with DPLA, and so newspaper materials in these repositories will be aggregated on the national level. With modeling focused around article-level objects (classified into specific genres such as news items, classified ads, editorials, etc.), it would be possible to develop additional search and visualization features to more efficiently present large amounts of information and offer an improved end-user experience. The resulting code will be available for use by any Hydra-based institution, allowing more institutions to provide access to newspaper collections, and/or to consolidate siloed digital collections into a single platform.

Project Design: The goals of the project will be to develop open-source software that provides functionality for (1) ingesting digitized newspaper content (supporting issue-, page-, and article-level modeling) and (2) enabling discovery and access functionality, using the Hydra framework. The software will be modular by design and provide both administrative and front-end components. This code will be packaged as Ruby gems, which will be pluggable into Hydra-based web applications. This approach differs significantly from many projects in that these deliverables are shareable pieces of code rather than a separate, stand-alone application, so an implementor would not forced to run a limited-focus piece of technology exclusively for their digital newspaper content. The code, which will

[1] http://www.neh.gov/files/grants/university_of_nebraska_image_analysis_for_archival_discovery.pdf
[2] https://www.digitalcommonwealth.org/

be hosted on GitHub, will be freely 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 newspaper content.

We have established a Hydra Newspapers Interest Group[3] (consisting of representatives from institutions like Princeton, Yale, and Johns Hopkins) to advise the project, help gather specific requirements, test the code, and develop a community of users who would be interested in sustaining this project in the long run. We will collaborate with other newspaper-focused open-source community groups (including the IIIF Newspapers Interest Group and Open ONI) to guarantee that our efforts meet current best practices. We will also work with DPLA to ensure that newspaper content can be easily aggregated, and build on the research from their recent newspaper-focused Knight Foundation Grant[4] so that our software addresses the newspaper-related needs of many types of institutions.

BPL and UU will coordinate the software development process by utilizing industry- and Hydra-community-adopted agile and test-driven development methodologies. BPL and UU 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[5] to actively contributing code to community-supported projects within the Hydra community. The initial phase (4 months) will involve working with the advisory groups to determine requirements, and hiring a developer (to be directly managed by UU). The following phase (14 months) will include designing, prototyping, and developing the software. The last phase (6 months) will involve testing, refactoring, documentation, releasing the code as official Ruby gems, and implementing the software in production digital repositories, as well as assessment, outreach and promotion.

National Impact: This project will have a substantial impact on the usage of digitized newspaper content. First, this project would help create a standard way to handle newspapers in the increasingly-popular Hydra framework. Second, this project would allow partners to establish a set of standards and protocols for disseminating and sharing newspaper content in large-scale aggregated contexts such as DPLA, and would serve as a model across the DPLA network. Third, it would open up access to newspaper content at a much more granular level for scholars, connecting them to relevant information more effectively. And fourth, it will open pathways for further research into automating the conversion of page-level newspaper content into article-level content. Since BPL and UU are active in the Hydra community and well-established DPLA content hubs, we strongly believe that the models and services established through this project will become best practices for managing and disseminating digitized newspaper content within the Hydra framework.

Organizational Information: The Utah Digital Newspapers (UDN) program[6] has been operating since 2002 and is a recognized leader in newspaper digitization. As the regional hub for newspapers, we have partnered 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 also currently involved in a massive regional newspaper digitization effort, with the goal of making these materials available via Digital Commonwealth, which has over 190 member institutions, representing a rich variety of unique local newspaper content.

Budget: Co-PIs Harish Maringanti (UU) and Steven Carl Anderson (BPL) will be overseeing the project. We estimate the total budget for the two-year project will be $249,999. Direct costs include: $131,300 to cover 20-month salary for a developer; $52,520 for fringe benefits; $14,592 for travel expenses for the project partners to meet and disseminate findings at various conferences. Indirect costs include $51,587 to cover F&A. We also estimate a minimum contribution of $130,000 in combined in-kind contributions from UU & BPL in terms of staff time devoted to managing the project.

[3] https://wiki.duraspace.org/display/hydra/Hydra+Newspapers+Interest+Group
[4] http://dp.la/info/2015/11/09/dpla-announces-knight-grant-to-research-the-potential-integration-of-newspaper-content/
[5] http://www.dlib.org/dlib/july14/neatrour/07neatrour.html
[6] http://digitalnewspapers.org