The Engaged Repository: New Tools for Sustainable Scholarship

Significance

Institutional digital repositories have solved a major problem in the preservation of digital research materials and scholarship, by providing for the long-term storage and curation of important digital assets. However, this solution creates another challenge. Digital repositories are strongly oriented towards the needs of libraries and librarians, and present a limited set of basic discovery and access functions to researchers. The researchers who have the greatest interest in these assets—as their creators and ultimate users—are thus least likely to be able to access them in meaningful and creative ways. Repositories serve well as data management and preservation systems, but for scholars who are developing digital research materials for use in a digital project, data preservation methods need to intersect effectively with research and publication platforms. As noted in the Environmental Scan (see Appendix), researchers are increasingly comfortable with using digital repositories as a way of storing and circulating research publications such as preprints and offprints, especially in the context of open access, but the role of the repository as a facilitator of research project development is much less apparent.

From the perspective of libraries and other institutional centers aimed at supporting digital scholarship, there is a complementary problem. The institutional repository is an effective long-term preservation strategy precisely because it consolidates preservation activities and technologies. Digital research and publication platforms, on the other hand, tend to be developed as distinct stacks of technologies that may be different for each project. The goals of innovation and the availability of startup funding make it possible for individual digital humanities projects to develop their own research and publication environments, but once those projects enter the long-term survival phase of their life cycle and become the responsibility of the institution (whether through the library, IT organization, or a digital humanities center) those individualized environments become a liability. The effort required to maintain a growing portfolio of projects, as many digital humanities centers can attest, can scale up quickly and consume resources that could otherwise be spent on developing new work.

Addressing these challenges, the Northeastern University Library’s Digital Scholarship Group (DSG) now seeks funding to develop a repository-based platform for digital scholarship: the Digital Repository Service Shared Research Platform, or SHARP. SHARP builds on the preservation promise of centralized data curation while also integrating effective, sustainable tools for research and publication. At the foundation of this system is the Digital Repository Service (DRS), a Fedora/Hydra-based repository which serves the university’s needs for long-term data preservation. SHARP will enable researchers, individually or in teams, to create research projects and publications that involve collaborative discussion, annotation, and the creation of new scholarship, with all scholarly content stored and curated in the repository. The underlying platform—the technical components that communicate with the repository and provide for annotation, visualization, and so forth—can be maintained, upgraded, and expanded coherently in ways that benefit all the projects using it, rather than having DSG’s development effort dispersed among the separate needs of individual projects.

DSG has already done some important initial work in this area. In 2014-15 we are developing the DRS Project Toolkit, a set of easy-to-use publication tools through which researchers can
gather, contextualize, and publish selected items and collections from within the DRS. The Toolkit, which is implemented in WordPress with a set of custom plug-ins, makes it easy for researchers to build simple digital publications that include core features such as searching, browsing, visualization, and the creation of exhibits and narrative pages, calling data from the repository through an application programming interface (API). Over time, the Toolkit will be extended to include more features (like mapping and visualization tools) and these features will become available to all Toolkit projects, enabling them to grow over time.

The Toolkit is an important first step, but its emphasis is on small-scale projects whose primary goal is to present and contextualize collections of digital objects (such as documents, images, and video), and it does not address the needs of larger projects. In the next phase of this work, for which we are now seeking funding in this grant proposal, we will develop SHARP as an additional tool suite aimed at large-scale humanities research projects which require more complex workflows and more dynamic systems of scholarly communication. The figure below illustrates the role and situation of SHARP in the overall ecology.

The key features of this research platform are:

- An interface and workflow for ingesting materials to the repository, supporting contributions from scholars and from project members and collaborators at other institutions; the platform will support the creation of project-specific ingestion workflows, including support for consistent entry of metadata.
- Long-term storage and curation of digital assets in the DRS, including images, audio, video, TEI and other XML files, metadata, annotations, and scholarly commentary.
- Support for the creation of research communities associated with projects (including provision for user accounts with accompanying access privileges).
- Support for annotation of objects in the repository by members of the project’s research community.
- Support for mechanisms of scholarly communication between members of the project’s research community, including discussion forums, short-form scholarly publications, and shared research spaces (on the model of the “open lab notebook,” discussed in further detail below).
Support for explicit links from these venues to repository materials, and for the use of controlled vocabularies and ontologies to enable precise discovery of relevant strands of discussion.

What we are seeking to create here is both a sustainable platform and a sustainable digital scholarship, focusing not only on the primary research materials but also on the associated scholarly discourse and on the systems through which those materials are published, contextualized, and consumed. As an important consequence, all of these materials—primary sources, scholarly commentary, linked open data—become part of an interconnected web of information that can be studied and mined as it expands over time.

As noted in the Environmental Scan (see Appendix), others are also developing repository-based scholarly infrastructure, and the presence of other efforts strengthens our work rather than reducing its importance. Some of these projects (notably TAPAS and TextGrid) are focusing specifically on TEI-encoded data, and SHARP will take advantage of their progress in developing systems for effective display and analysis of TEI resources. The Canadian Writers Research Collaboratory (CWRC) is developing a system analogous to SHARP, aimed at supporting their community of scholars and editors in a specific set of editorial and collaborative activities. As scholarly platforms, SHARP and CWRC are currently on parallel paths that we would like to see converge, but at this stage they represent different and equally useful experiments at developing workflows, discursive mechanisms, and tool sets. During the development of SHARP we will maintain close contact with CWRC to discuss our different approaches and share results, and in our planning for subsequent larger-scale development we will involve them as a potential developmental partner. The Scalar authoring platform offers an intriguing set of authoring options and is selectively able to draw data from repository sources, but does not currently emphasize the creation of scholarly communities; Scalar will be an important source of tested ideas about digital authoring environments and the kinds of new scholarly discourse they make possible.

We have identified four initial pilot projects to use in developing and testing SHARP. These projects are important both for their own scholarly value and also because they represent a range of needs and functional requirements that this platform will seek to support. With the completion of the initial development of SHARP, we will issue a call for proposals for further projects that we will use to test the effectiveness of SHARP itself and the accompanying documentation and training materials. After that point, SHARP will become part of DSG’s regular infrastructure for supporting digital research projects. The four pilot projects for this first phase of development are:

1. The Early Caribbean Digital Archive (ECDA) typifies the kind of digital project for which this new platform is designed. Organized around a rich and extensive collection of primary source documents, with a strong research community, the ECDA seeks to develop a highly interactive digital scholars lab for collaborative research, and to engage both scholars and students in a shared, critical study of the textual, material, and cultural histories of the Caribbean. An explicit part of the project’s mission is to explore innovative digital technologies and newly emerging discursive platforms. The needs of the ECDA thus extend beyond the basic publication tools provided by the DRS Project

---

1 The Text Encoding Initiative (TEI) is an XML language for representing humanities research materials, and it poses specific challenges and offers specific opportunities for display and analysis; see [http://www.tei-c.org](http://www.tei-c.org) for more information.
The Engaged Repository: New Tools for Sustainable Scholarship

Toolkit: the project needs a framework within which the research materials can serve as the basis for a fuller ecology of scholarly communication.

2. **Diva 360: Letters and Manuscripts of Pauline Viardot** seeks to construct a holistic view of one of the nineteenth century’s most important musical figures, Pauline Viardot (1821-1910), based on the extensive and diverse documentary record of her life, which also represents an unparalleled resource for research on the musical culture of the nineteenth century. In her varied roles as singer, composer, collaborator, entrepreneur, costume designer, teacher, and memoirist, she left behind substantial numbers of letters, costume designs, and musical manuscripts. The research team for this project, led by Professor Hilary Poriss, wants to develop a collaborative research project in which these primary sources can be contextualized by additional materials including chronologies, annotations and commentary, and resources for studying her life and work. The project’s complexity puts it out of scope for the initial DRS Project Toolkit development process, but it is ideally suited as a pilot for SHARP.

3. **The Desegregation Collaborative Collection (DCC)** is a multi-archive digitization project from the Northeastern University Archives and Special Collections focused on archival material on the history of desegregation in Boston. The goal is to create a digital library of material that can be widely disseminated for both curricular and scholarly use. The project will expand existing use of Northeastern’s University Archives and Special Collections by creating new opportunities for scholars to interact with materials, developing new narrative techniques, and working with the local community to find opportunities for collaborative exploration of Boston’s history. In addition to supporting scholarship, the project has a strong community focus, making these important collections accessible to community members, students and neighbors and engaging them in the process of annotation and contextualization. The project combines several key collections that include a variety of document types including organizational documents, newspaper articles, personal correspondence, publicity materials, and public reports. This diversity not only provides an exceptionally rich set of research opportunities but also makes this a strong pilot project for SHARP.

4. **The TEI Archiving, Publishing, and Access Service (TAPAS)** is a repository-based publishing and research platform for TEI-encoded data, aimed at providing long-term storage and publication services for TEI projects that lack access to such infrastructure. Hosted at the Northeastern University Library, TAPAS uses a repository-based Fedora/Hydra architecture that parallels our approach with the DRS. As a result, tool development for TAPAS and the DRS is proceeding along complementary lines, and SHARP will take advantage of existing TAPAS code for the display of TEI documents. TAPAS will serve as an environment in which to test the portability of Hydra components for annotation and the exploitation of linked open data developed under this grant.

SHARP will address an important need at Northeastern University, one that is also indicative of a broader need in libraries for scalable, sustainable tools to support large-scale digital research projects. Digital projects and publications are increasingly popular modes of scholarly research in the humanities, supported through programs like the NEH’s Digital Humanities Startup and Implementation grants. As these projects increase in ambition and sophistication they go beyond what can be supported through simple content management systems like WordPress or Omeka, or simple XML-based publication systems like eXist or XTF. A recent Boston-area...
workshop titled “Beyond the PDF: Born-Digital Humanities”\(^2\) brought together experts from the domains of publishing, libraries and archives, repositories, linked open data, and scholarly infrastructure to discuss the future of digital publication systems, with particular emphasis on how these systems could support the creation and publication of secondary scholarship. One outcome of the event was a rough feature list for an innovative research and publication platform to support sustained argumentation within specific “data communities.” The target features included:

- a strongly curated “aquifer” of primary sources and data giving the researcher access to the foundational materials of scholarship, with provision for version management, long-term data preservation, and permanent addressability
- support via an API for analytical modules that treat these materials as data, including tools for text analysis, structural analysis, network analysis, mapping, and other appropriate research activities
- the ability to create authored narratives (both shorter- and longer-form scholarship) that use and link into the primary data, with the ability to embed outputs from the analytical mechanisms described above
- support for connections with the apparatus of traditional scholarship including peer review, assessment of impact, and citation

Implementing this ambitious platform in full will require contributions and buy-in from multiple institutions, and needs to proceed in carefully planned stages. We intend the current grant proposal and the development of SHARP as a set of initial steps in the process of realizing such a platform.

The first of these steps is the involvement of important stakeholders in the conceptualization of the overall architecture, to ensure that we are not only taking into account local needs and requirements at Northeastern, but also anticipating the requirements that will arise when adding connections to components and systems developed elsewhere. We will begin this project with a two-day meeting involving representatives from the key areas of this ecology: other institutions working on Fedora/Hydra-based repository systems for digital scholarship, such as Brown University and the University of Virginia; innovative digital scholarly publishing and communication platforms (such as CWRC, Scalar, Manifold Scholarship, and Annotation Studio); professional organizations concerned with the evolution of scholarly communication in the humanities (such as MLA and AHA); the repository technologies themselves, including Fedora, Hydra, and new efforts such as the IMLS-funded “Hydra-in-a-Box” initiative; and organizations developing linked open data resources, with particular emphasis on the representation of bibliographic and scholarly entities (for instance, Social Networks in Archival Context [SNAC]). This meeting will help us situate the immediate development process for the grant within a timely intellectual framework and ensure that we are aware of the current plans and development efforts in all of these relevant domains. The meeting will also start a conversation with these stakeholders which will continue in a second meeting in the final year of the grant, at which point we will be considering the next steps for the development of the platform, probably through some form of inter-institutional funding proposal.

---

2 This workshop, funded by the Andrew W. Mellon Foundation and co-sponsored by the University of Pittsburgh, Northeastern University, Brandeis University, and Tufts University, brought together 30 participants from Europe and North America, including representatives from standards bodies, funding agencies, institutional libraries, and tool developers. A white paper is planned but has not yet been published.
Those conversations will also inform and launch the research agenda for this project: centering on the challenges of supporting humanities scholarship in a way that treats discourse as data while also retaining its vibrancy and engagement as human communication. There are three central sets of questions at stake here. The first has to do with the changing shape of scholarly writing. What forms of scholarly research and writing can actually be catalysed through a collaborative, repository-based platform? If scholars have access to significant collections of primary data through tools that support annotation and authoring, what scholarly communication practices emerge without artificial incentives, as a normal outgrowth of the community’s engagement with the source material? What kinds of scholarship do research communities find most valuable to encourage: what moves the shared conversation in the most fruitful directions? What scholarly interventions receive the most sustained attention? What is the half-life of value for different kinds of interventions? Can we extend the visibility and usefulness of scholarly work by embedding it in a platform where it forms an accumulating, compounding context? We will address these questions explicitly through discussions with the pilot project teams and project partners, and we will also seek answers based on what we can observe of researchers’ behavior and usage patterns as we test the platform.

The second set of research questions concerns the possibility of formalizing certain informational components of scholarly discourse, to support better analysis and discovery. Can we improve the quality and level of engagement through mechanisms drawn from linked open data, both formal and informal? Formal mechanisms include the use of existing formal ontologies and vocabularies; informal versions of these include user-regulated bottom-up mechanisms analogous to the Twitter hashtag. In both cases, these systems provide a way to anchor precise meanings that can be analyzed and mined. Research that mines and analyzes Twitter hashtags already illustrates the promise of this approach for social science research, and our goal here is to explore how a scholarly communication platform can exploit such techniques to create tighter and more precise, recoverable connections between ideas, sources, and authors. Some of the relevant entities and concepts are already being managed publicly, or will be in the near future. For instance, the Social Networks and Archival Context project is developing the infrastructure for shared public linked data about the persons, families, and organizations named in archival collections, a potentially crucial resource for this platform. Resources like the Library of Congress authority records (names, subject headings, genre terms, and so forth) and the Getty Thesaurus of Geographic Names are now available as linked open data. Some research communities will also need to develop and support discipline-specific or even project-specific vocabularies, and SHARP will provide for the authoring and maintenance of these local resources, as well as linking to public data sources.

The third set of questions concerns the engagement of students in research communities, and the possible pedagogical uses of SHARP. What forms of annotation and authoring by students might constitute meaningful, durable contributions to these communities, and what kinds of curricular framing would best support them? Both the ECDA and DCC projects include significant curricular components, and our work plan includes specific provision for recruiting

---

3 The “hashtag” is a mechanism arising in social media and most commonly used by Twitter users to identify themes or topics using a specific “tag” prefixed with a hash mark (#). Tags are created by users themselves and are used to collocate their specific content with the broader conversation occurring under that hashtag.

4 For early discussions of these methods see for instance Dellschaft and Staab 2008, and Huang et al. 2010. Two research groups at Northeastern University, the Lazer Lab (http://www.lazerlab.net) directed by David Lazer, and the NULab (http://nulab.neostate.edu) have strong research programs in this area.
faculty who will use the platform in the context of an undergraduate or graduate course, with part of the user testing process focused on these usage contexts.

The core development work of SHARP will further this research agenda. As noted above, this grant proposal represents the first development phase in a longer strategic plan whose later phases involve inter-institutional collaboration. Our goals for this first phase are to lay down the basic architectural systems through which transactions between the digital repository and the user interface are conducted, and to develop an initial tool suite that demonstrates the core authoring and discursive functions. Research groups will use the SHARP framework to build collaborative digital environments that support a specific research community in working with a shared set of source materials. Specifically, projects using the SHARP platform will have the following features and supported activities:

- Projects can create and manage user accounts through which the identities and roles of community members are represented, both for purposes of authentication and access, and also for purposes of referencing, citation, and community interaction.
- Projects can create workflows for ingesting research materials and metadata into the repository.
- Projects can create and manage local ontologies and controlled vocabularies, in addition to using the public linked open data resources with which SHARP is connected. Local ontologies are stored in the repository.
- Project members can create annotations on repository objects. Annotations follow the Open Annotation Data Model ([http://www.openannotation.org/spec/core/](http://www.openannotation.org/spec/core/)) which represents annotations as structures of scholarly discourse that may be either quite simple (a gloss about a single word) or quite complex (a coordination of multiple resources, scholarly commentary, and metadata). All annotations are stored in the repository.
- Project members can create personal working spaces, framed as "open lab notebooks." The lab notebook is organized around a reading interface through which users can search, browse, organize, view and annotate repository materials. The user’s’ own annotations on each item are visible to them, and they may also choose to make their annotations visible to others and to view the shared annotations of other users, with filtering options to help manage copia and focus on annotations that are topically relevant. The notebook thus becomes both a reading and research space and a way of putting one’s own research activities into intersection and dialogue with others. The notebook will also include visualization and discovery features so that researchers can find those points of intersection and debate.
- Project members can use public and local ontologies in their annotations and other communication to make formal reference to entities such as authors, texts, repository objects, other project collaborators, genre terms, and keywords.
- Project members and general readers can explore and analyse both the repository materials themselves and the aggregated annotations using interface tools that take advantage of both metadata (on repository objects and on annotations) and ontology terms. For example, researchers viewing a specific repository document (such as an archival letter) can choose to see the hashtags representing the topical keywords, citations, and links that other researchers have associated with the item, represented through visualizations such as a timeline, a tag cloud, or a network diagram showing relations between terms or contributors.
The results of this grant will be substantive: a powerful tool set with at least four initial pilot projects demonstrating its features and one demonstrating its portability. At the same time, we recognize the important ways in which this work is only preliminary, and one very important outcome of this project will be the groundwork it lays for a subsequent phase of development. Important as the pilot projects will be, their research communities can only take their full shape once their collections of research materials achieve critical mass, which requires digitization resources and effort beyond the scope of this grant proposal. Similarly, for many research communities their core materials may be scattered among multiple repositories, and their maturation will depend on our developing technical mechanisms and institutional agreements that permit SHARP to provide integrated access to any public, trusted repository via a standard set of APIs. Such an architecture requires both significant investment and also diplomatic leverage from a major funding body. At the start of the final year of this project, we will hold a second meeting of the project partners in which we will review the actual implementation of SHARP and consider how the state of play at peer projects has changed. The goal of these concluding discussions will be to frame the next steps in realizing the larger shared vision for this platform, and to create partnerships that can support a second, multi-institutional development phase and an appropriate funding plan.

Background of Applicant

The Northeastern University Library

Northeastern University Library is at the hub of campus intellectual life. Resources include over 900,000 print volumes, 500,000 e-books, and 60,000 electronic journals. The Snell Library building welcomes 2 million visits a year on the Boston campus and the Library’s web site serves users around the world. In addition to a growing focus on networked information and extensive special collections that document social justice efforts in the Greater Boston area, the library has an ambitious vision to expand its digital initiatives by developing its digital repository, digitizing unique collections, constructing integrated collaborative spaces, and fostering the adoption of digital media and the creation of new knowledge.

Digital Scholarship Infrastructure

As a university, Northeastern has a recent but very strong institutional commitment to the digital humanities that has resulted in several key faculty hires (including Ryan Cordell, Benjamin Schmidt, Nicholas Beauchamp, and Julia Flanders) and the formation of the NULab for Texts, Maps and Networks to provide key services and research partnerships across a range of digital humanities and computational social science domains. With ongoing plans to expand both faculty and staff positions in these areas, Northeastern is emerging as a recognized leader in digital humanities. Interest in digital scholarship, and in innovative platforms for digital research and publications, is widespread at Northeastern. A recent call for DRS Toolkit pilot projects yielded proposals from Engineering, Nursing, History, Judaic Studies, Music, Archives and Special Collections, Political Science, Criminology and Criminal Justice, and the Library’s Digital Media Studios.

Reflecting increasing campus expertise in digital scholarship and digital humanities, the Digital Scholarship Group (DSG) was founded in January of 2014, under the direction of Julia Flanders, as an applied research group within the Northeastern University Library. The DSG supports and educates researchers at all levels on new techniques of representation, analysis, and
dissemination, and builds tools to help Northeastern’s researchers publish and preserve their research. The DSG now includes ten staff positions that include expertise in repository development and management, XML programming, TEI, GIS, web applications programming, data visualization, and scholarly publishing. Despite DSG’s recent origins, the organization supports a number of major and long-standing digital humanities projects including the TEI Archiving, Publishing, and Access Service (TAPAS), the Women Writers Project (WWP), and Digital Humanities Quarterly (DHQ), as well as more recent projects including the Early Caribbean Digital Archive, Viral Texts, and Our Marathon. Many of these projects have established histories and programs of research that contribute to the expertise and intellectual richness of Northeastern’s digital scholarly community.

Repositories at the Northeastern University Library

At the foundation of University and Library support for digital scholarship is the Digital Repository Service (DRS). The DRS is the library’s primary digital object storage system, and it currently stores more than 60,000 unique digital objects ranging from faculty and student work to archival materials and administrative files. Northeastern University Library has supported digital object storage and preservation in various capacities for a decade, starting with the adoption of Sympsoia, an Innovative Interfaces Inc. repository product, then IRis, a Digital Commons product. In 2010 the library decided to move away from vendor supplied repository systems and began to design a homegrown repository system using Fedora. The resulting digital repository, the DRS, represents a major part of the Library’s substantial investment in new scholarly technologies.

The library has enthusiastically committed to using Fedora and Hydra to develop the digital object storage architecture, the user interface, and the repository index for the DRS. Our implementation of Fedora, which is an open source repository architecture that supports a variety of essential digital object storage needs, has been more successful than the vendor supplied options because we have been able to build and design multiple repositories that suit the needs of content creators and content consumers. Library technology staff is currently investigating the requirements for moving to Fedora 4 and preparing to upgrade repository components for migration. Fedora 4 promises to be an important upgrade that will enhance existing repository practice and enable the Fedora repository to become a true linked data environment.

Adopting Hydra enabled us to move away from single-repository support to a multiple repository environment. By using Fedora and Hydra, Northeastern University Library has the ability to support a wide range of projects that require independent repositories. The underlying code for the DRS, branded as Cerberus, allows for easy uploading, search and browse features, faceted result refinement, multiple download options, a save-for-later feature, and a My DRS page for collections and deposited files. Although the code for Cerberus was initially created using Penn State’s ScholarSphere Hydra Head, Sufia, the Cerberus code that Northeastern University Library shares in GitHub today is its own repository system. Cerberus already supports discovery, uploading, downloading, communities and groups, and user-created sets and activity pages. SHARP will be able to rely heavily on the work that Northeastern University Library technology staff has already done to support the features described above.
History, Scope, Duration

This project involves several converging strands of work with distinct histories.

At the core of the project is the development of the DRS, which as noted above has been under way in its current Fedora/Hydra form since 2010. With the founding of the Digital Scholarship Group in 2014, the DRS took on an increasingly important role as a way not only of preserving scholarly research assets (such as data sets, offprints, and research data), but also of serving and curating the materials from which digital scholarly projects would be built—thus avoiding the tendency towards dispersion and project-based data management that makes long-term project support more difficult. With this philosophy in mind, the DSG began planning a platform for digital projects that would use the DRS as its primary data storage mechanism, with APIs and Hydra heads providing access to data which could then be manipulated, displayed, and queried through a variety of interface tools. We envisioned the development of this platform as a phased process, in which we would produce new tools and modules at intervals, recruiting pilot projects and developing external grant proposals as needed to support major development effort. We began the process with a set of simple tools (the DRS Project Toolkit) using WordPress and WordPress plugins to enable researchers to create projects on their own, using data stored in the DRS, without requiring individualized development effort from DSG staff. A set of pilot projects was recruited early in 2015 and will be completed by summer 2016. The pilot project process forms an on-ramp to digital scholarship: as new pilots are accepted, we will introduce them to concepts such as scope, metadata, and preservation, and once pilots are complete they will become full-fledged “DSG Projects”.

The second phase of development (for which we are seeking support in this proposal) focuses on adding components to support project-specific ingestion workflows (to enable external collaborators to contribute materials) and to support an initial set of scholarly communication mechanisms, including annotation, the use of controlled vocabularies, and the open lab notebook. Other planned developments (for which we are preparing separate proposals) will add advanced support for TEI and other forms of XML data, building on work already done by the TAPAS project (described in more detail below), and support for detailed mapping and geospatial analysis.

Another important strand of this history is the work that has already been done on two of our pilot projects for this grant proposal, the ECDA and TAPAS. TAPAS is a multi-institutional collaboration begun in 2010 and now hosted at Northeastern, with the mission of providing repository and publication services to small TEI projects and individual scholars who lack access to institutional support. Formally launched in October 2014, TAPAS is now halfway through a second development cycle under a grant from the National Endowment for the Humanities, working on building an XML-aware Fedora/Hydra-based repository back end to the service. TAPAS is an important part of the landscape for DSG because its repository-based approach (using the Hydra head mechanism to create reusable publishing and analysis components) closely parallels our approach to digital project support; while TAPAS and DSG serve somewhat different audiences, there are important functional commonalities and the two development efforts will be able to share significant amounts of code and expertise. For instance, the TEI reading interface now being developed for TAPAS will be directly reusable as part of the DRS Project Toolkit. Conversely, the tools for scholarly communication and annotation that we describe in this current proposal will have direct relevance and reusability within TAPAS.
The ECDA first emerged out of the formation of the Early Caribbean Society. The ECDA’s early development has been supported by three seed funding grants from Northeastern University, starting in 2012, under which the project has developed a TEI schema for transcription, captured an initial exploratory set of texts, and developed a set of scholarly exhibits by members of the project’s research community. The project has been shaped by the DSG’s deep XML and TEI expertise gained from TAPAS and the Women Writers Project, and the ECDA has also contributed to that expertise in important ways by exploring more intensive ways of involving user communities and scholarly agendas in the design and use of markup. The project has also built a working partnership with the Digital Library of the Caribbean (DLoC) and developed a community of researchers and contributors. With this groundwork in place, the ECDA has reached the point where it urgently needs a research and publication platform that can support the next phase of its development: building and publishing a core collection of transcribed texts, exploring modes of digital analysis, and cultivating the collaborative work of its research team.

Diva360 and the DCC are both newer projects, but we have already learned much from initial meetings with their representatives about how SHARP’s deliverables can meet the needs of their collaborative dimensions. The Diva360 team of music scholars is spread across several continents, a situation that necessitates an accessible, central location where researchers can access, discuss, and eventually curate materials. The DCC’s range of collaborators is focused more on a particular geographic region (Boston and the surrounding New England region), but the range of institutional participants similarly calls out for a shared digital space where content can be stored, refined, and examined. The fact that Diva360 and the DCC are in earlier stages of development means that they can help us test SHARP as a platform for incubating new projects and research communities, and can also bring fresh eyes to SHARP’s best practices and design questions.

The project we now propose is scoped as a three-year development effort, but the long-term horizons for this project extend beyond the immediate scope of this proposal, both within Northeastern and through inter-institutional partnerships. As noted above, the Digital Scholarship Group’s strategy for the support of digital scholarship is centered on the digital repository, and over time we plan to build successive components to disseminate, analyse, contextualize, and visualize repository data. Some of this work can be undertaken with DSG’s existing resources, as with the DRS Project Toolkit where existing tools can be adapted for our use. In cases of larger-scale work requiring additional staff effort or expertise—as in the case of the current proposal—we will seek external funding for the initial development, with the longer-term maintenance and documentation being covered as part of DSG’s regular work. Following the conclusion of this project, the next set of high-priority components for DSG includes support for advanced TEI/XML publication, geospatial data, and text analysis. More broadly, we will be exploring opportunities to extend the SHARP platform to enable it to work with multiple repositories and to embrace a more extensive set of scholarly research and publishing features, working through partnerships established under this grant.

**Methodology and Standards**

**Repository Methods and Standards: Hydra/Fedora, RDF, and Ontologies**

Digital repositories are now a common piece of institutional infrastructure and there exist a range of different repository systems. They represent a spectrum from very simple systems that
offer basic functionality out of the box to very complex and powerful systems that require significant configuration effort but provide much more sophisticated forms of data management, curation, and access. Fedora is in this latter category: it requires a very substantial institutional investment in expertise, design and strategic planning, and development effort, but it offers unparalleled power and has the potential to support a very wide range of needs and functions. For this reason it offers an ideal foundation for a long-term repository-based publication platform like SHARP. On its own, Fedora provides data storage and management, but not systems for interacting with that data; Hydra is an interdependent set of components which work with Fedora to create full-fledged repository systems that include interfaces for ingestion, curation, and exposure of repository data. Both Fedora and Hydra are open-source projects supported by non-profit organizations and long-term sustainability plans.

The recently released version 4 of Fedora offers a number of significant improvements, and Northeastern will be migrating the DRS to Fedora 4 during the next two years, with an initial investigation of the migration path in 2015, and with migration scheduled for completion by the end of 2016. Fedora 4 is designed to be a repository and a linked data platform, and it is an important component of the platform proposed here. However, while in the long term SHARP will benefit from the improvements available in Fedora 4 (as described below), there are no strict dependencies that would affect the development plans for SHARP. Fedora 4 follows the basic principles set by the Linked Data Platform ISO standard, which pragmatically standardizes Tim Berners Lee’s initial rules for creating networks of linked data: identify entities/objects using URIs, link using HTTP, share information about the identified entity/object, and link to other entities/objects.

The critical new component of Fedora 4 is its full support for the Resource Description Framework (RDF). RDF is a structural standard for storing linked data components, particularly metadata, ontologies, and other authorized vocabularies, like name authority values. RDF takes the specified metadata or ontology values and serializes each into three machine-readable components: subject, predicate, and object. The grouping of subject, predicate, and object (referred to as a “triple”), can be queried, identified, and referenced by other systems using query languages such as SPARQL. By taking advantage of Fedora 4’s RDF/Linked Data capabilities we will be able to make the relationships between object stored in the repository (like TEI documents and annotations), semantically meaningful, creating a network of related objects and entities.

Using predefined ontologies (vocabularies and term sets used to create a taxonomy of elements to organize and define entities in a system of related entities) allows concepts and entities to be atomized into controlled statements. These statements are more easily understood by the machines we use to interpret the information, and can be leveraged to create the semantically meaningful connections that fuel linked data and the Semantic Web. Where existing ontologies are available as linked open data, we will use them: for instance, during the specification phase we will review resources including the Getty and Library of Congress linked open data, the Social Networks and Archival Contexts, and other ontologies that may meet the basic expressive needs of our pilot projects. In addition to these publicly maintained resources, SHARP will also support the creation of local ontologies defined by specific research communities, including mechanisms for documenting them, adjudicating conflicting terminology, and mapping them onto public ontologies.
The platform we are developing as part of this project will be expressed as a Hydra “head”: a self-contained toolset that can interact with the Hydra API. The core functions of the Hydra head can be expressed as Ruby “gems” representing generalized functional units (for instance, to transform and display a TEI file using an XSLT stylesheet); within the Hydra development framework, both the head and the gems are designed to be adaptable and reusable by anyone using the Hydra platform. Unlike the somewhat vague hope of reusability that comes from posting tool code in public venues like GitHub (where the work of adaptation and troubleshooting rests with the reuser), in the Hydra community the creation of gems is well formalized and entails a level of design and documentation that makes the ease and likelihood of reuse much greater. The results of this work will thus be extensible to the increasing number of libraries working with Fedora/Hydra repositories as the basis for their support of digital scholarship — at the most recent Open Repositories conference, the number of Fedora installations was placed at over 300.

In addition, the “Hydra-in-a-Box” project, a collaborative effort between the Digital Public Library of America, Stanford University, and DuraSpace supported by a National Leadership Grant from the Institute of Museum and Library Sciences (IMLS), is an initiative designed to support the wider adoption of (and migration to) Fedora-based platforms by U.S. libraries, archives, and museums. The support expressed here by both the IMLS and the institutions involved in Hydra-in-a-Box echoes our own belief in the value of committing long-term to this repository model. We expect the particular customizations specified in our proposal will be of great interest to the even-larger community of Hydra users likely generated by the implementation of Hydra-in-a-Box in the coming years.

Open Source and Open Design Methods and Standards

Open standards are a crucial aspect of DSG’s strategy for long-term project support as well as our participation in the broader research community of digital humanities. DSG’s standard development practice uses applicable open formats and standards including XML, TEI, the W3C Open Annotation Data Model, Linked Open Data, W3C Web Ontology Standards, and RDF. Our current system of user authentication and identity management uses a combination of Shibboleth and Grouper, both well-established and actively maintained open source systems. Northeastern’s current implementation of Shibboleth has the potential for cross-institutional capacity, but we can also provide fall-back authentication as necessary using project-specific identity management.

DSG is also committed to open source development and we educate our projects in the relevant practices and tools, with particular attention to the use of open standards for data and metadata, open access web publication, and long-term preservation. All DSG projects maintain their code bases in GitHub or use an alternative public distribution mechanism. We will develop SHARP and its components within the DRS’ Fedora/Hydra framework and make all code we create available for modification and reuse via our GitHub repositories.

This public sharing of code is backed by a design philosophy that emphasizes reusability. Our repository-based approach to project support forms the bedrock of a development philosophy focused on scalable, generalizable architecture for digital scholarship and publication. While individual projects each have their own needs and trajectories, this philosophy directs their attention, collectively, to what they have in common and as a result to what they can share with and learn from one another. We thus avoid bespoke solutions that are hard to scale, and focus
development attention on tools that generalize well within our own suite of projects and hence are more likely to scale usefully to the broader community.

Our end goal is not development of a new tool for Northeastern, but development of new tools that are shared with and receive major uptake from the wider scholarly community; however, widespread adoption of new tools is often more of goal than a reality. To address this problem, we will follow participatory design methodologies. Our opening meeting of expert stakeholders in scholarly communication will help us refine an agenda and vision for the project. During the development process we will work with multiple user communities (graduate student researchers, students in the classroom, scholars in TAPAS and DLoC) to develop user stories and a technical specification. Design methods may include ethnographic user study, focus groups, surveys, storyboarding, and paper prototyping; following the participatory philosophy, prototypes are iterative, with at least 1-2 rounds of feedback and further development involving users. Finally, once we have developed a prototype, we will test it with concrete research projects involving both the ECDA and TAPAS digital collections and communities. User testing methods may include direct observation with talk-aloud narration, online testing with post-test surveys, and questionnaires sent to our distributed user base. Following each user testing activity, the results will be fed back into the design process as feature requests and bug fixes.

**Work Plan**

**January-June 2016**

- announce project to relevant communities and at relevant conferences
- convene first meeting of expert stakeholders
- begin recruiting members of scholarly community (graduate student researchers, students in the classroom, members of ECDA, TAPAS and DLoC) for participation in design process
- profile specific user communities, create user stories, create detailed feature inventory
- prepare specification
- begin preparation of project data: generate document inventories and test document sets for ECDA, Diva360, and DCC
- begin Hydra development work, focusing on foundational components: identity management, community identification and management, moderation and administrative roles
- issue call for 1-2 additional pilot projects

**July-December 2016**

- continue Hydra development work: set up basic research environment framework including navigation, display of member information
- begin development of the open lab notebook, focusing on tools for annotation and highlighting of primary sources
- begin planning and design for support for named entity lists, focusing on their representation in the DRS
- begin planning and design support for community-based ontologies/keyword systems, focusing on the creation, storage, and documentation of keywords
• continue preparation of project data: digitization and encoding of sample documents from ECDA and Diva360
• set up project test sites for DCC, Diva360, and ECDA using the platform
• first phase of user testing, using ECDA community members and DCC focus group: review storyboards and usage scenarios, practice setting up accounts and groups, create notebooks and annotations
• release preliminary user scenarios and technical specifications back to user community where possible
• continue to announce project to relevant communities and present initial work at relevant conferences

January-June 2017
• continue Hydra development work, focusing on extended notebook features: enabling notebooks to be shared and compared, and enabling integration of named entity references in annotations.
• continue development of support for named entity lists, focusing on mechanisms for contribution of new entity records
• continue work on community-based ontologies/keyword systems, focusing on how these are made visible for use in annotations
• continue digitization and encoding of ECDA and Diva360 documents
• continue development of project test sites
• begin documentation of features
• second phase of user testing: further testing of notebook features

July-December 2017
• continue Hydra development work: focus on mining and analysis of named entity references and ontologies, and on mechanisms for referring to and citing annotations and notebooks
• present preliminary results at DH2017 and other relevant conferences
• third phase of user testing: testing of data mining features
• continue development of project test sites
• begin implementation of open notebook, open data, and annotation features in TAPAS
• continue digitization and encoding of ECDA and Diva360 documents
• continue work on user documentation: materials to support usage in classrooms
• recruit faculty partners to test use in graduate or undergraduate classroom
• present Hydra head at Hydra camp, DLF, and other appropriate meetings for feedback
• convene second stakeholder meeting to evaluate project progress and set next steps
• develop final punch list of remaining features and fixes; begin final feature development

January-June 2018
• comprehensive usability testing exercise covering usability and workflow for entire system
• continue implementation in TAPAS
- hold virtual events for ECDA, Diva360, Archives, and TAPAS communities to learn and use the open notebook feature
- conduct testing in classroom with faculty partners or other community partners, developing at least one real-life research project
- begin packaging Hydra components as Hydra gems for dissemination
- release final technical architecture plus results from initial usability testing, present at relevant conferences

**July-December 2018**
- package and document Hydra head for dissemination
- complete the creation of Hydra gems and disseminate
- complete implementation in TAPAS
- complete final phase of bug fixing
- prepare final white paper with technical specifications, summative evaluation, and results.
- present final results at Digital Library Federation Forum
- announce public release of code
- host public virtual launch meeting highlighting top beta users, presentation of pilot projects

**Staff**

**David Cliff** is the Senior Digital Library Developer in the Northeastern University Library. His work focuses on the developing and expanding the campus-wide Digital Repository Service, customizing a local installation of VIVO and the development of custom data loaders for VIVO, coordinating the work of other programmers and developers. He will be responsible for the overall technical architecture of the project, working closely with Patrick Yott, and for overseeing the work of the rest of the technical team. He will also be responsible for developing the prototype Hydra components into finished “gems.” He will work 5% time on this grant, contributed by Northeastern.

**Julia Flanders** is the Director of the Digital Scholarship Group in the Northeastern University Library and interim co-director of the NULab for Texts, Maps and Networks; she also serves as the co-director of TAPAS and the director of the Women Writers Project. She has served in positions of leadership in the digital humanities and TEI communities for nearly two decades, serving on the executives of the Association for Computers and the Humanities (president, 2008-2011), the TEI (chair, 2004-2005), centerNet, and the Alliance of Digital Humanities Organizations (Secretary, 2012-present). She has extensive experience developing and managing digital humanities projects and working with sustainable models for digital humanities services, and deep research expertise in digital scholarly communication. She will serve as the co-Principal Investigator for this project and will work 5% time on this grant, contributed by Northeastern.

**Giordana Mecagni** is the Head of Special Collections and University Archivist in the Northeastern University Library, and she oversees the design and dissemination of digital collections based on the university’s archival holdings. She has deep expertise in archival management with particular emphasis on materials in the history of social justice movements, local history, and community engagement. For this project she will work with the development
team on the Desegregation Collaborative Collection pilot, contributing to the needs assessment and document profiling. She will work 2% on this project, contributed by Northeastern.

**Linda Moss** is the Web Applications Developer in the Northeastern University Library where she works as part of both Library Technology Services and the Digital Scholarship Group. Her expertise serves as a bridge between the library’s repository and technical infrastructure and the scholarly tools that build on it. She currently serves as the lead developer for the TAPAS project, a repository-based platform for archiving and publishing TEI data. She has programming experience in Java, PHP, Python, and C++, and is currently studying Ruby. She also has experience with MySQL, and a background as a metadata specialist, with a BS in Computer Science and an MsLIS in Library and Information Science. She will have responsibility for developing the SHARP user interface and annotation systems. She will work 25% time in the first and third years of the grant, and 50% in the second year, paid by the grant.

**Amanda Rust** is the Assistant Director of the Digital Scholarship Group and the Digital Humanities Librarian in the Northeastern University Library, working closely with DSG’s research collaborators on project design and implementation. She has extensive expertise in library systems for the support of digital scholarship and a deep understanding of how the work processes of scholarship—including collaborative and discursive systems—intersect with technical systems and user interfaces. She also has expertise in assessment, participatory design, and communications and outreach. For this project, she will coordinate the development of pilot projects and the involvement of users in the design and assessment portions of the grant, and will also contribute to the development of DSG policies and documentation relating to the platform. She will work 5% time on this project, contributed by Northeastern.

**Sarah Sweeney** is the Northeastern University Library’s Digital Repository Manager and a member of the Digital Scholarship Group. Sarah brings extensive experience with metadata standards, repository design and implementation, and platforms for digital scholarship. She has led the organization of the DRS Project Toolkit pilot projects, and will serve as the project’s co-Principal Investigator. She will work 5% time on this project, contributed by Northeastern.

**Karl Yee** is the Systems Administrator for the Northeastern University’s Library Technology Services group. He administers 18 Linux-based virtual servers and coordinates library services needs with Information Services (Northeastern’s central computing group). For this project he will be responsible for provisioning, configuring, and maintaining servers and software platforms. He will work 2% time on this project, contributed by Northeastern.

**Patrick Yott** is the Associate Dean of Libraries for Digital Strategies and Services at Northeastern University. He has been working in the digital library arena since the mid 1990’s and has held strategic positions at the University of Virginia (1995-2001) and Brown University (2001-2010). He has taught numerous workshops on XML/XSLT, PHP/MySQL and other digital technologies for the Association of Research Libraries, the American Library Association, the New England Library Network and other regional consortia. Patrick is the lead architect of Northeastern’s DRS and related systems and led the initial development of the Brown University Digital Repository Service prior to coming to Northeastern in 2010. He will work 2% time on this project, contributed by Northeastern.

The **Digital Scholarship Programmer/Analyst** (to be hired) is a new position in the Digital Scholarship Group and provides software design and programming support for digital scholarship projects. The position works closely with the library and faculty colleagues to
develop a sustainable software ecology that can be applied to a range of scholarly projects in the humanities, social sciences, and related disciplines. He/she designs applications that interact with core components of the Library’s digital repository services (Hydra, Blacklight, and Fedora) and programs in such languages as Ruby, PHP, CSS, and JavaScript. The Programmer/Analyst works collaboratively with colleagues in the Library and across campus, and participates in an ongoing evaluation of emerging academic and library technologies. A full job description is included in the Appendix. For this project, the Programmer/Analyst would be responsible for developing the Hydra head through which the core functions of the research and publication platform are implemented, working under the direction of David Cliff. S/he will work 25% time in the first year and 50% time in the second and third year, paid by the grant.

**Sustainability and Evaluation**

The fundamental design of SHARP is oriented towards sustainability, at both the level of digital content and at the level of the platform itself.

With respect to content, all images and transcriptions, and as well as all of the other scholarly content developed by the pilot projects — user-contributed annotations, exhibits and commentary, ontologies, schemas, and supporting documentation — will be stored in the DRS. The Northeastern University Library has made a firm long-term commitment to the curation of scholarly data in the DRS, without cost to the project and without limitation in the case that the project staff changes or leaves Northeastern. This long-term curation also includes automated migration of standard formats (such as PDF, TIFF, and XML) into future versions and formats. Because all of the data formats used by SHARP are based on open standards and have no hardware or software dependencies, the only long-term curation concerns have to do with the evolution of those formats themselves, and the possibility that they may be replaced in common usage by other formats. Over the very long term, there may also be a need for more detailed curation and migration of data beyond what can be handled automatically (for instance, to future versions of the TEI Guidelines). The library’s Digital Scholarship Group is responsible for maintaining the viability of digital scholarly research at Northeastern and will assist projects with specialized data curation tasks.

Each project’s user-related data, including user accounts, discussion forums, personal collection notes, and other data related to specific users will be accounted for and stored as part of its framework, in a database that will be maintained following the library’s standard data management practices including regular off-site backup and appropriate security for personal information.

The SHARP platform itself is also being designed with sustainability as a primary goal. The Hydra head is planned explicitly as part of the library’s overall development and expansion of the Digital Repository Service as a system for scholarly digital publication and research. As part of the project’s design, this Hydra head and its component gems will serve as a model for other DRS-based projects at Northeastern. The library manages its software projects, including code management, issue tracking, documentation, and iterative releases in a public GitHub environment and offers its repository at https://github.com/NEU-Libraries. Code forked from external sources is also managed in this environment thereby maintaining a link to the original development trunk.
There are further sustainability issues connected with ongoing support for the ECDA, TAPAS, and other projects supported by the DSG. Many projects begin through a combination of internal and external funding — some with large initial grants, some with more modest beginnings. Eventually, projects reach the end of a funding and development cycle and enter a maintenance mode, where any ongoing activities are community-driven: acquisition and transcription of further texts for the collection, annotation and commentary (for instance through tools like SHARP), and development of exhibits and curricular materials by community members.

As projects enter maintenance mode, support for their systems will be provided by the DSG, which maintains the Digital Repository Service and associated publication systems, and the Library Technology Services group, which provides server-level monitoring and maintenance. In addition, the library will continue to develop its Hydra framework for repository-based scholarly publications and as new features and interface tools become available they will be integrated into the technology stack available to all DSG projects. Ongoing server costs are approximately $1000 per year; the library is currently developing a joint business model with Northeastern’s constituent colleges for the ongoing support of digital scholarly publications. Projects would seek external funding for any major new development work.

Our partnerships with outside groups — our expert stakeholders, our user communities — will provide one important framework for evaluation. We convene two meetings with expert stakeholders, the second with the express purpose of evaluating this proposal’s progress and looking forward to the next steps. The participatory design process will provide another important framework: as user scenarios and new features are developed, we will return to the user community for feedback, and careful analysis of that feedback will enable us to assess and adjust our development progress. Once we have a basic working system in place (by beginning of the second year) we will begin developing test sites for each of the pilot projects and will conduct usability testing on workflows and features as they are added. We will also recruit at least one faculty member to test usage of SHARP in a classroom context, which will give us valuable information about how the system works in the wild. Finally, we will present preliminary and final versions of SHARP to the developer community (through Hydra and Fedora camps, which are the chief mechanism for code exchange and review in those communities).

**Dissemination and Intended Audience**

Our communication and dissemination plan will focus on two audiences: the developer community (Fedora, Hydra, TEL, Open Annotation, Linked Open Data), and the user community (researchers, students, archivists, librarians). For the purposes of planning, we’ve separated out events into these two streams, but realize that our developer and user communities can overlap. The researcher we envision is one who undertakes scholarship focused on primary sources, who is comfortable working as part of a research community, and who wants to share their early and middle stage work with others; these researchers might be undergraduates, senior faculty, graduate students, independent researchers, archivists, or members of the public. We will plan to incorporate four main user types into our design process — faculty, graduate students, undergraduate students, and archivists — but realize that the initial process of user discovery may cause us to expand or refine our notion of audience.
In addition to national and international audiences, we also have a rich density of local scholarly and repository communities through which we expect to find additional partners, including: Northeastern’s own NULab for Texts, Maps, and Networks; BostonDH; New England Archivists (NEA); Northeast Fedora Users Group (NEFUG); and ACRL New England Chapter (ACRL-NEC).

Finally, to ground our practical on-the-ground design work, we plan to recruit within several existing scholarly communities representing our archetypal audiences: ECDA, DLoC, TAPAS, the University Archives, and Diva360, a burgeoning project based in Northeastern’s Music Department. Our pre-existing relationships with these projects puts us in a good position to design for a concrete audience, rather than an amorphous archetype.

Throughout all stages of the project, we will create and release documentation via our existing departmental wiki. Documentation will not only cover technical details, but also major decisions and lessons learned along the way. We will write at least two informal project updates each year to send out through our own publicity channels (campus news, departmental blog and Twitter accounts, library blog and Twitter accounts).

We will also regularly use existing disciplinary channels of communication to recruit participants, release results, and increase awareness; these channels include but are not limited to email lists, blogs, and Twitter accounts for organizations and projects like HASTAC, Fedora, Hydra, and ADHO.

We will recruit members from existing communities (ECDA, Digital Library of the Caribbean, Diva360, DCC) to pilot and complete a new research project using the new tools. We will also recruit at least one Summer or Fall 2018 instructor to use the new tools in a classroom assignment. We expect that the recruitment and testing process will serve to increase knowledge about these new tools in those communities.

The library staff involved with this project already attend and have an active presence in a wide variety of conferences, from local to national and international. We therefore have ambitious plans for using those conferences as a method to not only raise awareness, but also recruit prototype testers. We will submit presentations for at least three local conferences (New England Fedora Users Group, New England Archivists, ACRL New England) as well as three national/international conferences (Open Repositories, Hydra Developers Congress, Fedora Camp, ADHO, Code4Lib, Digital Libraries Federation, TEI Conference, Balisage, or Coalition for Networked Information).

We will hold virtual launch events for the ECDA and TAPAS communities, showcasing the work of early adopters, the completed pilot research project, and use in the classroom. We will finish with a widely-publicized virtual launch event open to the public, covering the same topics.

Finally, the completed code will be packaged and released via GitHub with broad email and Twitter announcements to the relevant developer communities and conferences, listed above, and we will write and release a white paper with technical specifications, a summative evaluation, and results from the design process.

A detailed schedule can be found in the Dissemination Plan in the accompanying Appendix.