Abstract

Over the last decade, society is seeing a nearly exponential increase in the volume of digital content. Researchers and educators in response see the potential that Big Data techniques bring to computational exploration or cultural and scholarly digital collections for organizing, accessing, and analyzing content. Libraries have long made a mission of provisioning access services to digital content to enrich and improve the lives of all Americans, however, when digital collections have access restrictions, provisioning services becomes a challenge.

We respond to this challenge with the Data Capsule service, developed in the HathiTrust Research Center, that enables remote access to restricted digital data in the HathiTrust Digital Library. Data Capsule is architected to be modular and uses application programming interfaces (APIs) for communication; this best practice in systems design plus proposed effort in packaging, will allow for faster integration into a new environment and ready contributions by third parties.

In this project, we intend to partner with 8 academic libraries across the country in a multi-method research project that draws from human computer interaction and experimental computer science to:

- Understand current library needs and practices in provisioning library services for computational access to special collections having constraints due to sensitivity or restrictions
- Extend the Data Capsule service to broader needs of provisioning for analytical access to restricted collections across a range of collections and uses,
- Study extensions of Data Capsule to cloud computing environments for broader uses
- Identify gaps in skills needed for librarians to enable secure data analytics and provide resources that can address those gaps.

This project proposal, responsive to the IMLS National Leadership Grants for Libraries program, is planned as a 2-year effort. If funded it will be carried out under the encompassing framework of Participatory Design and involve funded partners at Indiana University, University of Illinois, University of California at Berkeley, and University of Virginia; and engaged partners at Indiana University, Lafayette College, MIT, Rutgers University, Swarthmore College, and UCLA.

In response to reviewer feedback, we increased the number of library partners in the project from 3-5 to 8, and introduced the two-tiered partner model. Level 1 partners (2) receive direct funding through the grant. Level 2 partners (6) receive travel funds built into the Indiana University grant to participate in a regional community-building event. The change resulted in an increase of about 15% from the pre-proposal.

Sustainability is planned through utilizing an existing operational service, growing its adopter community (libraries), extending for broader collections and use cases. The service itself is grounded in the HathiTrust Research Center, which continues to support and endorse the Data Capsule service as its primary service for computational analysis on the nearly 15 million volumes of the HathiTrust Digital Library. HTRC deeply welcomes this initiative to involve more partners in use and sustainers of the software code base.

Data Capsule Appliance for Research Analysis of Restricted and Sensitive Data in Academic Libraries

1. Statement of National Need

Over the last decade, society is seeing a nearly exponential increase in the volume of digital content [1]. The new content is coming into existence on the cultural side through massive digitization efforts [2] or because content is increasingly born digital. Libraries have long made a mission of provisioning access services to digital content to enrich and improve the lives of all Americans [3]. When digitized collections (of letters, government papers, video clips, institutional records, annotated volumes) have access restrictions, however, provisioning services becomes a challenge. Collections can have access restrictions for a number of reasons: a set of papers that have not been properly accessioned; a collection of videos with mixed in-copyright and public domain content; material donated by a prominent researcher that contains sensitive information from ethnographic studies on aboriginal peoples. The data-side push for new services to meet the challenge of restricted and sensitive collections is being met with a corollary end user pull, as researchers and educators discover the potential that Big Data techniques bring to the humanities [4] and other areas, and begin to envision opportunity in their own research sphere to the exploration of both small or large collections of materials computationally for organizing, accessing, and analyzing content.

Traditional types of library services often inadequately address end user needs when a collection of materials is restricted or deemed to contain sensitive data. Secure data enclave pilots allow researchers to work with this unique type of data [5]–[9]. Yet such enclaves often are limited to analysis of microdata through common statistical packages, making them less-suited for other uses as there are hundreds of different computational content mining tools, for example, the text analysis portal TAPoR lists 493 of them [10]. Additionally, enclaves are frequently custom-built for a collection, or a small set of centrally located collections, making this solution not so easily portable to new institutions or collections.

Drawing on the most pressing themes of trust, access, infrastructure, and skills in providing data services [11], the overarching goal of this project is manifold: understand current library needs and practices in provisioning services for computational access to special collections, extend an existing service to enable intuitive and yet secure computational access to restricted data in libraries, and identify gaps in skills needed for librarians to enable secure data analytics and provide resources that can address those gaps. We aim to build upon a service that has been developed in the HathiTrust Research Center (HTRC) that enables end users to remotely access the HathiTrust Digital Library for computational use. We propose, as part of this grant, to package the service as an appliance so that it can be easily installed in a library technological environment, and extend the service to satisfy scenarios of different collections and end user needs driven by our library partners. The service is called Data Capsule [12], [13], and it derives from theoretical work on a concept called “storage capsules” [14]. Through a grant from the Alfred P. Sloan Foundation (2011 - 2015) the author of storage capsules, Atul Prakash, along with Plale and McDonald (latter two are leads on this proposal) developed the storage capsule concept into the working Data Capsule service, which became available in HTRC in 2015. The service in HTRC utilizes a tool called the Workset [15], which maintains an end user’s context.

Building on the earlier work of the DC service, we propose to extend and evaluate the system under the encompassing framework of Participatory Design with library partners from eight libraries across the country who have committed to serving as either being Level 1 testing partners, eager to engage in hands on evaluation, or Level 2 partners, ready to participate in discussions and studies. We propose, through this Participatory Design framework, to extend the service to:

- Be packaged as an appliance that can be run and managed locally at partner institutions
- Generalize the Data Capsule service to connect to broader types of restricted collections
- Deliver extensions to the Data Capsule service and Workset model that reflect partner needs obtained through intense partner engagement
- Deliver a design of Data Capsule that utilizes high performance and cloud computing resources that accommodates both large-scale needs of partners and partners with lighter technology resources available to them

As the Data Capsule service is architected using principles of well defined APIs and software component modularity, it is highly suited to extension and generalization for the broader use.

The conceptual framework guiding the architecture of Data Capsule (DC) in its current form can be explained in the context of fair use. Legal judgments of fair use have repeatedly returned to two key analytical questions [16]: First, "*did the use "transform" the material taken from the copyrighted work by using it for a broadly beneficial purpose different from that of the original or did it just repeat the work for the same intent and value as the original?*" And second, "*Was the material taken appropriate in kind and amount, considering the nature of the copyrighted work and of the use?*" In DC, the transforming work is carried out by an end user within a Capsule that they have at their disposal for use for an extended period of weeks to months. The service then enforces both questions as follows:

- Use is appropriate: the DC service assesses appropriateness of the content exported from Capsule:
  - Unintentional exportation such as through malware is stopped
  - Intentional exportation is reviewed through manual (or in future automatic) results review
- Amount of data used is appropriate: the amount of data used in creation of all exported data products is below a threshold of appropriateness of use
- Data types: the type of data used in the creation of new content is allowable for the need
- Intent is reasonable and identity is proven: through structures of policy and institutional infrastructure
- When a Capsule is used for analytical purposes, acceptable activities include but are not limited to a) image analysis and text extraction, b) textual analysis and information extraction, c) linguistic analysis, d) automated translation and language translation, and e) indexing and search.

Data Capsule thus enables transformative use of restricted and sensitive collections through a service that will be packaged as an appliance, will have options for hooking to a new collection with relative ease, and provides the needed assurances that the actions allowable by the service will protect the collection.

2. Project Design

The project is structured to bring together three distinct complementary bodies of expertise: human computer interaction expertise in community engagement, participatory design, and social-technical interactions (Kouper); computer science and technology expertise in data-driven architectures, data models, and trust (Plale, McDonald, and Downie), and library partners with expertise in technology services for special collections (Mitchell and Unsworth). The multidisciplinary team is critical to bring about a project of this nature.

The library partnership is designed at two levels. ***Level 1 Testing Partners*** identify a collection and an end-user need, and work with the Data Capsule team to implement a proof-of-concept demonstration for the collection. Level 1 Testing Partners also participate in the assessment, user study, and participatory activities. They include the libraries of University of California Berkeley and University of Virginia. ***Level 2 Partners*** engage in the assessment and user study, and contribute to participatory activities. Level 2 partners include the libraries of Lafayette College, Indiana University, MIT, Rutgers, Swarthmore, and UCLA.

2.1 Goals, methods, assumptions, and risks

The broad goal of this project will be accomplished through synergistic and mutually reinforcing activity in its two major foci of expertise: in participatory, design-oriented partner engagement and in software architecture and evaluation. The nature of the project is iterative within and between the two foci of expertise: "explore, approximate, and refine" [17].

Research methodologies: The project will employ research methodologies from both the domains of human-computer interaction to accomplish the goals assessment, partner engagement and evaluation, and experimental computer science to advance the Data Capsule design and Workset. This multi-method approach to research is increasingly important in successful technology adoption: active all-stakeholder engagement at the early stages ensures a good fit on the human capital side, and the experimental computer science ensures a good fit on the technological side. The methodologies of each are described in more detail below.

Project risks: Low library partner participation is a potential project risk. We addressed this risk during development of the full proposal by devoting substantially more resources to the library partners. We increased the number of library partners in the project from 3-5 to 8, and introduced the two-tiered partner model. Level 1 partners (2) receive funding through a subcontract that they use for engagement of technical or collections expertise. We additionally built funding into the Indiana University budget to fund travel for Level 2 partners (6) to participate in a regional community-building event. The change resulted in an increase in the overall budget of about 15% from the pre-proposal. We thought this action a necessary risk mitigation strategy. Our project has already built into it a program for constant support and interaction with the library partners on both levels to ensure the highest possible participation.

Assumptions: Our project has several assumptions, all of which we think are reasonable expectations in the environments of major academic libraries, though further study will be carried out for less well-equipped libraries. Data Capsule is an environment (a set of software services plus policies) that utilizes a cluster of computers located within a secure network. The code base is modular and utilizes Application Programming Interfaces (APIs) for extensibility and interoperability. A Data Capsule Controller runs on one of the cluster nodes. From there, it allocates to an end user a Capsule -- a virtual computer (virtual machine) that runs on one of the other nodes in the cluster. The Data Capsule service implementation will be extended in this project to utilize the operating system library, Libvirt[1], which allows DC to configure an end user Capsule for secure access. Our implementation plan thus assumes i) Level 1 testing partners support the existence of a library such as Libvirt running on their testing servers, ii) programmatic access to a collection is available through an API, and iii) there exists a trusted service in the library environment through which user authentication can be carried out.

Research Framework 1: The framework of *Participatory Design (PD)* informs the research questions and methodologies of the human-computer interaction research. A theoretical framework and a set of practices, PD explores conditions for deep user engagement in the design and implementation of computer-based systems at work [18]. User empowerment and democratic decision-making are crucial for successful PD as one of the main assumptions is that technology is being designed to facilitate skilled work and enhance rather than completely replace human labor [19]. Libraries recognize the need to engage their end users in the design of library spaces and technologies [20], [21]. We raise the questions of how librarians themselves can be involved in co-design of tools that use and enhance their skillsets, while, at the same time, enable library end users.

Research Framework 2: *Experimental computer science* as a discipline and methodology forms the framework for assessing and advancing the technological aspects of the project. Through iterative design and prototyping, we reflect user needs in the software development process. Through carefully controlled comparative evaluation studies that are designed to include performance evaluation, we accurately assess different technological tradeoffs. These studies, which are of a quality so as to be published in archival venues, contribute to the diffusion of the project results more broadly through libraries and through time.

Data Capsule is an environment that utilizes a cluster of computers located within a secure network. Capsules have two modes of running: an open mode during which a user can upload tools, data, and software of their choice. During open mode, access to the restricted or sensitive data is blocked. In the second mode, a closed mode, all access to the Internet is blocked, and the channels to the restricted data are opened. This is where the tools that need to work with the sensitive data can be started up. Upon completion of a task, the user stores the results they wish to export to a special directory, where they are queued for manual review, and, upon successful review, the user is sent a URL from which download can occur.

The existing Data Capsule system will be migrated to utilize the Libvirt virtualization toolkit. The Data Capsule Controller is delivered as either a virtual machine image or multiple Docker containers, together with a set of configuration files for partners to customize for their particular environment. The Data Capsule Controller expects two communication endpoints from the partner site: APIs and corresponding SDK/toolkit that can securely access the data collection to be used from capsules; and a trusted user authentication/authorization information relay to the Data Capsule Controller. Libvirt daemons are required to be running on all Data

[1] The virtualization API: https://libvirt.org; runs on Linux, Windows, OSX, FreeBSD

Capsule hosting servers. The Data Capsule Controller will provide RESTful APIs and a basic administration dashboard for partner site to build customized front-end user interface. A separate database is needed to store status of Capsules and their activities, as well as user computation results for the whole system. The Data Capsule Controller is expected to be rather lightweight to run as a single VM. The Docker container approach could provide further flexibility of packaging components and less system resource consumption, albeit be more complicated to deploy [22], [23].

One of the important tools in the Data Capsule environment is the *Workset*. As restricted collections cannot be moved outside of their secure storage and processing environment, users need a mechanism to save a persistent context of their sources that holds information about the state of their activities. HTRC uses the notion of the Workset - a machine-actionable personal research collection described using the Resource Description Framework (RDF) that consists of references to digital objects (e.g., volumes, pages, and so on) and metadata [18]. The Workset model combines pointers to, and metadata about, the generated resources and its selection procedures as well as metadata about bibliographic resources that went into its creation. It provides context and continuity through the research lifecycle, from its conception and creation to archiving, citation, and use by other researchers.

The research questions/issues that we propose to investigate are:

- What are the uses of restricted collections in the context of delivering computational analytical services? How do collection providers and users construct their needs of transformative uses of the collection?
- How do collection-specific services, policies and uses affect the design of DC, and how can DC appliance fit within the library and its technological and organizational models? How do differently positioned actors within an organization influence that?
- Quantify the performance implications of certain design tradeoffs in extending and generalizing the Data Capsule system to meet the needs of a broad set of library uses and environments.
    - Include in the study an assessment of tradeoffs when considering libraries with less well equipped technical infrastructures
- Evaluate the tradeoffs for extending the Data Capsule system to allow user Capsules to utilize high-performance compute resources inside or external to an institution, and run large analysis tasks.
- Evaluate the different models for Workset use in the Capsule for different use and collection needs.

2.2 Specific activities

Element 1: Assessment	Work with partners to map out collection specifics and the contexts of their use; prioritize needs in co-design and implementation; organize events to bring participants together as a community. Employ parallel theoretical reflection and continuous exchange of knowledge.

Tasks

- Research team interviews partners to gather information about collections and the context of their uses, identifies collection-specific characteristics as well as work practices that may impact development and implementation of DC. Access restrictions, storage, security, and analytical needs as well as the relationships between collection users, stewards, and technical support will be included. User needs as seen by librarians or taken from previous feedback of actual users (e.g., types of data analysis, tools used) will also be identified.
- Examine policies and other factors that affect the use of restricted data and DC. Collect and analyze documents that govern access and use of the restricted collections.
- Organize community-building events possibly co-located with regional HTRC UnCamp events to increase participation; organize regular information-sharing sessions.

Outcomes

- Effective coordination, sharing, and networking with all partners
- Taxonomic knowledge about restricted collections and their policies and contexts of use
- Emerging sense of community
- Community building meetings

Element 2: Partner Engagement	Engage the technical team, Level 1 testing, and Level 2 partners in close cooperation. Level 1 testing partners each have an installation of Data Capsule on an experimental set of machines of their choice.

Tasks

- Technical team and Level 1 testing partners engage in mutual exchange about collection constraints, infrastructure constraints, technology options, and solutions for prototype demonstrations with partner collections. Carry out continuous installation, evaluation, and feedback cycles to refine.
- Engage library partners in Participatory Design. Participatory activities and evaluation of appliance, which will include demo of Data Capsule prototype and Workset reflecting co-designed functionality; installation of Data Capsule at Level 1 partners; continuous install of extensions at Level 1 partners, evaluation of improvements for all partners.
- Visit workplaces of Level 1 and 2 partners for purposes of information exchange, assessment and learning.

Outcomes

- Shared knowledge and understanding
- Participant-influenced design of technologies
- Better fit of technology to needs
- Lowered barriers to adoption for partners

Element 3: Data Capsule	Extend existing Data Capsule service to enable intuitive and yet secure computational access to restricted data in libraries. Evaluate extensions through demos, prototyped functionality, and evaluative studies.

Tasks

- Design, develop architecture for packaging Data Capsule as an appliance
- Extend data capsule system's architecture to
  - i) Enforce proper access of restricted and sensitive collections,
  - ii) Support access to multiple collections having diverse formats and types,
  - iii) Support range of use models needed by partners. Implement selective changes in form of prototype demo for feedback.
- Design evaluative study of DC as capable of utilizing high performance or cloud computing resources to serve institutions with various resources including less equipped institutions. Carry out performance experiments evaluate different design tradeoffs

Outcomes

- Extended code base of Data Capsule packaged as an appliance with support for new collection types and use cases. Code base released with appropriate user and developer documentation.
- Published proof of concept study of how Data Capsule can be scaled to use large-scale compute resources at an institution or at a cloud provider such as Amazon Web Services
- Published study of design tradeoffs in enhancements to support new use cases and access modes to restricted and sensitive collections

Element 4: Workset	Evaluate Worksets within the context of the project's new uses and users to improve the utility and impact of Worksets in the scholarly research process.

Tasks

- Participate in assessment and participatory activities to gather information about the applicability of the current Workset model to specific collections.

- Design and carry out study that evaluates tradeoffs to extension of Workset model to accommodate the new uses of Data Capsules for computational access to restricted and sensitive collections.
- Bring Workset to state to participate in demos showcasing new Data Capsule functionality
- Actively engage library partners in exploring how best to educate users on optimal practices for Workset use and reuse.

Outcomes

- Educational materials for a researcher's best utilization of the Workset notion in the distant analysis that this project enables
- Publishable study of design tradeoffs for extending Workset to additional collections and uses

2.3 Project management

The project will be led by Beth A. Plale with direct oversight and responsibility for project success. Dr. Kouper and Robert McDonald will serve as co-Directors. The leadership team including J. Stephen Downie at University of Illinois will meet weekly, and be joined once a month by the Level 1 Testing Library partners. Decision making with is carried out through consensus building with the final decision resting with the PD.

Dr. Plale also brings technical expertise, and in this Plale will work closely with Dr. Yu (Marie) Ma, Dev/Ops manager of HathiTrust Research Center, to ensure that the technical staff members are tasked appropriately for the project needs and timelines. Dr. Inna Kouper will lead the project assessment and community building activities using Participatory Design methods and carried out in collaboration with partner libraries. Robert H. McDonald will coordinate the partner libraries. Level 1 partner libraries will supervise prototyping and testing of digital collections. J. Stephen Downie will coordinate expertise on the Workset.

Bi-weekly videoconferencing meetings carried out for community building will be held using the Zoom.us conferencing system that IU provides free to its research groups. Technical communication with Level 1 (and level 2 as interested) partners, which tends to be frequent and short during joint efforts, will utilize a Slack.com channel. Stakeholder interactions will be via regular teleconferences and phone calls. User studies will be conducted online using screen-sharing and recording tools such as Zoom in addition to in-person visits.

Issues raised by library partners needing immediate attention of the Data Capsule and Workset technical team can utilize the HathiTrust Research Center service desk built on the Atlassian Jira Service Desk and bug tracking system. Software development and project management computers, grants management staff, and office space needed for the effort at Indiana University are provided by the Data To Insight Center. The other funded universities will provide similar resources needed for accomplishing tasks. We will utilize computer resources such as Amazon Web Services as needed for testing.

As this is a research grant, evaluation and performance measurements are built into the outcomes. That is, published results are amongst the planned outcomes. The findings from assessment and Participatory Design will be shared and discussed with developer and librarian teams during regular meetings. Ongoing feedback will be incorporated into the findings.

2.4 Project dissemination and sustainability

Recommendations from this project can be adopted in diverse library settings; the surveys and community building efforts can bring together many stakeholders in data, including researchers, librarians, university administrators, and funding agencies. Results of the project will be disseminated through multiple professional, academic, and social media channels.

Community building is a key part of the project. Community building user meetings from this project will be considered to become part of the regular HTRC UnCamps -- hybrid conference-workshop events already a part of HTRC's community engagement plan. Changes to the Data Capsule code base undertaken during this project will be committed back to a new project branch of the existing Data Capsule code repository (https://github.com/htrc/HTRC-DataCapsules). As an intended outcome of the Participatory Design framework of this project, library partners, especially Level 1 partners, will be actively contributing to the code branch by the end of the project. This will create a broader community around the code base, thus giving a strong foundation for its sustainability. The changes to the Data Capsules system, including the Workset, are anticipated to also benefit the instance running in the HathiTrust Research Center, creating another pillar in the foundation of sustainability for the framework.

3. National Impact

The proposed project will have national impact through i) provision of a portable solution for accessing restricted and sensitive collections, ii) fostering a community and increased collaboration around the technical, organizational, and policy challenges of providing computational access to restricted collections, and iii) amplifying project outcomes through the connection to HathiTrust Consortium and its hundreds of member libraries. Our portable solution, once in shareable form, can be reused by other libraries around the country, where experts can improve the code and documentation as well as digital curation activities, and work with their users to develop new requirements and materials to use restricted digital collections in research and teaching. An emerging community will become part of the larger HathiTrust community and will continue stimulating libraries and research and non-profit organizations to join forces in further development and mutual learning and support. A strong sense of contribution and collaboration around community-sustained software will help to have a long-lasting impact.

Addressed needs: Through its development and participatory activities, this project will broaden access to digital collections that exist in libraries, including papers, letters, video-materials and many others. It will not only establish a community dedicated to working on solutions for restricted collections, but also develop a strong foundation for motivating and engaging future generations of library experts in developing innovative software and services. Project outcomes will address the library needs of providing scalable tools for working with digital collections, while respecting privacy, copyright, and confidentiality restrictions, and contribute to building the National Digital Platform as a distributed set of software applications and professional expertise that provide library content and services to all users in the US [24].

In addition to providing a strong prototype, we will help train librarians and professionals involved in developing technology via support from and collaborations with our technical team and via targeted community events. We will support communities of practice and strengthen libraries as partners in addressing the research and scholarship needs of computational research.

Resulting products: This project will result in the tangible products of extensions to the existing code base for Data Capsule, to guidelines and educational materials, and publications. The intangible product is community buy-in towards adoption and community involvement in ongoing contributions to the DC code base. The tangible products enable proliferation of experience and facts beyond the immediate library partners to increased adoption. Publications, for instance, are a tangible outcome that facilitates trust in technology and human work. Research is grounding for assessments of use.

Sustaining the benefit: The sustainability of the benefits of the proposed activity extends well beyond the period of funding. It is an important point that this activity will vault an existing and successful service into broader use through study and extension, and will do so in a way that builds its adopters (libraries) into the process thus growing the sustaining community through the grant duration.

Growing adopters and a sustaining community around the software code base can take time, likely more time than the short grant duration. This risk is mitigated because the service itself is grounded in the HathiTrust Research Center, which stands behind the Data Capsule service as its primary service for computational analysis on the nearly 15 million volumes of the HathiTrust Digital Library. HTRC deeply welcomes this initiative to involve more partners. As expected outcome of this project is to have partners outside the HTRC technical team making contributions to the code base, the HTRC commits to incorporating those changes back to the main branch of the Data Capsule code base and use the extensions in future releases of Data Capsule for its own and broader use.

Schedule of Completion

	2017			2018				2019	
	Apr - Jun	Jul - Sep	Oct - Dec	Jan - Mar	Apr - Jun	Jul - Sep	Oct - Dec	Jan - Mar	Apr - Jun
Task									
Award - May 2017	⟺								
Task/element I: Assessment									
Preparation for assessment	⟺								
Assessment of collections, policies and contexts of use		⟺							
Preparation for community building events			⟺						
Community building events				⟺	⟺				
Carry out publishable analyses of collected assessment and participatory design data						⟺			
Support stakeholder / community interactions			⟺	⟺	⟺	⟺	⟺		
Conduct online user studies							⟺		
Publish training materials								⟺	
Publish results									⟺
Task/element II: Partner engagement and evaluation									
Plan DC install	⟺								
First install in test environment		⟺							
Partner campus visits			⟺	⟺					
Guided hands on experience and cross institution learning			⟺	⟺					
Co-design and evaluation of appliance				⟺	⟺	⟺			
Demo DC and workset reflecting participatory design functionality							⟺		
Continuous install, evaluation of improvements					⟺	⟺	⟺	⟺	
Integrate project developments into DC code base and release									⟺
Task/element III: Data capsule development									
Design for appliance architecture		⟺							
Development: code changes to package as appliance			⟺	⟺	⟺				
Using feedback from assessment, refine design plans				⟺	⟺	⟺			
Carry out publishable study that evaluates different design tradeoffs						⟺	⟺		
Design evaluative study for DC as thin client to HPC resources							⟺		
Carry out development study of DC as thin client							⟺	⟺	
Evaluate and integrate changes in main DC branch							⟺	⟺	
Publish results									⟺

Develop and release user and developer guides									↔
Task/element IV: Workset study and development									
Develop study of workset in this setting		↔							
Conduct study of workset			↔	↔	↔				
Using feedback from assessment, refine design plans				↔	↔	↔			
Carry out publishable study that evaluates different design tradeoffs						↔	↔		
Evaluate and integrate changes in main workset/workset builder branch							↔	↔	
Publish results									↔

DIGITAL PRODUCT FORM

Introduction

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

PART I: Intellectual Property Rights and Permissions

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

The formal products produced as outcome of our proposed effort are software, training materials, user and developer documentation, and studies. We anticipate intermediate products emerging as well in the form of datasets derived from testing of the connections to restricted and sensitive collections. The formal materials and software products resulting from this effort will be licensed using open and free licensing, e.g., Creative Commons and Apache 2.0-style licenses, following the best practice established by the HathiTrust Research Center (HTRC). Intermediate products emerging as a result of testing and experimentation will be discarded by the end of the project life. While operational use of a Data Capsule service at a partner institution is not anticipated over the course of the project, should it occur, or should use of HTRC’s operational Data Capsule service be used for training, then the data products emerging from end user use of a Capsule will follow the HTRC policy of not imposing licensing restrictions on the products assuming that the Data Capsule service that the end user is using is fully operational and the data products pass the review process (run by HTRC). If the conditions are not met, the data products are considered intermediate products and will be destroyed by end of project life.

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

Software products developed in this project will be openly shared and accessible via an open software repository (Github). As to access to the Data Capsule service, during the course of the project there will be test instances of Data Capsule service running at the Level 1 library testing partner institutions, and an operational instance running at Indiana University as part of HTRC. We anticipate the test instances of Data Capsule service having no end-user uses during the course of the project as they will be under development. Training will be carried out on the operational HTRC instance of the Data Capsule service.

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

As part of this project, we will be conducting interviews and taking notes during ethnographic with explicit consent from participants. Additionally, restricted collections that will be used during testing in computational analysis in Data Capsules may raise copyright, privacy or other concerns. These concerned will be addressed through policy discussions with library partners; these discussions may be guided by HTRC's policy developed to address similar concerns.

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

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

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

In the course of this project the following digital content will be created:

1. Extensions to Data Capsule service. The extensions will start from the existing HTRC codebase, which is organized in approx. 50 modules. It is expected that modifications will touch 10-20% of the code for partner customization.
2. Enhancements to the Workset model. This resource is an Ontology that can be expressed in RDF and/or XML formats. Enhancements will comprise about 10% of the resource.
3. Interview recordings and transcripts and fieldnotes. See Part IV Datasets for more details.
4. Online manuals and training materials. Installation, testing and use of Data Capsule will be documented in online manuals and training materials, which will be openly accessible via the web.
5. Publications and presentations. Findings from the project will be disseminated via journals, conferences, and other venues. PDF documents and slides will be openly shared with the community, unless publishing restrictions apply.

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

The project activity will be to develop software extensions to existing code bases and conduct human-computer interaction studies. Activity does not extend to the creation of digital collections. We intend to use computers at Indiana University, University of Illinois, University of Virginia, UC Berkeley, and UCLA for testing and development. We expect Level 1 partners to have test servers available on which we will install the software (Data Capsule).

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

Software will exist in development formats, predominantly Java files, Python scripts, and XML configuration files. Partner libraries who will use the operational Data Capsule service at HTRC for analyzing their restricted collections, may have derived products in other formats that are appropriate in their respective user disciplines, such as tabular files or images. Quality standards for those derived products as well as quality challenges will be discussed during participatory design activities. Software quality will be monitored and evaluated by using "fitness for purpose" and structural analysis techniques.

B. Workflow and Asset Maintenance/Preservation

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

For details on software quality control, see Part III.

The assessment is carried out by a PhD research faculty member who is highly trained in carrying out quality processes. Dr. Kouper has a strong record of publication quality research in this area. Software development will use HTRC's software development processes, including oversight by a DevOps Manager, helpdesk, and bug tracking. Studies of Data Capsule and Workset will be under the supervision of Plale

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

Software products will be shared, preserved and maintained using the open software repository Github. Technical documentation will be stored on GitHub as well as on the open HTRC wiki pages. We will encourage HathiTrust community and the emerging Data Capsule community to further contribute to curation and preservation of the software. Products of research (publications, datasets, and presentations) will be preserved in Indiana University institutional repository IUScholarworks, which will serve as an additional preservation layer to traditional publication venues.

C. Metadata

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

README files, user and developer guides are the form of documentation used to preserve software metadata. For datasets we will use Dublin Core to record description, administrative, and preservation metadata.

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

Metadata will be maintained as part of the software and data maintenance, i.e., it will be stored and migrated along with the digital products.

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

As the project is not concerned with creating a digital collection, we will rely on other larger resources for widespread discovery and use, including HathiTrust Research Center networks, academic publishing databases, and software and institutional repositories.

D. Access and Use

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

Software and study products will be openly available online, unless the latter is restricted by the publishers.

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

The Data to Insight Center has its own group repository on GitHub where all software products are made available to the public: https://github.com/Data-to-Insight-Center Most recent examples include Data MatchMaker https://github.com/Data-to-Insight-Center/Data-MatchMaker and PRAGMA Data

https://github.com/Data-to-Insight-Center/PRAGMA-Data-Reposit

Additionally, D2I contributions to HTRC code are made available via separate repository https://github.com/htrc, where the existing Data Capsule codebase can be found https://github.com/htrc/HTRC-DataCapsules.

Part III. Projects Developing Software

A. General Information

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

To accomplish the goals of this project, we will extend the Data Capsules service code base. HTRC Data Capsule works by giving a researcher a virtual machine (VM) that runs within the HTRC domain. The researcher can configure the VM as they would their own desktop with their own tools. After they are done, the VM switches into a “secure mode”, where network and other data channels are restricted in exchange for access to the data being protected. Currently, Data Capsule works only with the HathiTrust Digital Library and within HTRC architecture. We will generalize the architecture to work with other collections and evaluate design, secure access and scalability options to work in specific library environments.

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

Comparable conceptual frameworks that intend to perform similar functions include Data Enclaves and Storage Capsules. Data Enclaves rely on customized virtualization software and pre-defined set of tools to enable access. To the best of our knowledge, no working software exists that addresses the need to perform computational analysis on documents and resources using a researcher-defined set of tools. As the need for computational research on restricted collections using a large variety of tools grows, the development of such software is undoubtedly significant and necessary.

B. Technical Information

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

Data Capsule software is in Java, Python, and shell scripts.

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

The software extends the Data Capsule service.

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

Data Capsule uses open source virtualization infrastructure (QEMU and KVM), which needs to be installed for the capsule to work.

MySQL relational database system is used to store capsule metadata and results.

Data Capsule is provided for Ubuntu (Linux) environment.

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

The code will be forked in GitHub repository, creating a new branch. Contributing developers will be using their environment to write code and then commit the code back to GitHub. We will use HTRC documentation and bug-tracking services (Atlassian Confluence and Jira) for maintaining and updating

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

The Data to Insight Center has its own group repository on GitHub where all software products are made available to the public: https://github.com/Data-to-Insight-Center Most recent examples include Data MatchMaker https://github.com/Data-to-Insight-Center/Data-MatchMaker and PRAGMA Data

https://github.com/Data-to-Insight-Center/PRAGMA-Data-Repository

Additionally, D2I contributions to HTRC code are made available via separate repository https://github.com/htrc, where the existing Data Capsule codebase can be found https://github.com/htrc/HTRC-DataCapsules.

C. Access and Use

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

We will use Apache 2.0 license to release Data Capsule. The license allows to reproduce and distribute copies of the software and its derivatives with or without modifications. The license text is put to use by adding it to the header of a software file (see https://www.apache.org/licenses/LICENSE-2.0 for a copy of the license).

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

The source code extensions to the Data Capsule will be made available via GitHub https://github.com/htrc as a separate branch of the primary branch.

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

Name of publicly accessible source code repository: GitHub

URL: https://github.com/htrc

Part IV: Projects Creating Datasets

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

Data will be collected via phone interviews and ethnographic observations, which involve note-taking, recording, and photographs. Phone interviews will be conducted at the beginning of the project. Follow-up interviews and additional recordings of conversations and note-taking will take place throughout the project as a need to document participant interactions will arise.

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

Data collection involves human subjects and requires IRB approval. IRB application will be prepared and submitted when/if the project is approved for funding.

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade

suppression PII, or synthetic data).

Participants can be identified in phone interviews, notes, and recordings. Personally identifiable information will be stored securely and only PI and co-PIs will have access to it. Before public release of the dataset all PII will be removed (participants will be assigned coded numbers and any information that may identify them individually will be obscured in the interviews, notes, and transcripts).

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

Participants will be provided with informed consent forms, which they will sign. The forms will be stored securely and separately and the relationship to the collected data will be maintained via a study ID that will be recorded in the informed consent forms and in the data files.

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

The data will be collected via interviews and observations and will consist of text files, audio and video files, and photographs. Common word processing software and multimedia players may be used to display the data. Processed data may consist of additional spreadsheets and visualizations, which will be stored in non-proprietary formats (e.g., CSV or PNG).

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

Codebooks will be created as part of the analysis of qualitative data (e.g., in the thematic coding procedures codes will be developed in the inductive manner, after close iterative reading of the interviews). Codes, their descriptions and other documentation that describes when and where the interviews and observations took place will be stored in text formats along with the data. The documentation will be associated with the datasets through consistent file naming and through identifiers that refer to each data collection effort separately.

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

The data will be managed and archived using Scholarly Data Archive (backed-up storage for long-term archiving) and institutional Google Drive at Indiana University (for active work with data). Folders with appropriate permissions for data, processing scripts, IRB documentation, and publications will be created. For dissemination, we will use IU Scholarworks repository and one of the publicly available repositories, such as Figshare or Mendeley.

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

Name of repository: IU Scholarworks; Figshare; Mendeley Data

URL: scholarworks.iu.edu/dspace/; fighare.com; data.mendeley.com

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

PIs will monitor the implementation of this data management plan. The plan will be reviewed every 6 months and adjusted according to the amounts and types of data generated.