HathiTrust Research Center: Computational Research on the HathiTrust Repository

PIs (exec mgt team): Beth A. Plale, Indiana University; Marshall Scott Poole, University of Illinois Urbana-Champaign ; Robert McDonald, IU; John Unsworth (UIUC) *Senior investigators:* Loretta Auvil (UIUC); Johan Bollen (IU), Randy Butler (UIUC); Dennis Cromwell (IU), Geoffrey Fox (IU), Eileen Julien (IU), Stacy Kowalczyk (IU); Danny Powell (UIUC); Beth Sandore (UIUC); Craig Stewart (IU); John Towns (UIUC); Carolyn Walters (IU), Michael Welge (UIUC); Eric Wernert (IU)

Project duration: Sep 01, 2010 – Aug 31, 2016

The HathiTrust Research Center (HTRC) is dedicated to the provision of access to a comprehensive body of published works for scholarship and education. This great public good will provide sustained access to works in the public domain and, on more limited terms, to publications in the copyright domain for computational research purposes. The HTRC will provide a persistent and sustainable structure to enable original and cutting edge research. It will stimulate the development of new functionality and tools to enable new discoveries that would not be possible without the HTRC.

Of HTRC volumes that are subject to copyright laws of the United States and other countries, some are additionally subject to the terms of the Google Books Settlement. The outcome of the proposed Settlement is unclear, but the general approach of the HTRC is to be in compliance with Settlement standards and specifications for security in order to readily support Settlement-covered materials when the Settlement is approved.

The HTRC will evolve in two phases. In the first phase, Phase I, attention will be directed towards building a prototype facility that meets the basic needs of the several classes of uses that we have identified using the public domain (noncopyright) volumes in the HathiTrust Repository. This phase will build the relationship between UIUC and IU in this effort and produce a system that spans both universities. The success metrics for this phase will focus on demonstrated capability with an architectural design able to fully scale to meet anticipated needs and sufficiently rigorous to serve as a foundation for Phase II. Phase II of the project involves scaling up to full operational capacity through support of in-copyright works in the manner of *non-consumptive research*. Because of the uncertainty of the timing and outcome of the Settlement, Phase I will also be used to demonstrate success upon which additional funding beyond the Settlement can be sought.

The HathiTrust Research Center is being designed to make the technology serve the researcher - to make the content easy to find, to make the research tools efficient and effective, to allow researchers to customize their environment, to allow researchers to combine their own data with that of the HTRC, and to allow researchers to contribute tools. The proposed HTRC is founded on a distributed cyberinfrastructure (CI) at Indiana University and University of Illinois Urbana Champaign that meets the specific needs of long-term secure research and analysis of the core HathiTrust text corpus. The CI is shaped by three dominant factors: anticipated modes of use, large volumes of data, and the requirement of non-consumptive research. The CI we propose utilizes existing infrastructure and analysis tools and data access services as the foundation for the a distributed, persistent service infrastructure that integrates HTRC capabilities with software actively used in digital humanities research and into the context of ongoing research community discussions with the HTRC Advisory Board.

Access to the HathiTrust Repository we anticipate being through a bridge between the HTRC data store located at IU Bloomington to the mirrored HathiTrust collection at the IU data center in Indianapolis. The HTRC data store will consist of full or cached copies of the public domain volumes and in Phase II, the in-copyright works. Computational resources include resources located primarily at UIUC, and to a lesser degree at IU. The security infrastructure will be designed by the CyberSecurity Directorate of NCSA. To ensure a long-term foundation of policy and practice for administering the HTRC we will examine both governance and sustainable funding models for the center.

I. Introduction

The HathiTrust Research Center (HTRC) is dedicated to the provision of access to a comprehensive body of published works for scholarship and education. This great public good will provide sustained access to works in the public domain and, on more limited terms, to publications in the copyright domain for computational research purposes. The HTRC will provide a persistent and sustainable structure to enable original and cutting edge research. It will stimulate the development of new functionality and tools to enable new discoveries that would not be possible without the HTRC.

This proposal brings together expertise in cyberinfrastructure (CI) and CI research, data mining, information retrieval, and digital libraries at Indiana University and the University of Illinois Urbana-Champaign to solve the difficult challenges of increasing access to the public domain and copyrighted material in HathiTrust (HT). We propose to achieve this through pursuit of the following ***long-term goals***:

- Support ***innovation in cyberinfrastructure*** to deliver optimal access and use of the HathiTrust corpus. The sheer size of the corpus demands innovative thinking about the architecture and the optimization at all levels of the software infrastructure from disk to tools. Research and development could focus on reducing reads, intelligent caching, and delivering maximum cycles at minimal costs, among other things.
- Explore innovation in delivering efficient access to copyrighted material that preserves and shapes the restriction of "non-consumptive research", that is restricting the ability to reassemble pages from the collection. ***"Non-consumptive" as a research challenge*** is approached through deeper study of the constraint and recommendations for tooling adaptations to satisfy it.
- Identify and ***host existing data analysis, text mining and retrieval tools***, such as MONK and SEASR, that either work on the public corpus or qualify under terms of "non-consumptive research". In addition, the HTRC can stimulate the development of new analytical methods and tools.
- Seek ways to ***enhance the value of HathiTrust through additional integration indexes and analysis tools*** that integrate HathiTrust data and non-textual time-series works, such as the Media Preservation Study's identified collection of 560,000 audio and video recordings on the Indiana University campus that are being digitized.
- Explore innovative methods for creating a ***sustainable research center*** to provide ongoing access to the HathiTrust corpus by seeking new short term grant funding as well as investigating models of sustainability.

At the outset it is important to delineate the structure of the effort with respect to the HathiTrust repository. The repository offers long-term preservation and access services, including bibliographic and full-text search and reading capabilities for public domain volumes and some copyrighted volumes. The Research Center on the other hand, provisions for computational research access to the HathiTrust collection. Limited reading of materials will be possible in the Research Center to accommodate needs for reviewing results, etc., but the primary destination for reading-based research will be the HathiTrust repository.

The HTRC will evolve in two phases. In the first phase, Phase I, attention will be directed towards building a prototype facility that meets the basic needs of the several classes of uses that we have identified using the public domain (noncopyright) volumes in the HathiTrust Repository. This phase will build the relationship between UIUC and IU in this effort and produce a system that spans both universities. The success metrics for this phase will focus on demonstrated capability with an architectural design able to fully scale to meet anticipated needs and sufficiently rigorous to serve as a foundation for phasing in non-consumptive research involving the copyright material in the HathiTrust in Phase II. The computational and storage resources in Phase I will be modest because of the limited access to public domain works only. Phase II of the project and beyond involve scaling up to full operational capacity.

II. A User-Centric Vision for Research Technology

The HathiTrust Research Center is being designed to make the technology serve the researcher - to make the content easy to find, to make the research tools efficient and effective, to allow researchers to customize their environment, to allow researchers to combine their own data with that of the HTRC, and to allow researchers to contribute tools.

The Call for Proposal has an extensive analysis of user needs that include structural, semantic and/or syntactic element extraction and analysis; creating new resources such as dictionaries or indices; character, location, concept, mood, topic, or political analysis and aggregation; linguistic analyses such as parts of speech, word counts, and changes in usage and meaning over time. These analyses needs have been validated in a number of workshops and venues held over the last several years (Unsworth et al. 2006, Summit 2005, Cohen et al. 2009). We additionally anticipate that the HathiTrust collection will be used as a test bed for developing new text and image processing tools. In service of these broad needs, we have identified three classes of users: the *humanities scholar* who expects an intuitive and simple interface to hide the underlying technologies, the *informatics user* who will use the HTRC corpus to test new algorithms, data mining techniques and information retrieval processes, and the *independent operator* who will download data for processing within a separate environment. We expect that researchers may cross from one category to another during the course of their work; that is, a humanities scholar may become an informatics user or an independent operator depending on the nature of the research task at hand. The HTRC infrastructure will not build barriers between these categories but will facilitate the flow of data and functions. We introduce several use cases in the humanities and informatics domains to illustrate the type of advancements HTRC can bring.

Humanities Scholar 1 from History and Philosophy of Science. A scholar is interested in determining how the work of Nicholas Flamel influenced Isaac Newton. She has a complete bibliography of Flamel's published works, another bibliography of Isaac Newton's published works, and access to the text of Isaac Newton's alchemical manuscripts encoded in TEI (the Text Encoding Initiative) [http://www.tei-c.org]. She logs in to the HTRC. She uploads her two bibliographies, verifies that all of the titles are present, and then uploads her set of TEI files of the Newton's manuscripts. She chooses to find all verbatim quotes in the manuscripts from Flamel. She selects a text comparison tool, provides the parameters such as the names of the collections to compare and the precision requirements for the match. After the comparison tool has completed, she reviews the results online and can, if she finds the results useful, save the results in the HTRC infrastructure for further processing. She adjusts the precision parameters until her results are satisfactory.

Humanities Scholar 2 from English Literature. A professor studies a specific poet and wants to find all references in the collected works to water. He uses the HTRC to create a collection of the poet's works and chooses the Thesaurus tool in the HTRC to create a set of synonyms for water. He executes the function to find references, saves his results for further exploration and discovery.

Humanities Scholar and Informatics User from Linguistics. A scholar interested in stochastic grammars has developed a new statistical method that he expects to more realistically describe the construction of long sentences. Primarily a humanities scholar, he will use some of the functions envisioned for the informatics user. He has an R program that he wants to test. He logs into the HTRC and uses its structural analysis tools to create a personal collection of very long sentences. He then registers his R program with the HTRC. Because his new algorithm is experimental, he has marked this program as private so no other researcher can access it. He executes his program against his selected collection. He is able to plot his results using the HTRC visualization tools and can download the resulting graphs to his computer.

Informatics User 1. A researcher develops a new algorithm for automatic structural feature detection in images resulting in XML encoded text. She identifies a set of images that are useful as a testbed, specifies her request, and is informed that she will receive an answer to her request within 48 hours. After her request is granted, she submits her C program to HTRC within a self-contained virtual machine (VM) container which is executed in parallel on her behalf on the HTRC computing resources. She spends several weeks testing and refining her program and gathering results. When her work is completed, the VM image is removed from the HTRC.

Informatics User 2. A researcher develops a new concept clustering algorithm and wants to use the HTRC to test. He has completed his initial testing and would like to offer his tool for use by others. He registers his tool with the HTRC as an experimental tool that is available to all HTRC users. The HTRC "Tools" Committee investigates his tool and approves inclusion of it in the portal. When other researchers find his tool, they can easily identify the experimental nature of the tool and can choose to participate in the further testing of the tool. They can contact the creator of the tool, engage in a dialog, and provide feedback.

Independent Operator. A researcher has developed a new search engine and wants to test the indexing and retrieval performance in a controlled system environment. He logs into the HTRC and downloads the entire public domain text to his file system and is able to run his tests with this huge corpus.

The HTRC will conduct regular user evaluations to test the usefulness and usability of the tools, the interface, the policies, and the overall research experience. We will use the data to plan and prioritize functional and technical enhancements.

III. Cyberinfrastructure

The proposed HathiTrust Research Center is founded on a distributed cyberinfrastructure that meets the specific needs of long-term secure research and analysis of the core HathiTrust text corpus. We use the accepted definition of cyberinfrastructure as "computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible" (wikipedia 2009).

The cyberinfrastructure of the HTRC is shaped by three dominant factors: anticipated modes of use, large volumes of data, and the potential requirement of enforcing non-consumptive research. The current collection of OCR documents in the HathiTrust Repository at the time of writing is approximately 225TB in size (uncompressed). Even under a strategy where portions (but not all) of the collection is cached in HTRC for research access, the storage requirements are formidable. To that requirement is the additional need for indexes that provide rapid response to queries for data, and assets like thesaurus, all of which can serve as an endpoint for inquiries themselves. An index can be anywhere from half the size of the collection to 1.5 times the size of the collection meaning storage needs could be as high as 750 TB[1]. The public domain OCR documents are a 37TB collection that together with indexes has a requirement of 100TB. We discuss user modes in Section III.A and non-consumptive research in III.B; here we outline the cyberinfrastructure.

The cyberinfrastructure (CI) we propose utilizes existing infrastructure and analysis tools and data access services as the foundation for an infrastructure designed for evolution in response to increasing user sophistication and emerging technology and information research. Specifically, we propose a distributed, persistent service infrastructure that integrates HTRC capabilities with software actively used in digital humanities research and into the context of ongoing research community discussions with the HTRC Advisory Board. Access to the HathiTrust Repository we anticipate being through a bridge from the HTRC data store located at IU Bloomington to the mirrored Hathi collection at the IU data center in Indianapolis (Fig. 1). The HTRC data store consists of full or cached copies of the public domain volumes and in Phase II, the in-copyright works. These will reside in a file system structured in the parallel tree structure used in the HathiTrust. The indexes will be maintained in the data store as will space for temporary storage of user data that is needed to carry out user computations, and a database to maintain versioning and audit activity. Computational resources will be located primarily at UIUC, and to a lesser

[1] By employing a storage strategy that caches only a portion of the collection instead of the full collection, the 750TB estimate could also accommodate anticipated need to store accumulated derived products (i.e., research results) as well.

extent at IU. FutureGrid resources have been made available to HTRC for prototyping data-driven solutions such as MapReduce (Dean and Ghemawat 2004). The underlying resources are brought together into the distributed HTRC through a wide area file system. The several use modes we propose, shown in Figure 1 sitting above the WAN file system, are discussed further below. The cyberinfrastructure is accessed through a web-based portal component that provides a friendly pre-defined suite of analysis tools and more advanced data analysis tool sets that can be command line executables or modules. End-user capabilities will include programmatic interfaces for direct retrieval operations from within community data acquisition and content generation, analysis and visualization software.

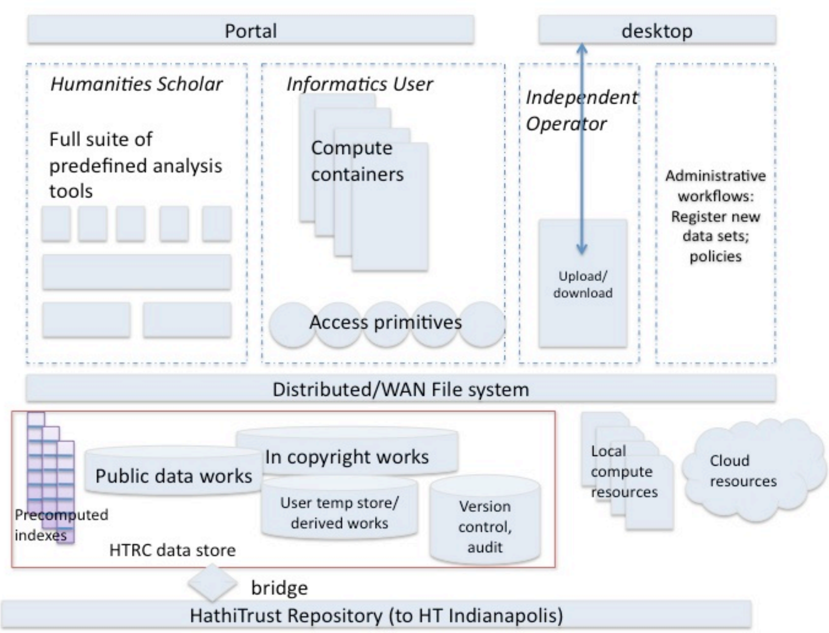

Figure 1. Logical architecture of the proposed HTRC system.

The HTRC would like to explore the possibility of mirroring the HathiTrust Repository Solr indexes for data mining of in-copyright materials. Pending legal agreement between the HTRC and the University of Michigan, the HTRC may be able to use copies of the Solr indexes for the entire corpus of materials, including in-copyright works to enhance access to the HTRC researchers. While these indexes will not permit unconstrained access to the content, they will facilitate very flexible and fast word and phrase queries.

The hardware and software systems and technologies utilized in this project leverage the excellent computational, networking, and storage infrastructures in place at Indiana University (IU) and the University of Illinois primarily through the National Center for Supercomputer Applications (NCSA) and builds on those strengths. The partner universities have substantial technology infrastructure upon which the HathiTrust Research Center can build. We have developed an initial network architecture of the HTRC federation that is designed to leverage the combined resources and expertise at the two universities utilizing three distributed yet tightly-coupled data centers. Two of these data centers are located at the IU campuses in Bloomington and Indianapolis. The other data center is at the campus of UIUC (NCSA Data Center), which has a similarly robust operation and long-term operational experience. The configuration, showing existing and planned network connectivity between the sites, is shown in Fig. 1. As both sites connect directly to the Internet2 Network (http://www.internet2.edu/network/), each has excellent connectivity with a minimum of 10 Gbps WAN capacity. In addition, both are connected to the TeraGrid backbone at 20 Gbps.

III. A. Use categories

Based on the assessment of user needs in the HathiTrust RFP and independent interviews with potential users, we define three distinct modes of use as mentioned earlier. The first class is the *humanities scholar* who expects user-friendly services. The technical details are largely transparent, and the workflow model mimics the user investigative process. The *informatics user* is technology-astute, and is developing a new data mining task for large corpuses for example. The *independent operator* simply wants to download as much of the corpus as bandwidth and their storage solution allows and work with it in their own environment.

Figure 1 depicts the cyberinfrastructure organization reflecting these modes. To the left in support of the humanities scholar is the full tool suite enabling research on the collection to be carried out in a friendly user-facing way. In Phase I of the center, this functionality will be provided by the SEASR/Meandre environment (Llora et al. 2008). As part of SEASR, a set of text analytics components will be deployed to process text using existing natural language processing tools like OpenNLP and Stanford NER, in addition to transformation and visualization tools like Simile Timeline, GoogleMaps, Flare, and Protovis. In Phase II the functionality of SEASR/Meandre will be improved to address additional text analytics, transformations, and visualizations. Consideration of and support for other analysis tools and frameworks will be ongoing. See section III.E Systems and Tools for other examples.

Support for the informatics user is envisioned as layered on top of a set of basic file access primitives that prohibit inappropriate access, viewed both as a snapshot and historically, to the corpus. By routing all access to the corpus through the primitives, which could be a thin layer on top of the file system API, one can more easily validate non-consumptive research. Higher level access primitives could be the indexes themselves. Access is through parallel Virtual Machines (VMs) or through Meandre, which has a Do-It-Yourself layer that could also be suitable. Finally, the independent operator users will be supported through existing upload/download protocols for exporting, compressing and downloading the public datasets.

Research Results Management. Managing the research results will be an important part of the infrastructure of the HTRC. Research result sets can originate from a number of HTRC activities: searching the OCR text, searching for images, executing HTRC analysis tools, executing user provided tools and programs, as well as other research actions. We distinguish between the results of an individual's research activities, and selected results that could benefit the center as a whole or could refine the HathiTrust collection so are considered for inclusion as Center assets. The research results sets of an individual may need to be saved for a short period. Result sets need t reviewed by HTRC before download is allowed to ensure compliance with the restrictions of non-consumptive research. All research result sets will be protected via the authorizations and authentication mechanisms detailed in the security section III.C below.

III. B. Data Use Restrictions and Non-Consumptive Research

As a mater of both policy and practice, the HTRC will comply with all legal requirements for the data including any requirements that are yet in negotiation such as the Google Settlement. The HTRC will be a strong advocate for open access to researchers to all of the materials in the HathiTrust Repository and the HTRC. The HTRC will work with rights holders to secure permissions for additional access to copyrighted materials for researchers.

The HTRC will be designed with the capability to provide access to a body of digital works originating from different sources and carrying different access and use restrictions. The HTRC will leverage the rights database in HathiTrust. But we anticipate that the HTRC will need additional data to that in the HathiTrust and will record in a database any additional access or use restrictions per volume. As part of this project, clear policies will need to be developed and communicated to the researchers who use the HTRC. Access and use restrictions are anticipated as follows:

1. ***Public domain volumes scanned by Google.*** There are no legal restrictions on use of these materials, and thus no restrictions for use within the HTRC. There are, however,

restrictions that Google imposes on their further distribution. Under current agreements, public domain texts that Google has digitized may not be redistributed without Google's permission. Those that are distributed may not be re-hosted or used commercially. The HTRC will be responsible for making the connection between researchers interested in using these materials and Google. In general, the HTRC will be responsible for managing records related to this distribution process and whether, for example, the researcher wishes updates to the corpus.

2. ***Public domain volumes not scanned by Google.*** There are no legal restrictions on non-Google-digitized public domain volumes that HTRC will receive. Contributing institutions are not expected to place restrictions on public domain volumes they deposit.

3. ***In copyright volumes both those scanned by Google and those digitized via other means.*** Of HTRC volumes that are subject to copyright laws of the United States and other countries, some are additionally subject to the terms of the Settlement. The outcome of the proposed Settlement is unclear, but the general approach of the HTRC is to be in compliance with Settlement standards and specifications for security in order to readily support Settlement-covered materials when the Settlement is approved. Further details are included in Section II.F. For materials covered by the Settlement (and perhaps for other materials whose rights holders), the HTRC will develop a *non-consumptive research* environment. We use the following definition of *non-consumptive research*:

> *Non-consumptive research -- No action or set of actions on the part of HTRC users, either acting alone or in cooperation with other users over the duration of one or multiple sessions can result in sufficient information gathered from the HathiTrust collection to reassemble pages from the collection.*

III. C. Secure Cyberinfrastructure.

A number of challenges exist around the safeguarding of the digital documents including digital rights management which may include copyright protection enforcement, limitations on redistribution, access control on restricted documents, and privacy issues related to the tracking of data access by individuals. As documented in the above section there are three data policy categories that will need to be enforced, which include 1) Public domain volumes scanned by Google, 2) Public domain volumes not scanned by Google, and 3) In copyright volumes both those scanned by Google and those digitized via other means. In addition the HTRC data must be secured from all unauthorized access regardless of intention. The HTRC will deploy sufficiently strong security measures in place to guarantee that persons who are not approved HTRC users will not be able to obtain access to the holdings.

The CyberSecurity Directorate (CSD) of NCSA, at the University of Illinois, will design and deploy the HTRC security cyberinfrastructure. CSD are experts in this field and are involved with and or lead a number of similar efforts including CILogon (that provides a credential translation service for the National Science Foundation), TeraGrid, Open Science Grid, GENI, Blue Waters, and NCSA's own production HPC environment. CSD will perform a threat and risk security analysis and from that design the security architecture, document policies, and design, develop and deploy the security infrastructure and procedures. From this thorough approach we will understand in detail the requirements, as well as the threats and risks to HTRC. We will fully document all policies and define those that are missing, and then this process will culminate in the design of the security architecture which then will lead to the specific implementation plan. What follows is a high-level description of a few of the obvious security protection categories and our suggested approach to these however. However implementation details will be finalized only after we have completed the architecture process just described.

Modern approaches to protecting on-line access of data or services involve mechanisms to authenticate (identify) users and then use that identifying information to link to that user's access rights. The access rights are then used to grant rights to services, directories, files, functionality (read/write) and potentially more. Logging of activities such as login/logout, what data was accessed, or what services were requested can also be important for auditing and an important information source when responding to security incidents.

Authentication and Authorization. The HTRC will authenticate and authorize all users, which will require that all users of HTRC be registered. During an online registration process new users will present digital identity credentials that will then be mapped to their HTRC account. From that point forward the original identity credentials can then be used to authenticate to HTRC. The digital identities we are considering include OpenID, that provides users with a unique identifying URL, and SAML, which relies on trusted identity providers and allows the user to use their institutional identifier. The OpenID based identity is available typically through commercial providers for anyone to request an identity. The typical implementation of OpenID however relies on a self-asserted identity vetting process, meaning the user self certifies their identity. The OpenID approach could provide a simple, lightweight approach to non-privileged users to browse and read unprotected documents, or use the provided tools while still allowing for the tracking of their actions. The typical SAML based approach, for example in the academic InCommon federation, is much heavier in that it typically requires organizational participation and a more thorough vetting of the user's identity. The InCommon approach could work well for providing access to restricted documents, as the confidence in the user's identity is higher. Another approach we are investigating is the use of one-time-passwords (OTP) via a hardware token that is provided to the user. This is a fairly heavyweight approach as tokens need to be registered with individual users, handed out, tracked and supported. However, it can provide a high level of identity assurance and may be appropriate for high-level access to restricted documents or services. OTP is a form of two factor authentication, in that it requires both something you have, the token, and something you know, the PIN or password to unlock the token, and as a result can provide a high level of assurance and may be most appropriate for the highest level assurance requirements such as for system administrators.

Registered users of the center may only access HTRC data, and these users will be classified by role: privileged users and general users (formerly known as researchers). There will be additional administrative roles that will be restricted to trusted center staff that are tasked with the system administration of the center. There may be additional roles that will be uncovered as we do a more thorough requirements gathering and threat/risk assessment. Privileged users will have access to the data, will be able to upload and execute programs and tools against the data and can download data. General users will be able to use the HTRC provided tools and will not have access to materials covered by the settlement or with other special restrictions. Researchers whose primary affiliated institution is outside the United States are prohibited from accessing books covered by the settlement will be granted general user access. Privileged users will undergo an additional one-time authorization process in the form of a policy agreement and must sign that agreement to uphold all of the policies of the HTRC including data distribution restrictions and the rules of non-consumptive research. The signed policy agreements will be archived and the privileged account will then be activated. Privileged users will need to be monitored to ensure they are not violating policies and their programs may need to be vetted to protect against malicious intent or policy violations such as downloading all of the HTRC data for redistribution.

Audit. The center will log the actions of each user in order to provide an audit trail. There are obvious privacy concerns when capturing and storing user actions and to address this we will utilize privacy enhancement techniques that include access protection on the files, directories and audit services that limits access only to authorized administrators. We will also protect the identities of the users by utilizing techniques such as the FLAIM [http://flaim.ncsa.illinois.edu/] log anonymization routines developed by the NCSA at the University of Illinois.

Secure Sessions. Protection of the restructured data will also include providing assurances that the data is protected in transport. All access to restricted data will be supported via secure sessions such as SSL/TLS. HTTPS is a common SSL/TLS implementation that provides an encrypted communication channel between a web browser and the server thereby reducing the chance that an unprivileged individual could capture restricted data while it is being transported over the network.

Physical facility, storage and network. Disk/tape storage will be in a physical facility that has secure, restricted access. Tapes will be utilized for less-frequently accessed content, and for backups as part of the disaster recovery process. Multiple copies of tape content will be created with both on-site and off-site storage and access to all these archives will be restricted physically to only authorized center administrators, further all access to the tapes will be monitored.

All systems, networks, and services within the center will be monitored and will require secure logon in order to access them. Wireless access to machines in the facility will require secure login as mentioned above. Wired or wireless network protocols must prevent eavesdropping and tampering, such as through SSL/TLS. Both IU Data Centers provide this level of security, as does the NCSA data center. Further all systems will be managed in a secure manner with timely security updates. The compute resources and software (such as the operating system) upon which operations on the collection execute cannot be compromised. Procedures must exist to ensure there are no vulnerabilities in the machines that execute center operations. This is provided today by competent system administrators who monitor new viruses, keep machines up to date, and generally stay on top of things.

File system. Reads/writes to the collection will be through a file system so the file system must be secure. Authorization controls will trigger off user identities as outlined above. File systems will read/write either a local disk or to network accessible disks. Network read/write access utilizes a network protocol to read/write/copy file chunks. That protocol has to ensure no eavesdropping or tampering. In the first phase of operation we intend to use Lustre for WAN file system access to resources across IU and UIUC. This is a decision of convenience as UIUC and IU both have Lustre installed and have done performance tuning on it to get good performance. Lustre however has no data encryption; hence we will examine Glustre for adoption in Phase II.

Security Monitoring and Incident Response. The HTRC facility will be monitored both for physical access as well as digital access. We will employ the latest monitoring techniques to ensure against malicious actions including the establishment of intrusion detection systems (IDS), network, host and application firewalling to block known attacks and vulnerabilities. When security incidents are detected our incident response security team will respond with a thorough investigation that will result in a complete understanding of the incident, and its impact. Measures will then be taken to restore the systems and protect them against future similar attacks.

III D. Content Ingest, Versioning, Visualization

HTRC will receive periodic, automatic updates of pages and volumes from HathiTrust. The frequency of updates must strike the right balance between maintaining the freshness of the collection, and conserving computational resources as indexes must be rebuilt. A volume received from HathiTrust that is a correction /refinement to an existing volume will be distinguishable from a new volume. It will also include reference to the earlier version of the volume. The new version will be recorded in a database maintained by HTRC that identifies every page and volume to which it belongs. Currently HathiTrust references operate at the volume level, but it will be desirable to move to the page level when possible. The HTRC will work with the HathiTrust repository to determine how this may be best accomplished. Each page will be assigned a globally unique identifier (URI, UUID, DoI). The persistent database containing this information will be secured and maintained as the primary HTRC data. The database will be replicated using either regular backups or if access speeds becomes an issue, using a master/slave replication model.

We propose to build into the data management system support for *repository versioning* to provide researchers with a permanent citable link to a snapshot of the HathiTrust collection taken at the time in which the research was carried out. HTRC will always provide the most recent version of a page for research use. However, in addition, the system will support landmark events (Santry et al. 1999), which snapshot the collection at a moment in time. The snapshot is a list of pointers to the most recent version of each page over the entire collection. Once created, it is assigned a global ID (URI or DoI) and these global IDs are published. The question exists of how often a snapshot should be taken. We anticipate the rate of new versions to be small. That is, we anticipate the total number of new versions generated over the course of a year to affect 1% of the collection. With this low rate of change, landmark snapshots taken monthly should be adequate. Additionally, while it is feasible to reconstruct a view of the collection at the time of a snapshot, support for this operation will only be put in place if user demand is high as it is a highly costly operation to implement. Landmark events would be scheduled by and monitored by system administrators.

Visualization tools will play an essential role in helping researchers explore, understand, and communicate large-scale data from the HTRC repository and from computational analyses against this data. Creative visual representations may also play a supporting role in helping to expand the types of analyses that may be conducted under the banner of "non-consumptive research". The first phase of visualization development will survey and incorporate commonly utilized desktop information visualization applications and environments such as NetworkWorkbench, R, and GGobi along with web-based APIs such as Simile Timeline, Flare, and ProtoVis. Such APIs can be used for generating small standalone visualizations as well as integrated "mash-ups" with other tools like Google Maps. Service offerings in this phase would include a common visualization software stack at all HTRC sites, file format conversion tools, educational resource listings, sample tools built with the APIs, and methods for researchers to share their custom tools. The second phase of deployment will focus on inherently scalable tools. The Titan toolkit is a collaboration between Sandia National Labs and Kitware that builds on the proven success of VTK and Paraview to create an API (Titan) and client-server application (Overview) for scalable, grid-enabled information visualization (Titan URL 2010). ParaText is a recent extension and customization of this framework to enable scalable text modeling and analysis (ParaText URL 2010). As we deploy these tools in the second, scale-out phase of HTRC, we will encourage users to transition to them whenever possible as a way to guarantee scalability as their research and the collection's data stores grow.

III.E Systems and Tools

The following systems and tools have been identified for evaluation for use in the HTRC. A plan for deploying systems and evaluation of other possible services and tools during the two phases of development of the HTRC is outlined in Section II.F.

- **IU DataCapacitor –** the primary storage resource utilized for HTRC will be the NSF funded Data Capacitor that offers a 339 TB file system available to the project using LUSTRE-WAN which is mounted at both NCSA/UIUC. The long-term storage at IU will be provided on a 3.6 PB HPSS hierarchal storage management system that provides redundant replication between the IU data centers. It consists of Dell 2950s running LUSTRE with a measured sustained I/O rate of 14.5 GB/s. A bridge will need to be built to support data replication from the HT Node in Indianapolis that supports HT content using the Isilon OneFS distributed file-system.
- **Lustre-WAN –** the Lustre file system supported over a wide-area network such as Internet2 or TeraGrid.
- **GlusterFS –** GlusterFS uses the idea of using clusters of independent storage units and combine them into one large storage server. This concept can have performance increase compared to other networked file systems, because every node has its own CPUs, memory, I/O bus, RAID storage and interconnect interface. These units can have an aggregated performance increase. GlusterFS is designed for linear scalability for very large sized storage clusters.
- **Isilon OneFS –** OneFS is Isilon's fifth-generation operating system software that provides the intelligence behind all Isilon® Scale-out NAS systems. It combines the three layers of traditional storage architectures—file system, volume manager and RAID—into one unified software layer, creating a single intelligent file system that spans all nodes within a cluster. In addition to enabling industry-leading scalability of performance and capacity, OneFS combines mission-critical reliability and high availability with state-of-the-art data protection to help storage administrators do more with less. This file-system is used to manage the two nodes of the HathiTrust repository at IU and UM.

- **HubZero –** is a platform used to create dynamic web sites for scientific research and educational activities. With HUBzero, you can easily publish your research software and related educational materials on the web. This could be the front-end portal for a variety of web-based services for the HathiTrust Research Center.
- **Blacklight** – is an end user discovery interface especially optimized for heterogeneous collections. It can be used as a library catalog, as a front end for a digital repository, or as a single-search interface to aggregate digital content. Blacklight is an open source system built with ruby-on-rails that uses Solr for its search engine (Blacklight URL).
- **Karma** – The Karma tool is a standalone tool that can be added to existing cyberinfrastructure for purposes of collection and representation of provenance data. Karma utilizes a modular architecture that permits support for multiple instrumentation plugins that make it usable in different architectural settings. It can synthesize a provenance picture across the activities of a tools and services at different levels in the software stack. It can be used in HTRC to capture information about user activity, and can be used to capture metadata automatically to aid in sharing and discovery of information (Simmhan et al. 2006).
- **Audit Control Environment (ACE) –** A system developed at the University of Maryland that incorporates a methodology to address the integrity of long-term archives using rigorous cryptographic techniques. ACE continuously audits the contents of the various objects according to the policy set by the archive, and provides mechanisms for an independent third-party auditor to certify the integrity of any object.
- **Sector/Sphere –** supports distributed data storage, distribution, and processing over large clusters of commodity computers. Sector is a high performance, scalable, and secure distributed file system. Sphere is a high performance parallel data processing engine that can process Sector data files with very simple programming interfaces. Sector/Sphere can be broadly compared to Google's GFS/MapReduce stack (Grossman and Gu 2008).
- **Meandre –** is a semantic-web-driven data-flow execution infrastructure to construct, assemble, and execute components and flows. Meandre is designed to: (1) provide a robust and transparent scalable solution from a laptops to large-scale clusters to clouds to petascale platforms, (2) create an unified solution for batch and interactive tasks in high-performance computing environments, and (3) encourage applications and software component sharing though a centralize application repository or Do-It-Yourself rapid application environment. Meandre also has a scripting language called ZigZag that allows users to easily describe, with Python like simplicity, data intensive flows which can be complied into self contained applications. Meandre will provide the infrastructure for performing the data access and analysis of HT data as well as returning results to users. This environment can be used to create or interact with an interface for the humanities scholars' analysis. The informatics scholars who want to create their own analysis can leverage existing components for creating custom solutions.
- **SEASR –** the Software Environment for the Advancement of Scholarly Research (SEASR) is a software instantiation of Meandre for the humanities. The SEASR software 1.) Enhances humanities researchers' ability to use digital humanities applications for knowledge discovery, and 2.) Provides digital humanities developers with an improved environment for advancing and innovating applications. As part of the SEASR project, a collection of text and data mining, as well as visualization techniques have been componentized for use in Meandre. We will deploy a set of analysis flows that were created during SEASR in the HTRC for the humanities scholars. The existing components can be used and recombined for the informatics user.
- **XMC Cat –**is a metadata catalog that provides search and retrieval of research results through storage of rich metadata on behalf of a user. It coexists in a cyberinfrastructure environment as a standalone web service, and plugs into the architecture's security

mechanisms. Its advantages include adaptability to domain schemata through configuration instead of code changes, support for automatic capture of metadata through curation plugins, and search and browse capabilities through a web-based GUI. (Jensen and Plale

- **Zotero –** Zotero is a Firefox extension that allows researchers collect, manage, cite, and share research sources. The Zotero environment can be used to create and manage collections of documents that will be used for analysis in the HTRC environment. A SEASR extension to Zotero provides users with the ability to submit items / collections for analysis through the SEASR Analytics environment and retrieve, display, and store the results in Zotero itself. The existing SEASR analysis deployed in Zotero could be customized to access the HT data and return the results (user authentication will need to be added for non-public data).
- **MONK –** Metadata Offer New Knowledge is a digital environment designed to help humanities scholars discover and analyze patterns in the texts they study. MONK consists of a datastore, middleware, an analytics engine, and various user-interfaces, of which the MONK Workbench is the most developed. The MONK project also spent time on some related proof-of-concept work (like faceted browsing for selecting worksets from large collections, or using Zotero to pass those collections into the MONK Workbench). The datastore was produced by an ingest process that used XSL routines collectively referred to as Abbot, a part-of-speech tagger called Morphadorner, and a database loader called Prior, all of which were developed wholly or in part during the MONK project. MONK middleware handles traffic passing back and forth among the user interface, the datastore, and the analytics engine. The analytics engine is SEASR (the Software Environment for Advancement of Scholarly Research), and it takes information from the user (for example, ratings of texts in a supervised learning scenario), combines that with the actual data from the datastore, and runs user-specified statistical routines (in particular Naive Bayes, Naive Bayes with Decision Tree, and Dunning's log likelihood ratio, for comparing and classifying different works or collections) to produce text-mining results.

IV. Governance

Executive Management. The HTRC will have an Executive Management Team that is responsible for day-to-day decision-making and fiscal and administrative duties. The proposed membership of the Executive Management team membership is Beth Plale, Marshall Scott Poole, Robert McDonald, and John Unsworth. The executive management team will engage with the *HathiTrust Executive Committee* through quarterly or semi-annual meetings and through monthly activity reports. This close engagement provides HTRC an opportunity for feedback on expenditures, direction, and progress, and allows HathiTrust Executive Committee an opportunity to identify synergies that could benefit HTRC.

Executive Board. HTRC will have an Executive Board that has both an advisory capacity of taking a longer-term view of the center and advise on strategic direction, and a decision-making capacity of taking major strategic decisions for the HTRC. Its membership will include representation from the legal community, researchers that have need for the Center or who are sensitive to the research context, members of the library community, and primary members of the HTRC Consortium. Primary members of the HTRC Consortium are defined as institutional members who make the standard contribution to maintenance of the HTRC [See Sustainability, Section IV]. The Executive Management Team will be nonvoting ex officio members of the Executive Board.

The Executive Board will consist of 20 or more members and will convene at least twice yearly. In its advisory capacity the board will recommend long-term directions and new projects for the HTRC. It will also inform the HTRC about issues the users find important, representing another channel for user requests and concerns in addition to the networks of the Executive Management Team. In its decision-making capacity, the board will make decisions on major matters confronting the HTRC, including approval of the yearly budget, major capital expenditures, and new service initiatives. Only the primary members of the HTRC Consortium will have a vote in these decisions, though other members may contribute to the discussion. At each meeting of the Executive Board the Executive Management Team will make a report on the state of the HTRC, usage levels, and initiatives and problems.

Possible members of the Executive Board include: Dan Cohen GWU; Greg Crane, Tufts; Bill Neuman, IU; Ray Siemans, U of Victoria; Susan Schreibman, Digital Humanities Observatory, Royal Irish Academy, Ireland; Jay Kesam, UIUC School of Law; Brett Bobley, NEH

Human resources for management of content access. The plan for managing content depends upon dedicated human resources committed to the task of interpreting restrictions in the context of research use, translating expanded definitions of non-consumptive research into technical and cyberinfrastructure implications, and vetting requests for research. When the settlement comes through, HTRC will be beholden to the Book Rights Registry to ensure appropriate use of the content. We will also want to fully exploit technical advances in non-consumptive research.

Assessment. The Executive Management Team will prepare an annual report and submit it to the Advisory Board for discussion at one of the two yearly meetings. A subcommittee of the Advisory Board will then conduct a thorough review and submit recommendations for the HTRC.

Engagement of university legal council. Legal contacts at IU, UIUC, and the University of Michigan, which bears legal responsibility for the HathiTrust repository, will be engaged to manage legal issues and questions.

Management processes. The following management processes will be developed by the Executive Management Team, in consultation with subject and technical experts at IU and UIUC, during the first year:

- Audit trail and provenance collection. Auditing captures a system's level of compliance with respect to a set of procedures. Provenance collection captures the lineage of a research results as a function of the processes and data objects that effected its generation. Both are required to establish and enforce the requirement of non-consumptive research.
- Security implementation plan
- Authentication and access controls
- Process for documenting responsibility for granting access
- Policies and procedures for handling derivative results

V. Sustainability Models for the HathiTrust Research Center

Sustainability during the initial phase of the HTRC will focus on codifying the relationships of the institutions involved in developing and sustaining the long-term leadership, management, policy, and resources associated with the research center. To ensure a long-term foundation of policy and practice for administering the HTRC, we will need to look at both governance and sustainable funding models for the center.

It is difficult to predict a single strategy that will provide reliable funding for a HathiTrust Research Center on a continuing basis, so it makes sense to experiment with a diverse set of strategies, and learn from the results. Any business plan, though, will have the effect of optimizing for some outcomes and discouraging others, so we need to be clear about our goals. A HathiTrust Research Center should encourage and support the use of its resources. It should provide incentives for its users to share code, share intermediate work products, and share research results. Finally it should partner as early as possible in the development of grant proposals for projects that would draw on its resources.

Libraries also have an interest in what happens in a HathiTrust Research Center—not only those libraries whose collections are represented in the HathiTrust, but any research library that represents a user community interested in computational work with text. For this constituency, then, the HathiTrust Research Center could

- Relieve libraries of the need to build local research environments
- Provide access to high-performance computing resources
- Allow users to enrich descriptive metadata for the collections on which they work, and return that metadata to the libraries
- Create a mechanism for crowd-sourced proofreading of scanned text, returning corrections to the authoritative source, held at the University of Michigan

These goals speak to the research community, but in order to develop sustainable operations, the Center also needs goals that speak to publishers and others with a business interest in the content in the Center and/or the outcomes of research performed there. So, for example, the Center could

- Provide publishers with insight into how their products might be used in the future, and how new derivative products might be used as well
- Allow publishers and other research service providers to observe and participate in the building of new tools and services for users of text collections

With respect to copyrighted material, security is an obvious interest for publishers, and an obvious cost for the Center, but there could be more to it than that. For example, the Center could provide a testbed for new techniques for protecting copyright while still allowing use, as digital content becomes the norm in publishing. Or it could participate in tracking and promulgating rights information.

What business strategies would speak to the motivations of these three constituencies—research users, research libraries, and publishers? Here are some suggestions:

- Advertise a charge for grant-funded projects using the resource and/or staff support associated with it, so researchers can build that charge into their budgets, and invite them to consult on technical matters, at no cost, as they develop their proposals.
- Offer free access to those who contribute or update software at least annually, and promote the use of a standard open-source license for software developed in this research environment.
- Offer free access for teaching, but require deposit of syllabi and sample assignments.
- Offer free or discounted access for those who contribute data that others can use, but charge those who want to keep private the data they upload or produce.
- Develop a two-tiered subscription model for research libraries: one tier for those who contribute material to the HathiTrust and would like to get back improved metadata or corrected texts, and another tier for libraries that have not contributed materials, but would like to provide access for their researchers.
- Establish a standard set of operating procedures for bidding on the right to commercialize software, techniques, or services that have been prototyped in the research environment, and make the right to bid contingent on membership.

Depending on the outcome of the Google Books Settlement, there may be other functions the Center could play. For example, if the settlement does not go through, publishers may still be very interested in a rights registry and rights clearinghouse for orphan and copyrighted works. The Center itself would need to have good information about rights in order to allow researchers appropriate access to large parts of the collection, so there may be some mutual interests to be served by collaborating with publishers on establishing and maintaining a registry, and if so, it is possible that the cost-recovery model proposed in the settlement—a foundation funded by sales proceeds—could work outside of the settlement as well. In fact, the value of this activity, all by itself, could be sufficient to carry not only its own costs but some substantial part of the costs of the Center as well.

Beyond all of these strategies, there will be opportunities for the Center itself to apply for grant funding for developing its own infrastructure, for experimenting with new kinds of scholarly communication, and for prototyping new kinds of library services. All in all, from a diverse array of sources, the Center probably needs to generate several million dollars a year through activities that will require a business office, fiscal technicians, and legal counsel. HathiTrust member universities may be able to provide some of these services, perhaps in exchange for other considerations in the form of subscription waivers, etc.

In order to take advantage of leveraged administrative support for the sustainability of the HTRC, one model for development would be to form a 501c3 or join a 501c3 not for profit organization that fits with the research goals of the HTRC. Note it has been discussed at the level of the HathiTrust of developing this type of group. One discussion path would be whether the HTRC could be a part of a HathiTrust developed 501c3 organization or whether this should be decided outside of that discussion? If decided outside of this discussion, then it is likely that one institution would need to volunteer in the short term to manage the joint finances and act as a lead institution of the HTRC until such time that an independent 501c3 organization is developed. In light of this, what follows are models that describe both 501c3 and institutional bound sustainability models. 501c3 sustainability models include: the Kuali Foundation (http://www.kuali.org); the Educopia Foundation (http://www.educopia.org); the Sakai Foundation (http://sakaiproject.org); the DuraSpace Foundation (http://duraspace.org); HubZero (http://hubzero.org); and Internet2 (http://www.internet2.edu). Institution linked non-profit sustainability models include: LOCKSS/CLOCKSS (http://lockss.stanford.edu/http://www.clockss.org); PORTICO (http://www.portico.org); and the Text Encoding Initiative (http://www.tei-c.org). Some trust-bound sustainability models include: the Digital Library Federation (http://www.diglib.org/); and the Bill and Melinda Gates Foundation (http://www.gatesfoundation.org).

Initial Plan for Sustainability. All of the possibilities above will be explored as the HTRC develops, but in terms of specific plans (which may change depending on circumstances):

1. The first year of phase 1 (startup) will be carried out using resources supplied by IU and UIUC (equipment and personnel) and will be devoted to submitting grant proposals that can support additional personnel.
2. If the settlement comes through then it will be used to execute Phase II. If the settlement does not come through then the HTRC will develop a copyright registry service to market to publishers.
3. Additional members will be invited into the HTRC during Phase II and there will be a two-tiered subscription system, as described above. The rates will be set after consultation with possible members on a basis similar to the current HathiTrust arrangement.
4. Rates for supporting research using the book corpus will be determined, and those who write grants to use the corpus will be asked to include support for the HTRC in their proposals.

VI. Education, Outreach, and Training

Effective outreach and training are central to HTRC's vision and supports its mission through a multi-faceted program to engage and train our researchers, educate future researchers, and create partnerships that build on existing education programs. The outreach program consists of: 1) In-person and self-paced *training workshops* for researchers and educators to enable and motivate HTRC usage and 2) creative and proactive outreach. The outreach program will *broaden participation* and reach out to under-represented communities and disciplines and will employ *evaluation criteria* to assess and ensure the success of this and other efforts and activities. An outreach coordinator will lead the HTRC outreach program with responsibility for training workshops, informal education, and outreach. We will seek additional grants to support the EOT program, including Research Experiences for Teachers (RET) and Research Experiences for Undergraduates (REU).

Training workshops. Training is the core of our outreach program. In Year 1 we will conduct domain-specific content-development sessions to identify the requirements of researchers, students, system developers, and educators. As HTRC facilities become available, the training workshops will begin. The workshops will provide feedback to the HTRC developers on the usefulness and usability of HTRC tools and systems and supply information necessary for iterative design as user needs change. We target a program of two intensive 2-day workshops the first year, and three annually thereafter, each reaching 25 users, and using online materials and shorter training sessions at conferences and other venues to reach a much larger audience. We will create synergies between HTRC workshops and other training programs that are already underway at the partner institutions.

Sakai-based learning systems at IU will provide a mechanism for wide delivery of content as will dissemination through The Alliance of Digital Humanities Organizations (ADHO).

Outreach. This activity will involve a strategic variety of online and face-to-face activities to promote awareness of HTRC, grow the user communities, and form new partnerships and relationships. In addition to effective, but standard, approaches such as attending stakeholder meetings and conferences to publicize HTRC, we will use the power of social network analysis to promote HTRC accomplishments as well as recognize contributors and users. For example, we will analyze usage data to highlight the most downloaded and used data over particular time periods, and provide information on popular HTRC tools. We will disseminate information and services as widely as possible by publishing both on the web and in journals as mean to broaden the impact of HTRC.

VII. Development and Management Plan

The project will be managed by co-leads Beth Plale of IU and Scott Poole of UIUC in close consultation with the other members of the executive management team, John Unsworth of UIUC and Robert McDonald of IU. We will have regular weekly leadership and technical meetings to ensure milestones are being reached.

HTRC will be developed through two phases. The first phase is an 18-month development cycle with a 12-month demonstration deliverable that will utilize existing tools and infrastructure to enable HTRC functionality among partner sites (IU and NCSA). Primary areas of work include the core-cyberinfrastructure and data analysis tool setup, end user services and portal, support center capabilities, and minimal support for derived research data capabilities. In Phase I only the public domain works in the HathiTrust will be utilized, as security and other policies must be developed and put in place.

Phase I implementation includes the bridge and caching strategies between the HathiTrust Repository and indices and the HTRC data store. This will involve the IU Data Capacitor on the HTRC data store side. Distributed interoperability between IU and UIUC will be through Lustre-WAN; but within the 18-month Phase I, we will target testing of the GlusterFS cluster storage file-system because of its security support. Work on the versioning database will also begin. Web capabilities will be developed utilizing functionality such as the popular HubZero. The decision of a portal infrastructure will be made based on an analysis of the relative development and integration costs. This infrastructure will be used in Phase I to begin ingest of data from the HathiTrust IU Node. To complete the core infrastructure, LUSTRE/GlusterFS Client capabilities will be implemented at UIUC to provide distributed access to the initial HTRC services. The CSD threat and risk security analysis will be performed in Phase I and development on the security infrastructure and procedures will get underway.

Phase I staffing for the first year is 1.25 FTE contributed by IU and another 1.25 FTE contributed from UIUC. IU devotes 0.5 FTE to working on the bridge between the HathiTrust Repository and the HTRC data store, and on Lustre testing between IU and UIUC. IU 0.5 FTE is focused on initial design of portal access, supporting use workflows through the underlying infrastructure layers, and on design of the access primitives and compute container in support of the informatics user. Both staff members will work on identifying indexes, providing support for those in the data store, and in helping establish auditing procedures. Stacy Kowalczyk will devote 10% of her time to establishing engagement with users. Robert Ping will devote 15% of his time for project management on the IU side and UIUC will provide 15% time for project management. Plale and McDonald (IU) and Poole and Unsworth (UIUC) will devote time for oversight, for policy and process establishment, for setting up the advisory board, and for engagement with researchers and the outside community. Plale and McDonald will also work with Poole and Unsworth to develop other funding options including writing proposals and developing the sustainability program. Eric Wernert's group at IU will investigate visualization support with extensions to scalable visualization in Phase II. Michael Welge's group at UIUC will devote 0.75 FTE for portal development and mounting SEASR and other tools in the portal, as well as proposal development related to tools. NCSA will provide security for the portal and computing cycles and support on an as-needed basis for estimated .33 FTE. Other listed senior investigators will engage and contribute where their expertise can help. There are no funds provided for this latter purpose.

It is anticipated that by the end of the first year funding will be in place for months 12-18 of Phase I. The planned funding will cover .5 FTE Project Managers at IU and UIUC (1.0 FTE total). These project managers will work with funded projects and support preparation of additional proposals. Two system developers, one at IU and one at UIUC and two 11-month 50% graduate RAs, one at IU and one at UIUC, will support the funded projects. NCSA will design, develop, and deploy the security system for the HTRC. This will include: (1) design, development and deployment of security infrastructure and services which involves a risk/threat security analysis, security architecture design, development of an implementation plan, and documentation; (2) design, development, and deployment of authentication systems which involves an on-line user registration process and establishment of an OTP server; (3) deployment of monitoring systems including IDS, firewalls, and logging; and (4) development and deployment of privacy protection systems for auditing user actions. This will require a 1.0 FTE Senior Security Engineer, 1.33 FTE Security Engineers, and .5 FTE for a security engineer for the OTP for the six month period. IU will house 1.0 FTE to continue and ramp up storage and databases and 1.0 FTE for engagement with external audiences. UIUC will house 1.0 FTE for maintenance of computing infrastructure and docents to assist users of the HTRC and 1.0 FTE for cyberenvironment design, expansion, and maintenance, including adding and maintaining tools. We also anticipate about $50,000 for equipment to expand the reach of the HTRC.

Phase II, expected to begin at the end of the first 18-month cycle, will involve development of an operational research center that will provide ongoing and up to date access to the HTRC research corpus and associated indices. This will involve the development of widely accessed service component for all researchers and the development of enforced auditing controls and review processes that will allow compliance for non-consumptive research of the HathiTrust content. System monitoring capabilities will be developed as part of the auditing component that are open and transparent for use in demonstrating the security and audit compliance that is required by the Google Books Settlement under U.S. law. Deployment of more advanced capabilities, services, and tools that are driven by the domain research partners and the HTRC Advisory Board, and guided by the HTRC research program will also begin in Phase II.

Phase II will additionally focus on long-term sustainability, selection of best and current technology platforms for long-term use and enriched funding capabilities for expansion either from the Google Books settlement or other associated granting agencies that provide funds for this type of research. During Phase II, a "Tools" steering committee will test, evaluate, and determine additional tools to be added to the suite of tools available in the portal. This will include tools and applications from the previous list that have not been mentioned in this section and additional tools brought or developed by users.

Early adopter user-generated derived analysis data will be resident in the HTRC portal and available for re-use within the system and if deemed non-consumptive of the HathiTrust Research Corpus available for use in non-HTRC system. By this time, we anticipate development use of the Meandre environment in a way that is consistent with linguistic analysis research and that can enable the use of new algorithms within the HTRC environment. We expect to develop procedures to allow researchers to add domain, project or methodology specific tools to the infrastructure.

The IU budget in Phase II devotes 25% of Stacy Kowalczyk's time to engagement and training of users. Robert Ping will devote 20% of his time, or 1 day a week, to project management. A graduate student is allocated to investigate a solution to versioning of the HathiTrust collection by means of snapshots taken of the data store at points in time. IU will employ a data collections compliance specialist, and as this is funded at 50% will seek a qualified candidate associated with the IU Library. IU budgets for a 0.5 FTE person to work on aspects of data storage and wide area file system support. A 1.0 FTE will be responsible for developing out the data store, versioning system, access primitives, extensions to incorporate new tools and algorithms, including experimental evaluation on FutureGrid. The 0.6 FTE applications programmer will work with users to facilitate research outcomes. Overall management consisting of tasks outlined above but extended to interaction with legal offices, HathiTrust Executive committee, and Google if necessary are the responsibility of Plale and Poole.

UIUC will budget for 0.5 FTE Project Manager each year. This project manager will work with funded projects and support preparation of additional proposals. UIUC will also employ 1.5 FTE system developer and 1.5 11-month 50% graduate RA to support the funded projects. NCSA will budget for 0.5 FTE for a security engineer who will manage security and vet applications for the HTRC. UIUC will house 1.0 FTE for maintenance of computing infrastructure and docents to assist users of the HTRC and 1.0 FTE for cyberenvironment design, expansion, and maintenance, including adding and maintaining tools. We also anticipate about $30,000 will be devoted to equipment refresh and expansion each year.

References

1. Blacklight URL, http://www.lib.virginia.edu/digital/resndev/blacklight.html
2. Cohen, Dan, Neil Fraistat, Matthew Kirschenbaum, and Tom Scheinfeldt. 2009. *Tools for Data-Driven Scholarship, Final Report*. Turf Valley Resort, Ellicott City, Maryland, 3. http://mith.umd.edu/tools/final-report.html.
3. Dean, Jeffrey and Sanjay Ghemawat 2004. MapReduce: Simplified Data Processing on Large Clusters, OSDI'04: Sixth Symposium on Operating System Design and Implementation.
4. Grossman, Robert and Yunhong Gu 2008. Data Mining Using High Performance Data Clouds: Experimental Studies Using Sector and Sphere, SIGKDD 2008, Las Vegas, NV, Aug. 2008.
5. Jensen, Scott and Beth Plale, Schema-Independent and Schema-Friendly Scientific Metadata Management, *4th IEEE International Conference on eScience*, pp. 428-429, 2008.
6. Llorà, X, B. Ács, L. S. Auvil, B. Capitanu, M. E. Welge, and David E. Goldberg. 2008. Meandre: Semantic-Driven Data-Intensive Flows in the Clouds. In *Proceedings of 4th IEEE International Conference on eScience*, 238-245. IEEE Press.
7. ParaText URL. ParaText: Scalable Text Modeling and Analysis, http://titan.sandia.gov/paratext.html
8. Santry, Douglas S., Michael J. Feeley, Norman C. Hutchinson, Alistair C. Veitch, Ross W. Carton and Jacob Ofir 1999. Deciding when to forget in the Elephant file system, 17th ACM Symp. on Operating System Principles (SOSP), Operating Systems Review 34(5), pp. 110-123.
9. Simmhan, Yogesh, Beth Plale, and Dennis Gannon 2006. A Framework for Collecting Provenance in Data-Centric Scientific Workflows, *IEEE Conference on Web Services (ICWS'06).*
10. *Summit on Digital Tools for the Humanities 2005*. University of Virginia, 28-30 9. http://www.iath.virginia.edu/dtsummit/SummitText.pdf.

11. Titan URL. The Titan Informatics Toolkit, http://titan.sandia.gov/index.html

12. Unsworth et al. 2006. *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences,* American Council of Learned Societies. http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf