Abstract

As part of the IMLS national digital platform funding priority, the Digital Public Library of America (DPLA), together with its partners Stanford University and DuraSpace, seeks to foster a new network of collaborative institutions and open access for the twenty-first century, one that will enable management, discovery, interoperability, and the reuse of digital resources by millions of people from this country and around the world.

Over the proposed 30-month time frame, the partners will engage with libraries, archives, and museums nationwide, and the global open source digital repository community, including Hydra, Fedora and DSpace. They will collaboratively extend the existing Hydra project codebase to build, bundle, and promote a feature-complete, robust, flexible digital repository that is easy to install, configure, and maintain—in short, a next-generation digital repository that will work for institutions large and small, and that will incorporate the capabilities and affordances to support networked resources and services in a shared, sustainable, nationwide platform.

The target audience for the initiative are institutions such as DPLA's "hubs," or those that wish to do similar work (and possibly become hubs) in managing, preserving and providing access to digital content. Many of the repository systems in place in these institutions today were designed, built, and deployed for a previous generation of technology, and do not take advantage of current web environments, user expectations, or curation needs for rich digital objects. These institutions are actively searching for new platforms that are easy to use, easy to integrate, and offer best of breed technologies and methods that match new and emerging requirements for managing and providing access to digital information resources.

This proposed initiative will satisfy a clear need by extending the extraordinary promise—and community—coalesced around the Hydra Project, an open source digital repository solution. Activities will be to radically advance and accelerate development on Hydra, to meet the demonstrated needs of DPLA hubs and similar institutions. Following a rigorous user-centered design approach and Hydra's proven agile software development methodology, the project will improve Hydra's capabilities to support advanced digital curation and management workflows, and enable publication and object syndication, including metadata aggregation with DPLA for national access. Development will be done in concert with both the larger Hydra community and the growing network of DPLA hubs, bringing the advantages of both broad-based, participatory development and long-term sustainability through shared investment.

In terms of project outcomes, the effort will equip libraries, archives, museums, and cultural heritage institutions with evolved digital repository capabilities; enable the efficient and scalable publication, aggregation, discovery, and reuse of digital content via DPLA; help form a global network of interoperable, digital content; and forge a robust community of common practice among institutions charged with stewarding and serving their digital assets.

Fostering a New National Library Network through a Community-Based, Connected Repository System

A Proposal from the Digital Public Library of America, Stanford University, and DuraSpace

I. Statement of Need

The management and publication of digital collections for libraries, cultural heritage, and memory institutions are poised for a generational change. Current digital repository systems were conceived with the main objective of putting collections online, during an earlier phase of the web in which unifying collections at a national scale was not feasible nor a priority, integration with other web-based services novel, and devices such as tablets and mobile devices considerably rarer. Since the advent of early-stage systems, digital content, the curation workflows applied to it, and mechanisms for publishing content to the web have all become far more sophisticated. The web has been able to demonstrate its capacity for decentralized discovery, sharing, and reuse of resources across the network. The right set of tools for managing, publishing, and sharing digital collections, marshaled by the right collaborators, has the potential to make the dream of a national digital platform for cultural heritage institutions a reality.

The Digital Public Library of America's hubs (institutions that host the material that comprises the aggregate national collection), as well as much of the library community in general, lament having to shoehorn the needed functionality into older systems to make this happen. Legacy systems are unable to take advantage of contemporary web affordances to describe, transform, preserve, and serve digital objects to audiences. For instance, they are often not based on an architecture that can natively support linked data, which is rapidly becoming critical for the cultural heritage sector to manage relationships between digital collections and their real world context. With almost two years of experience harvesting over eight million items from 24 hubs, DPLA has also realized that these systems lack rigorous and scalable ways to export their metadata into aggregation systems. This holds equally true for other external applications with which digital collections systems must interoperate, including discovery environments, metadata enrichment systems, content management systems, and crowdsourcing platforms. Legacy systems are also missing reliable and scalable mechanisms to import or synchronize improvements, such as geocoded place names, back from entities such as DPLA.

While some institutions have chosen to undertake the development work necessary to address these gaps, many have discovered aging software is also increasingly hard to maintain given older code bases and architectural assumptions, and painfully difficult to extend, integrate, and build new services upon. DPLA's hubs and similar institutions with major repository needs are using aging tools for the evolving needs of stewarding their digital assets, and for the entirely new job of making digital cultural heritage materials first-class citizens of today's web. Overall, these gaps make management of digital collections and their metadata both more complicated and more expensive for organizations like DPLA, as well as for all organizations seeking to share and add value to their content.

In a recent survey of DPLA hubs, nearly half of the respondents noted they or their partners are considering implementing a new digital collections management system in the near future. Several hubs also identified the importance of providing tools that facilitate easy integration of metadata enriched by DPLA back into their systems, both at the hub and partner level. Current systems in use by DPLA hubs and their partners for digital collections management include 47 instances of CONTENTdm, 14 instances of home-grown or locally developed software, and a smaller number of Islandora, DSpace, Omeka, Luna Insight, and bepress Digital Commons instances. Numerous hubs highlighted that they or their partners were actively seeking to move away from CONTENTdm, but as yet there were no "obvious choices" in terms of replacement systems. One hub noted that their possible move away from CONTENTdm was complicated by their reluctance to invest development time and resources to install, configure, and migrate to an immature open source solution. In short, there is a dissatisfaction with the current state of digital collection storage and display, and a great willingness, albeit tempered by the hurdles of any locally created solution, to move to a new environment.

Many members of the DSpace community—whether or not they currently contribute content to DPLA—find themselves in a similar position. With over 1,700 known installations worldwide, DSpace has had tremendous success as a repository service for institutions—especially those with limited technical means. The DSpace architecture and codebase, however, is not well structured for the leap to modern web standards and curation methods. An increasing number of DSpace institutions are thus looking to other systems to accommodate the growing range of their digital content, and the increasing sophistication of their stewardship and publication needs.

Fortunately, a rapidly growing set of libraries and developers are converging on a shared solution to address these new needs. The Hydra project[1], both a suite of open source software and a community dedicated to furthering its development, provides a flexible and robust framework for managing, preserving, and providing access to digital assets. The project motto, "One body, many heads," speaks to the flexibility provided by Hydra's modern, modular architecture, and the power of combining a robust repository backend (the "body") with flexible, tailored, user interfaces ("heads"). Co-designed and developed in concert with Fedora 4.0[2], the extensible, durable, and widely used repository software, the Hydra/Fedora stack is arguably the most technologically advanced and most promising solution for digital asset and metadata management available to the cultural heritage community. Since its founding in 2008 with three initial partners (Stanford University, University of Virginia, and University of Hull), the Hydra partnership has grown eightfold, with 26 core partners, and scores of adopters nationally and internationally—including national libraries and laboratories (the Royal Library, Denmark; the National Library of Ireland; the Digital Repository of Ireland; Los Alamos National Laboratories). The second worldwide Hydra Connect meeting[3] in Cleveland in October 2014 had 170 individuals from 47 institutions in attendance. As such, Hydra is also the centerpiece of a thriving and rapidly expanding open source community. This community of adopters includes participation by DPLA hubs and their content partners, including the University of Virginia, Digital Commonwealth[4] (the DPLA service hub for Massachusetts), Boston Public Library, the Lowcountry Digital Library[5], and WGBH.

While Hydra offers a robust software framework and a lively community to sustain and advance it, it currently falls short as an easy-to-implement solution. The grassroots and organic nature of the distributed community of Hydra partners and adopters has led to the development of impressive, but disparate, implementations to address local needs. While these are rich and functional, components of these implementations are not always easily reusable by other Hydra adopters. A frequent lament of

[1] http://projecthydra.org/

[2] http://fedorarepository.org/

[3] https://wiki.duraspace.org/display/hydra/Hydra+Connect+2+Program

[4] https://www.digitalcommonwealth.org/

[5] http://lcdl.library.cofc.edu/

smaller institutions without in-house IT expertise is that Hydra is a beautiful framework, but beyond their reach until there is an out-of-the-box application they can install without having to do extensive local development.

Hydra implementations are also missing key pieces that allow aggregators such as DPLA and its hubs to easily harvest, reuse, and add value to digital collections and their metadata in a scalable, modern manner. Conventions and standards, such as Schema.org[6] and Schema Bib Extend[7], ResourceSync[8], and the International Image Interoperability Framework[9] have emerged, and can support many core interoperability needs for digital collections. Finally, despite the introduction of a small but growing number of "solution bundles" within the Hydra community, such as Sufia[10] (a self-deposit institutional repository) and Avalon[11] (a management and access system for large audio and video collections), the Hydra partnership faces an ongoing issue in that there are no service providers available to offer hosted service options for Hydra-based repositories. Together, these gaps hinder broad adoption of Hydra, despite the merits of the framework's flexibility and technical robustness, and the strength of the community that develops and maintains the software ecosystem.

II. National Impact and Intended Results

Clearly, cultural heritage institutions are actively searching for new solutions that are easy to use, easy to integrate, and offer best-of-breed technologies and methods matching current web environments and evolving curation workflows. Widespread adoption of new and emerging standards across the DPLA network by current and potential hubs, and other content partners, will greatly reduce the technical and economic barriers to making their cultural heritage resources easily repurposable, aiding aggregation and discovery both within and outside the context of DPLA. Availability of a robust, turnkey open source solution for digital collections management will make it significantly easier for organizations to not only share their content with aggregators like DPLA, but will allow digital collections to be not just *on* the web, but *of* the web. In other words, improving and modernizing core infrastructure for digital collections will allow the cultural heritage sector to build on advances of technology not only in libraries, archives, and museums, but with the web at large.

We propose a tripartite partnership to address these increasingly pressing needs, and intend to produce a turnkey Hydra-based solution for the management of digital collections that can be widely and easily adopted by institutions nationwide. The partners will also produce a "cloud-ready" version of the turnkey application and offer a hosted service both to meet the needs of institutions without local IT capacity and to help sustain the effort over the long term. This partnership will feature major contributions from DuraSpace[12], the stewardship organization for Fedora, DSpace, and VIVO, and a non-profit with deep technical expertise in providing hosted services and managing open source projects; Stanford University, a founder and leading institution in the Hydra community, and home of one of the Digital Preservation Network (DPN) nodes; and DPLA itself, emerging as an expert in metadata

[6] http://schema.org/
[7] http://www.w3.org/community/schemabibex/
[8] http://www.openarchives.org/rs/toc
[9] http://iiif.io/
[10] https://github.com/projecthydra/sufia
[11] http://www.avalonmediasystem.org/
[12] http://duraspace.org/

aggregation, transformation, and enhancement, and on its way to becoming a national digital library of the future with public access for all.

The impact of this project will be substantial and widespread. First, it will allow institutions with aging installations of traditional software to have access to a best-of-breed replacement, one that is not only free and open source but also supported by a diverse, active community. Second, this project will provide an easier technological pathway for institutions to join DPLA; the turnkey solution will be recommended to new service hubs, as well as established ones—the state and regional digital libraries that power DPLA's network. Thirdly, these proposed activities advance the priorities in Hydra's strategic plan,[13] and will accelerate, bolster, and complement development of the open source community. Specifically, Hydra's strategies call for the development of solution bundles; development of turnkey applications and hosted solutions; growing a Hydra vendor ecosystem; ensuring the technical framework allows code sharing; and expanding the community of users and adopters. Beyond Hydra, because this project brings together a number of promising next-generation software initiatives, it can serve as a magnet for additional development around advanced services for digital collections and metadata management.

By providing hosting options, this project will enable a simple way for institutions without significant technical staffing or infrastructure to benefit from Hydra and to participate in DPLA. As we note below in the dissemination section, DuraSpace will host a version of the turnkey Hydra for its many members or other institutions that wish to move their digital collections to a hosted environment. Similarly, DPLA can work with its constituent hubs to help existing hubs move to a new codebase, and to provide hosting services for those who need the extra help. A hosted model also benefits the Hydra project not only through its expansion of the framework's user base, but through the potential to broaden the types of interests and capacity represented within the Hydra community.

III. Project Design and Evaluation Plan

The proposed project has four goals:

- produce a polished, feature-complete, easy-to-install, turnkey, Hydra-based application for next-generation digital asset management;
- improve and generalize DPLA's metadata ingestion system into an "aggregator system in a box," lowering the bar for DPLA contribution for users of the turnkey system and other Hydra applications;
- connect these key infrastructural pieces with DPLA hubs, current Hydra partners, and prospective Hydra adopters—creating a vibrant, participatory community of adopters and contributors; and
- work toward a hosted service, offering a cloud-based version of the application for use across multiple domains.

These goals strongly support the fundamental objectives of developing a national, digital platform. It will equip libraries, archives, museums and cultural heritage institutions with advanced digital repository capabilities; enable much more effective publication, aggregation, discovery, and reuse of digital content via DPLA; help form a global network of interoperable, digital content; and forge a robust community of common practice among institutions charged with stewarding and serving their digital assets.

[13] https://wiki.duraspace.org/pages/viewpage.action?pageId=58360054

To achieve these goals, project partners will form a dedicated, virtual team over the proposed 30-month project term. The cross-organizational team will use Hydra community's proven agile methodology, with test-driven development, transparent design and development practices, (relatively short) sprints, continuous releases, and ongoing documentation. All development will be done in an open repository in Github, and pull requests will be subject to peer review, as part of standard Hydra community development process. All code will be released as open source under an Apache 2.0 license (the Hydra community's norm), with contributor license agreements to the Hydra Project.

The project team will actively and continuously seek to incorporate design, code, and feedback from Hydra community members and the DPLA Hubs network and their partners from outside the funded IMLS effort; conversely, the project will actively and continuously disseminate its results to help ensure awareness and adoption by Hydra users worldwide. Given Hydra's modular nature, the "release early and often" philosophy will begin accruing benefits with the project's first release. Hydra community members will be able to see code while it is under design and development, contribute input and code, and incorporate incremental advances into their own work on a continuous basis, rather than waiting until a final deployment.

Transparency, continuous engagement, and integration with the existing Hydra community are fundamental to the project's design. This approach will leverage expertise and contributions from the Hydra community at large, making the project's designs, specifications and code richer, more extensive and better-tested than if it were developed in isolation. An open development process well integrated with the community's current efforts and structure will provide continuous validation of the project's assumptions, code quality, and release compatibility. Finally, by incorporating these specific Hydra developments into the main branch of Hydra community work from the project's onset, it will ensure a widespread base of adopters and contributors, each with a vested stake in maintaining and enhancing the software at the end of the project (see the Sustainability section, below).

Unlike most other, current Hydra development efforts, which are typically organic or locally driven, this effort will be run as a directed project, with a unified and integrated project team spanning three different organizations and multiple locations. There will be defined roles for a project manager (managing across functional, technical, and outreach activities, and supporting project administration); a product manager (who will serve as the product owner for the sprints and determine the product's feature set and roadmap); and a technical manager (who will arbitrate direction for architecture and data models, and ensure alignment with the community's overall technical direction). All hands will follow an agreed project plan, product roadmap, and technical architecture as determined by the respective leads. This is necessary in order to achieve the more ambitious goals and scope that a feature-complete, easy-to-install, cloud-ready, DPLA-friendly, Hydra solution demands.

The project team will follow a user-centered design approach, with development being done based on identified use cases grounded in user interviews and persona development, and supported by market and technical analysis. Customers/users will be both current (and would-be) Hydra users, as well as DPLA hubs. Led by DPLA, the project team will reach out to hubs throughout the project. Stanford will ensure alignment and incorporation of the effort with the Hydra community's interests and activities, including integration with Hydra's emergent interest and working groups, which provide an organizing structure[14] for activities ranging from digital preservation with Hydra to repository service management. DuraSpace will broker ongoing outreach and communication with those in the DSpace community interested in the features and community approach under development in this project. Initially this will all be for requirements-gathering, awareness, and expectation-setting; later in the project, it will be to validate designs and code, to further adoption, and to facilitate the bilateral exchange of rich, interoperable and reusable metadata and digital content with DPLA and other aggregation and interoperability sites. Section 4, below, includes more information on each partner's contributions, commitments to, and benefits from, the project.

Many of the ingredients for the project's success already are available; it is building on extensive, existing work. Over the course of several years, the members of the Hydra community have done repeated reviews of Hydra functionality relative to that of DSpace[15, 16]. Reasonable estimates are that 95% of DSpace functionality currently exists somewhere within the Hydra ecosystem, but not yet all in a single Hydra head (a single instance of a Hydra-based application) and none with the ease of installation and management that comes with DSpace. Pulling these features together into a single Hydra application that is also easy to use, while non-trivial, is also straightforward.

A promising starting point for developing a turnkey Hydra application is Sufia,[17] a component that is a focal point for much of the Hydra community's current development activities since it is the closest to providing a generic digital repository solution bundle within the Hydra ecosystem. Using Sufia as a starting point, the project partners will leverage the Hydra community process to gather requirements, specify and design, and then code features necessary to make our next-generation digital repository platform. Beyond supporting complex digital objects, known milestones on the roadmap to a feature complete Hydra repository include support for: sets/collections; permissions, access control, and administrative metadata; mediated deposit and workflows; management dashboards; bulk operations (especially important for data and metadata transformations and enrichment); very large files (such as video or scientific data); RDF/linked data; code modularity; ease of installation; and ease of maintenance. Note that this project has the rare opportunity to accelerate existing and proven work in these areas to build a national digital platform.

In addition, DPLA will do significant work under this project plan to further refine its toolset for metadata aggregation and harvesting, so that it can be married with the Hydra efforts. In a prior IMLS grant, DPLA's tech team was funded to move its Metadata Application Profile forward from version 3.1 to version 4.0, which included much richer metadata and linked data, and to begin reshaping the entire ingestion system into a more modular and reusable state. These efforts point toward ways to expand and generalize the DPLA ingestion workflow into an "aggregation system in a box," which could be set up as a cloud service and hosted for new service hubs and other aggregations across the country.

We plan to complete and release this "aggregation system in a box," integrating it with the Hydra development within this project, in the timeline of this present grant. This further development of the DPLA metadata ingestion system will allow for support of new and emerging standards not currently supported by DPLA's ingestion process, like ResourceSync. The intention is to extend the same modular

[14] https://wiki.duraspace.org/pages/viewpage.action?pageId=67241635
[15] LDCX Session Notes, 2013: https://lib.stanford.edu/node/12912
[16] Hydra Partner Meeting notes, 2013, https://wiki.duraspace.org/display/hydra/IR+Solution+Bundle+Requirements+Breakout
[17] https://github.com/projecthydra/sufia

infrastructure that DPLA is currently developing for its new ingestion system, to improve aggregation and content management, with similar capabilities to define ingestion profiles, mapping, transformation and enrichment processes and to generate quality assurance reports. Discussion with DPLA content and service hubs suggest a minimally restrictive open source license will allow for the best possible reuse and contributions of such a system. As such, this system will be released under the MIT License.[18]

Project partners will use several fundamental metrics to gauge the success of the multipronged effort: 1) the overall growth in Hydra community (as measured by the number of partners, known adopters, community email list members, licensed code contributors, code contributions made); 2) the usage rate of the new supported Hydra components (as measured by voluntary surveys and registration); 3) the number of DPLA Hubs that become new Hydra partners and adopters over the grant term; 4) the number, ease of transmission, and relative quality of metadata records submitted to DPLA via Hubs using the emergent platform; and 5) the number of inquiries and pilot users for a hosted Hydra service.

IV. Project Resources: Budget, Personnel, and Management

This proposed effort over 30 months will draw on the following roles (some full-time, some part-time) from the three partner institutions. This mixture of staffing will help season the team with dedicated contributors to do the "heavy lifting," especially in critical roles, while also providing coverage and complementary skill sets with a sizeable cast of contributing experts.

Personnel (Grant-funded and Externally Funded)

Project Directors ([redacted]). DPLA, as the lead institution on the grant, will take overall responsibility for the grant. Among the three organizations, the Directors will track finances, agree timeframes and deliverables, and assign local resources. They are ultimately accountable for success of individual components from each institution, as well as overall success of joint effort.

The **Project Manager** ([redacted]) will help create and manage the project plan, produce a work breakdown structure; coordinate across functional, technical, and delivery teams; track progress against timelines, and drive project goals to completion. He will also provide general project administration and support.

The **Product Manager** ([redacted]) will determine the application's feature set and roadmap based on user interviews, personas, landscape surveys, and knowledge of marketplace, and will take a lead role in coordinating across functional, technical, and delivery teams. She will produce and groom a feature backlog for the developers, and perform acceptance-testing of features; help produce documentation and marketing material; and help deliver training and outreach. She will work with early adopters to test the application, and work with community leads to expand the installed base.

The **Technical Manager** (to be hired) will serve as technical lead for development of Hydra components and architecture; ensure alignment of development with existing Hydra community code base, development practices, and directions; drive diffusion and adoption of the project's development into the wider Hydra community; and ensure incorporation of innovations and development into the project's development plan.

[18] http://opensource.org/licenses/MIT

User Experience (UX) Designers & Specialists ([redacted]) will conduct user-centered design for requirements gathering and to help determine feature needs; generate personas to guide development, and create UI specifications based best practices in human-computer interactions. They will also generally perform analysis, technical writing, support, training, outreach, and metrics tracking and reporting.

The **Data Modeler / Metadata Architects** ([redacted]) will define data models and patterns that support current and emerging needs for digital objects and metadata; ensure compatibility with DPLA metadata best practices and standards; incorporate linked data practices into Hydra model, and help ensure alignment of model needs with code and supporting systems (Fedora and Hydra).

Developers ([redacted], and one more to be hired) will perform software engineering to realize the project's feature backlog, functional and non-functional requirements. Principally working in Ruby on Rails, they will employ test-driven development and make use of and contribute to the existing Hydra code base to the greatest extent possible. Note that 1-2 developers will focus on enhancing and packaging Fedora 4 (based in Java) to optimize its compatibility with the project, and on the devOps features of Hydra, ensuring it is "cloud-ready" and can easily be deployed and run in a multi-tenant, virtual environment on commodity hosting services.

Community managers ([redacted] will serve as the primary contacts for outreach to communities such as DPLA hubs and their contributors, DSpace and Fedora users; gather feedback to inform development priorities; help deliver training and disseminate information about the project; and work with early adopters to test and to establish an install base for the application.

The **Documentation Specialist** ([redacted]) will undertake writing and production of documentation, training materials, and possible workshop content aimed at users of the systems developed within this project.

The **Business Development Manager** ([redacted] will be responsible for identifying and pursuing opportunities for cross-sector partnership, adoption and enhancements with both commercial and cultural heritage entities. S/he will work to identify potential vendors, service providers and contractors who can assist DPLA hubs and Hydra adopters with implementation services. S/he will also help explore tie-ins for online services that will further enrich the solution's utility.

Marketing & Communication Specialist ([redacted]) will help support broadcast communication and build a public profile for the project. They will also be responsible for two-way communication with the larger repository community, helping to gather requirements and disseminating news and updates about the development progress.

Management

The partners bring unique and complementary skills and strengths to the project. Stanford will use its existing leadership in the Hydra project to develop core Hydra components and ensure alignment with the greater community; it will resource product and technical management roles, as well as UX and development. Stanford benefits directly from a stronger Hydra community and a more complete

Hydra-based application for its use in the Stanford Digital Repository.[19] DPLA will focus on the development of infrastructure necessary to support metadata mapping and crosswalking, the harvesting and synchronization between hubs, partners, and DPLA, and infrastructure to support metadata enhancement and remediation. DPLA will also manage outreach to and analysis of hubs' needs. DPLA benefits directly by strengthening the national network of content contributors and increasing the quality of data and metadata available on the web. DuraSpace will use its expertise in building repositories, and doing so at scale, to construct the back-end systems for Hydra hosting, enabling the platform to run on cloud infrastructure. It benefits by helping build a more feature-complete Hydra, which serves as one of the principal front-ends to Fedora, and by offering the DSpace community a compelling and contemporary alternative to that platform. DuraSpace supports a community of over 2,000 repository installations all over the world, and has 140 member organizations that provide financial support and in-kind contributions to both the Fedora and DSpace platforms.

V. Communication Plan

The three partners have extraordinarily broad connections that will be leveraged to disseminate the results of the project widely. The Digital Public Library of America, Stanford, and DuraSpace are all looked to for technology leadership, and regularly participate in large national and international meetings. For instance, we will present the outcome of this project at the Coalition for Networked Information meetings, which reach hundreds of university librarians and CIOs; the Digital Library Federation, Open Repositories, and Code4Lib meetings, which together reach more than a thousand librarians and technologists who are most likely to implement a solution such as this; DPLAfest, the Digital Public Library of America's annual meeting, which has many members of the public library community in addition to university libraries; and other large-scale and niche-interest meetings.

Current marketing and outreach channels will be employed to leverage the national impact of project results. With both knowledge of and deep connections to the Hydra community and academic library communities DuraSpace, DPLA and Stanford will develop a marketing plan to build awareness and adoption of this advanced Hydra solution. Ongoing widespread distribution vehicles and tactics include: serial publications (newsletters, reports; 9,000+ DuraSpaceDigest monthly circulation; 3,000+ DPLA Newsletter circulation); Face-to-face event marketing (speakers at national and international conferences, exhibit tables, DPLAfest, Open Repositories); web seminars (free synchronous and recorded asynchronous access; thousands of attendees and views for both DuraSpace and DPLA); Blog posts, news items, and multiple social media channels (1,700+ followers on DuraSpace Twitter feed; 17,000+ followers on DPLA's Twitter feed).

For the version of Hydra that will be supplemented with the ingest and data management needs of the Digital Public Library of America and its hubs, DPLA will take the lead to push the resulting code out to its current and existing partners. Using its regular connections to the hubs via media ("Hupdates") and in-person visits, the DPLA Content Team will assist the Tech Team in getting information and deployment options out to its national network. The project will also make extensive use of the existing Hydra community and its communication channels as part of its dissemination events. By channeling

[19] http://library.stanford.edu/research/stanford-digital-repository

requests for input, code contributions, and validation of code through the Hydra project email lists (with hundreds of individual members), Interest and Working Groups, and face-to-face meetings, the project will have a ready-made audience and contributors with a natural interest and capacity to participate at a deep level. The dissemination will in turn lead to pathways for sustainability.

VI. Sustainability

During the past several years, over 75 organizations have come together to work collaboratively to build a variety of applications under the "Hydra" project. This "new" configuration of Hydra being proposed will naturally become one of the Hydra applications and fit under the current umbrella of Hydra projects. The community of members currently supporting the Hydra projects have committed to work together to sustain and advance the projects, ensure all code contributed to Hydra is open source under the Apache license, and use a standard framework to build a variety of Hydra applications. The Hydra community is excited about this proposal and new development, as it will help to advance all Hydra projects in providing a standard, easy to use configuration which is currently lacking across the community projects.

At the time of this grant writing, DuraSpace has recently become a "fiscal sponsor" of the current set of Hydra projects. In this relationship DuraSpace will provide a legal home and administrative support for the projects. Both the DuraSpace organization and Hydra members will be evaluating the expansion of this relationship, in the current year, such that Hydra becomes a fully recognized project under the DuraSpace organization, similar to the Fedora, DSpace, and VIVO open source projects. DuraSpace has many years of experience providing stewardship and sustaining both DSpace and Fedora projects, and more recently VIVO. Elements of sustainability include building a robust and engaged community of users and contributors; building a governance model to allow for different levels of participation and decision-making; providing a funding stream to provide any necessary dedicated resources to advance the project which are not provided through the current volunteer contributing organizations. Over the course of the next several years, DuraSpace will be working closely with the Hydra community to map out a longer-term strategy for ongoing sustainability and support with the help and guidance of DuraSpace. This proposed project will directly benefit from this work.

In addition, DPLA will work directly with its hubs to ensure organizations running those digital libraries can sustain their efforts. Since dozens of hubs will begin to use, or migrate to use, this central piece of infrastructure, and given the tightly connected nature of DPLA's national network, there will be strong pressures to sustain the aggregation- and metadata-enhanced version of this new turnkey Hydra implementation. DPLA's core technical staff of four is committed to ongoing maintenance and improvement of this constellation of technology well beyond the phase of the grant. Having a common infrastructure among organizations in the DPLA network is crucial since without it DPLA would return to the suboptimal state of affairs that a highly heterogeneous and aging set of repositories currently presents. In addition, DPLA has made sustainability of the central organization and its hubs a top priority in its recently released three-year strategic plan, which coincides with the term of this potential IMLS grant. A $600,000 grant from the Andrew W. Mellon Foundation will be instrumental in helping us achieve that sustainability, which will extend to our Hydra-related work.

Schedule of Completion

May 15, 2015 - November 15, 2017

Our effort will be divided into three distinct phases. In Phase 1 (months 1-6), we will conduct a landscape survey (ContentDM, DSpace, existing Hydra heads, DPLA hub profiles); perform initial UX interviews & persona generation; do preliminary UX design; formally express the project's high-level requirements and specifications; conduct intensive data modeling (specifying models for works, collections/sets, permissions & administrative structures, all in RDF-based models). Finally, we will dedicate significant effort to team formation: an important element given the distributed nature of the project, and the need to integrate with both the wider HydraSphere and DPLA hub requirements. This phase will also be used to hire new staff. This phase is critical to grounding the project's development in user-centered needs and developing the data models before coding starts—a necessary prerequisite. Outreach and communication during this phase will include preliminary announcements of the grant award, project goals, and framing of community conversations, at events such as the Digital Library Forum and Hydra Connect 3.

Phase 2 (months 7-18) will focus primarily on core application development for both the turnkey Hydra system and the DPLA metadata aggregation system. During this year, the primary focus will be refactoring and integrating existing code from across the Hydra community in different institutions' extant applications and components to develop a minimum viable product (MVP). By the end of this phase, we do not expect the product to be feature complete; however it will have the core features necessary to satisfy basic needs, and serve as a useful foundational product for the community. Note that the agile principles of frequent releases will apply here, and the project partners will be making continuous deployments of individual components for incorporation into the installations of current Hydra partners (and DPLA hubs, as their interest and capacity allows). Technical documentation will be developed in parallel with software development within this phase. During this phase, we will continue our outreach and community activities, through presentations and community-focused interactions with the DPLA network and Hydra adopters.

Phase 3 (months 19-30) will be dedicated to four goals: 1) adding functional enhancements to the applications; 2) bundling them for turnkey installation; 3) standing up a hosted service; 4) completing user documentation for the applications. During this phase, we will undertake aggressive outreach to DPLA hubs and early adopters in the Hydra community (both current partners and new users) to test the applications' ease of installation and maintenance.

SCHEDULE OF COMPLETION

OUTPUT	ACTIVITY	May-15	Jun-15	Jul-15	Aug-15	Sep-15	Oct-15	Nov-15	Dec-15	Jan-16	Feb-16	Mar-16	Apr-16	May-16	Jun-16	Jul-16	Aug-16	Sep-16	Oct-16	Nov-16	Dec-16	Jan-17	Feb-17	Mar-17	Apr-17	May-17	Jun-17	Jul-17	Aug-17	Sep-17	Oct-17	Nov-17
A,B,C,D	Project team development	■	■	■	■	■	■																									
A,C,D	Landscape survey	■	■	■																												
A,C	UX interviews		■	■	■																											
A,C,D	Hiring of new staff		■	■	■	■																										
B,C	Initial outreach to DPLA Hubs		■	■	■	■	■																									
A	Initial UX design		■	■	■	■	■																									
A,C	Persona development			■	■	■	■																									
A	Development of high-level requirements			■	■	■	■																									
A	Initial data model development				■	■	■																									
A,C	Initial outreach to Hydra community					■	■	■																								
A,B	Iterative software & technical documentation development (initial phase)							■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■						
A,B,C	Outreach to Hydra community (initial phase)							■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■						
A,B	Incremental software release schedule (initial phase)									■		■		■		■		■		■		■		■		■						
A,B,C,D	Outreach to DPLA hubs coordinated with major releases (initial phase)										■				■				■				■				■					
A,B,C	Development of user documentation																						■	■	■	■	■	■	■	■	■	■
C	Development of preliminary hosting platforms																										■	■	■	■	■	■
C, D	Outreach to DPLA hubs for turnkey application and hosting services																										■	■	■	■	■	■
A,B	Iterative software & technical documentation development (enhancement phase)																										■	■	■	■	■	■
A,B	Incremental software release schedule (enhancement phase)																											■		■		■
A,B,C	Application packaging for turnkey deployment																											■	■	■	■	■

OUTPUTS
A - Turnkey Hydra software development
B - Aggregator in a box development
C - Hosting service development
D - Community development

OMB No. 3137-0071, Exp. Date: 09/30/2015

DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction:
IMLS is committed to expanding public access to IMLS-funded research, data and other digital products: the assets you create with IMLS funding require careful stewardship to protect and enhance their value. They should be freely and readily available for use and re-use by libraries, archives, museums and the public. Applying these principles to the development of digital products is not straightforward; because technology is dynamic and because we do not want to inhibit innovation, IMLS does not want to prescribe set standards and best practices that would certainly become quickly outdated. Instead, IMLS defines the outcomes your projects should achieve in a series of questions; your answers are used by IMLS staff and by expert peer reviewers to evaluate your proposal; and they will play a critical role in determining whether your grant will be funded. Together, your answers will comprise the basis for a work plan for your project, as they will address all the major components of the development process.

Instructions:
If you propose to create any type of digital product as part of your proposal, you must complete this form. IMLS defines digital products very broadly. If you are developing anything through the use of information technology – e.g., digital collections, web resources, metadata, software, data– you should assume that you need to complete this form.

Please indicate which of the following digital products you will create or collect during your project. Check all that apply:

	Every proposal creating a digital product should complete	Part I
	If your project will create or collect	**Then you should complete**
□	Digital content	Part II
☑	New software tools or applications	Part III
□	A digital research dataset	Part IV

PART I.

A. Copyright and Intellectual Property Rights

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the copyright or intellectual property status of the content you intend to create? Will you assign a Creative Commons license to the content? If so, which license will it be? http://us.creativecommons.org/

All documentation and training materials produced as part of this project will released under the Creative Commons Attribution 4.0 (CC-BY) license. Software developed by this project will be under the copyright of the primary institution(s) that developed it and released under open source licenses (see Part III).

A.2 What ownership rights will your organization assert over the new digital content, and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users of the digital resources.

We will assert no additional ownership rights over new digital content other than that listed in A.1.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

This project has no concerns regarding cultural sensitivities or privacy.

Hydra-related projects are subject to the Hydra Project Intellectual Property Licensing and Ownership Guidelines and Contributor License Agreements (see Part III.C.1 below).

Part II: Projects Creating Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and the quantities of each type and format you will use.

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

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

B. Digital 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 grant period (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: Storage and publication after the end of the grant period may be an allowable cost.

C. Metadata

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, preservation). 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 and/or collected during your project and after the grant period.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content created during your project (e.g., an Advanced Programming Interface, contributions to the DPLA or other support to allow batch queries and retrieval of metadata).

D. Access and Use

D.1 Describe how you will make the digital content 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 URL(s) for any examples of previous digital collections or content your organization has created.

Part III. Projects Creating New Software Tools or Applications

A. General Information

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

The proposed activities include the development of (1) a polished, feature-complete, easy-to-install, turnkey, Hydra-based application for next-generation digital asset management, (2) "aggregation system in a box" based on DPLA's improved metadata ingestion system, and (3) the development of a hosting options for the turnkey Hydra application. The primary audiences include institutions with aging installations of traditional digital repository software to have access to a best-of-breed replacement; the network of DPLA Hubs; and institutions that need additional support in adopting a modern digital repository environment.

A.2 List other existing digital tools that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

While Hydra offers a robust software framework and a lively community to sustain and advance it, it currently falls short as an easy-to-implement solution. While these are rich and functional, components of these implementations are not always easily reusable by other Hydra adopters. A frequent lament of smaller institutions without in-house IT expertise is that Hydra is a beautiful framework, but beyond their reach until there is an out-of-the-box application they can install without having to do extensive local development. While DPLA has a new metadata aggregation system under development, it is not yet designed for general usage. The "aggregation system in a box" will use a simplified domain-specific language and user interface to create and manage ingestion profiles and mappings.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your new digital content.

The turnkey Hydra application will be built using the Hydra framework, Apache Solr, and Fedora 4.0, a best of breed digital repository platform. Hydra applications are written in Ruby, using Rails; additional integration components written in Java may be necessary. The aggregation system in a box will be an expansion and refinement of the DPLA metadata ingestion system, a Ruby on Rails application built on top of the Apache Marmotta triple store, Apache Solr, and the PostgreSQL RDBMS. Both systems share some common components, such as the ActiveTriples library for implementing RDF graph models in Ruby and the Blacklight discovery interface framework.

B.2 Describe how the intended software or system will extend or interoperate with other existing software applications or systems.

The systems will be designed to integrate with one another, and more broadly, with the metadata aggregation process used by DPLA to make freely available cultural heritage resources available to all.

Specific mechanisms will be identified during the requirements gathering process, but we are investigating the possibility of using emerging standards such as ResourceSync (http://www.openarchives.org/rs/toc) and the International Image Interoperability Framework (http://iiif.io).

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

Specific dependencies are listed in B.1. The turnkey system will include a “bundled” distribution of all necessary dependencies and software.

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

Technical documentation will be developed iteratively alongside the systems to be developed within this project. Developers will be directly responsible for maintaining documentation within the code; project staff will share broader technical documentation, either through Github or the Hydra wiki, hosted by DuraSpace. This documentation will be made freely available.

B.5 Provide URL(s) for examples of any previous software tools or systems your organization has created.

DPLA: https://github.com/dpla/KriKri ; https://github.com/dpla/platform
DuraSpace: https://github.com/duraspace/dfr
Stanford University: https://github.com/sul-dlss/spotlight ; https://github.com/sul-dlss/triannon

C. Access and Use

C.1 We expect applicants seeking federal funds for software or system development to develop and release these products as open source software. What ownership rights will your organization assert over the new software or system, and what conditions will you impose on the access and use of this product? Explain any terms of access and conditions of use, why these terms or conditions are justifiable, and how you will notify potential users of the software or system.

All code developed directly for in support of the turnkey Hydra-based system will be released as open source under an Apache Software Foundation License, version 2.0 (the Hydra community's norm), subject to the Hydra Project Intellectual Property Licensing and Ownership Guidelines and Contributor License Agreements (https://wiki.duraspace.org/x/eSHjAQ). The Guidelines and Contributor License Agreements address the requirements for permissions and rights. The DPLA "aggregation system in a box" will be released as open source under a MIT license, per request by a small number of DPLA hubs. As needed DPLA will either develop its own Guidelines and Contributor License Agreements or will adopt the Hydra Project's guidelines.

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

The turnkey Hydra-based system and the DPLA "aggregation system in a box" will both have their source code freely available for download from Github. We will also develop an "installer" or bundled version of the turnkey Hydra-based system for ease of deployment. This project also aims to provide a hosted version of the turnkey Hydra-based system. Specifics regarding the hosting agreements will be developed during the course of the project.

Part IV. Projects Creating Research Data

1. Summarize the intended purpose of the research, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

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

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

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.

5. What 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).

6. What documentation will you capture or create along with the dataset(s)? What standards or schema will you use? 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?

7. What is the plan for archiving, managing, and disseminating data after the completion of research activity?

8. Identify where you will be publicly depositing dataset(s):

Name of repository: ____

URL: ____

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

Original Preliminary Proposal

Fostering a New National Library Network
through a Community-Based, Connected Repository System

The coast-to-coast system of libraries that arose out of Carnegie's initiative over a century ago was essential to fostering democratic access to our culture—a radical idea with enormous social impact. As part of IMLS's national digital platform funding priority, the Digital Public Library of America (DPLA), Stanford University, and DuraSpace seek to foster a new network of collaborative institutions and open access for the twenty-first century, a network of hubs that will enable discovery and interoperability, and the reuse of digital resources by millions of people from this country and around the world. At the core of this transformative network are advanced repositories that not only empower local institutions with new asset management capabilities, but also interconnect their data and collections through a shared platform.

In the terminology of the DPLA, these critical nodes in the national network are called "hubs." They bring together and host the content and metadata from smaller collections across their state or region. Yet DPLA's hubs and similar institutions doing this key work are using aging tools for an entirely new job. Current digital repositories, as useful and good as they were when the main objective was to put collections online, were conceived in an earlier phase of the web in which unifying collections at a national scale was not feasible or a priority, integration with other web-based services was novel, and devices such as tablets and mobile devices were considerably rarer. DPLA's hubs, and much of the library community in general, lament having to repurpose older tools (such as ContentDM), which do not have natural, rigorous ways to export their data into a central discovery service, or re-ingest enhanced and corrected metadata from external sources like the DPLA. These systems are also not natively based on a linked-data architecture—critical to take advantage of relationships and related data streams—nor do they take advantage of contemporary web affordances to describe, transform, preserve, and serve digital objects to audiences. These tools are also increasingly hard to maintain given legacy code bases and architectures, and painfully difficult to extend, integrate, and build new services upon.

Institutions such as DPLA's hubs, or those that wish to become hubs or do similar work, are actively searching for new solutions that are easy to use, easy to integrate, and that offer best of breed technologies and methods that match current web environments and evolving curation workflows. Digital asset management for libraries, cultural heritage, and memory institutions is poised for a generational change, and the right set of tools marshaled by the right collaborators has the potential to make a huge impact—to provide much more streamlined and scalable network nodes for DPLA, and to make it easier for all institutions to host their digitized content in an up-to-date fashion and exchange what they have with others.

Fortunately, a rapidly growing set of libraries and developers are converging on an ideal solution: the Hydra project. The Hydra open source software suite provides a flexible and robust framework for managing, preserving, and providing access to digital assets. The project motto, "One body, many heads," speaks to the flexibility provided by Hydra's modern, modular architecture, and the power of combining a robust repository backend (the "body") with flexible, tailored, user interfaces ("heads"). Co-designed and developed in concert with Fedora 4, the extensible, durable, and widely used repository software, the Hydra/Fedora stack is arguably the most technologically advanced and most promising solution available to the cultural heritage community. With 25 core partners and scores of adopters nationally and internationally, and with 170 individuals from 47 institutions having just attended the second Worldwide Hydra Connect meeting in Cleveland in October 2014, Hydra is also the centerpiece of a thriving and rapidly expanding open source community.

Yet while Hydra offers a robust digital asset management framework and a vital community to sustain and advance it, and while it is poised to be fruitfully combined with DPLA standards and technology needed by hubs, it currently falls short of an easy-to-implement solution. Due to the grassroots and organic nature of its distributed community, individual partners have implemented impressive—but disparate—instances to address local needs. While these are rich and functional, they are not readily portable and do not have the key pieces that aggregators such as DPLA need. And as of yet, no providers have started offering hosted service options—hindering adoption further.

We propose a tripartite partnership to address these opportunities, and produce a turnkey, Hydra-based "hub-in-a-box" solution that can be widely and easily adopted by institutions nationwide. This partnership would feature major contributions from DuraSpace, the stewardship organization for Fedora 4 and a non-profit with deep technical expertise in providing hosted services and managing open source projects; Stanford University, a founder and leading institution in the Hydra community, and home of one of the Digital Preservation Network (DPN) nodes; and DPLA itself, emerging as an expert in metadata aggregation, transformation, and enhancement, and on its way to becoming a national digital library of the future with public access for all.

In the first year of our proposed, 30-month effort, the partners will form a dedicated, virtual team that will use the Hydra community's proven, agile methodology to streamline Hydra's data models and architecture; refactor and integrate existing code from across the community in different institutions' extant heads; and develop functionality that aligns with expressed needs from DPLA service hubs (current and prospective) and Hydra community groups. In the second year we will bundle a complete Hydra-based solution into a turnkey application that requires only local installation and configuration (i.e., no further development); build integrations for the solution that allow for easy harvesting and metadata mapping via DPLA-supported protocols and metadata models; and stand up a Hydra-based hosted solution for those institutions without the means or the need for a local instance. In the final six months of the project, DPLA will implement the new infrastructure with its existing hubs and work with new hubs so that they can join the network.

Given its considerable ambition and broad impact, this is a substantial project and we anticipate it will require $2 million in funding from IMLS, which we expect to match with an equal amount in cost share. We understand IMLS's potential funding constraints, however, and so like Hydra itself the project is modular. Even partial funding would help us make great strides; we would attempt to seek out additional funding from other interested parties to supplement IMLS funding, if necessary.

Stanford will use its existing leadership in the Hydra project to develop the core Hydra components. DPLA will focus on the connective tissue between hubs, mapping, and crosswalks to DPLA's metadata application profile, and infrastructure to support metadata enhancement and remediation. DuraSpace will use its expertise in building and serving repositories, and doing so at scale, to construct the back-end systems for Hydra hosting.

The impact of this project will be substantial and widespread, and felt immediately across the cultural heritage landscape. First, it will allow institutions that have aging installations of traditional software to have access to a best-of-breed replacement, one that is not only free and open source but also that is supported by a diverse, active community. Second, given the connections of the partnership, it will provide an easy pathway for scores of other institutions to join the Digital Public Library of America; indeed, DPLA can recommend this turnkey solution to new service hubs (as well as established ones)—the state and regional digital libraries that power DPLA's network. Finally, this project can serve as a magnet for advanced library services development, and as a coordinating body for that development, which is now happening in a piecemeal fashion.