Workset Creation for Scholarly Analysis and Data Capsules (WCSA+DC): Laying the foundations for secure computation with copyrighted data in the HathiTrust Research Center, Phase I.

1. Executive Summary

The HathiTrust Digital Library comprises the digitized representations of 13.68 million volumes, 6.84 million book titles, 359,528 serial titles, and 4.79 billion pages. Approximately 39% of the items in the HathiTrust corpus are digital representations of print volumes in the public domain. The remaining 61% are works under copyright. Because of copyright restrictions, scholars have come to see this 61% of the HathiTrust collection of volumes as sitting behind a "copyright wall" that makes it next to impossible for them to have meaningful access to their content.

The HathiTrust Research Center (HTRC) is the research arm of the HathiTrust. The HTRC is a collaboration between the University of Illinois and Indiana University. HTRC has been developing models and tools to help scholars conduct interesting new analyses of works found in the HathiTrust corpus. To maximize accessibility to the entire corpus (regardless of copyright status), the HTRC has been prototyping tools to facilitate large-scale analyses under a "non-consumptive research" paradigm. Under this paradigm, analytic algorithms can be applied to that 61% of the HathiTrust collection that has been blocked off by the copyright wall. Once the analyses are run, only results are returned to researchers. Thus, restricted material is never directly "consumed" by scholars.

The project being proposed here builds upon, extends and integrates two developmental research threads that HTRC has been working on for the past several years aimed at making non-consumptive research using the HT corpus a reality. The first thread originates from work that was conducted in the *Workset Collections for Scholarly Analysis (WCSA): Prototyping Project*, funded by the Andrew W. Mellon Foundation (1 July 2013 - 20 September 2015). The second thread continues the work of the *Data Capsules (DC)* project, previously supported by the Alfred P. Sloan Foundation (2011-2014).

Informally, worksets can be understood to consist of two parts: 1) References to the actual data that is used in a given computational analysis. The actual data could be a whole volume, a given page, an image, or anything other type of possible input; and, 2) Metadata elements that describe the workset itself. This metadata helps in the management of worksets through the research cycle, from their conception, their various stages in the analysis process, their archiving, their citation, all the way to their retrieval and subsequent use by later scholars. HTRC Data Capsules provide the scholar with a virtual machine with two modes: a maintenance mode during which a user can access the network and install software freely, but cannot access copyrighted data; and secure mode where copyrighted texts become accessible to the user while the network access and file system access is highly constrained. (The Data Capsule intentionally drops network access for the virtual machine once the environment is configured to prevent data leakage during data analysis. The running analysis software cannot open network channels and can only access limited, predefined, areas of the storage system to prevent data copying and the loading of malicious code.)

The primary objective of the WCSA+DC project is the seamless integration of the workset model and tools with the Data Capsule framework to provide non-consumptive research access HathiTrust's massive corpus of data objects, securely and at scale, regardless of copyright status. That is, we plan to surmount the copyright wall on behalf of scholars and their students.

Notwithstanding the substantial preliminary work that has been done on both the WCSA and DC fronts, they are both still best characterized as being in the prototyping stages. It is our intention to that this proposed Phase I of the project ***devote an intense two-year burst of effort to move the suite of WCSA and DC prototypes from the realm of proof-of-concept to that of a firmly integrated at-scale deployment.*** We plan to concentrate our requested resources on making sure our systems are as secure and robust at scale as possible.

Phase I will engage four external research partners. Two of the external partners, Kevin Page (Oxford) and Annika Hinze (Waikato) were recipients of WCSA prototyping sub-awards. We are very glad to propose extending and refining aspects of their prototyping work in the context of WCSA+DC. Two other scholars, Ted Underwood (Illinois) and James Pustejovsky (Brandeis) will play critical roles in Phase I as active participants in the development and refinement of the tools and systems from their particular user-scholar perspectives: Underwood, Digital Humanities (DH); Pustejovsky, Computational Linguistics (CL).

The four key outcomes and benefits of the WCSA+DC, Phase I project are:

1. The deployment of a new Workset Builder tool that enhances search and discovery across the entire HTDL by complementing traditional volume-level bibliographic metadata with new metadata derived from a variety of sources at various levels granularity.
2. The creation of Linked Open Data resources to help scholars find, select, integrate and disseminate a wider range of data as part of their scholarly analysis life-cycle.
3. A new Data Capsule framework that integrates worksets, runs at scale, and does both in a secure, non-consumptive, manner.
4. A set of exemplar pre-built Data Capsules that incorporate tools commonly used by both the DH and CL communities that scholars can then customize to their specific needs.

For the two years (especially in the initial eighteen months) of Phase I, WCSA+DC will focus relatively inwards with our main goals being the refining, testing and then deployment of the necessary data, tools, and systems in preparation for real-world use at scale. We hope after we have laid the solid foundation in Phase I that we would then be able to turn more outwards in Phase II to actively engage with scholars and other developers on exploiting, refining, and enhancing the data, tools and systems of HTRC.

Project Leadership
Principal Investigator: J. Stephen Downie (University of Illinois)
Co-Principal Investigator: Beth Plale (Indiana University)
Co-Principal Investigator: Tim Cole (University of Illinois)

Key Research Partners
James Pustejovsky (Brandeis University)
Kevin Page (University of Oxford)
Ted Underwood (University of Illinois)
Annika Hinze (University of Waikato)

Projected Budget: $1,169,893.93 USD
Time Frame: 1 January 2016 to 31 December 2017

2. Proposal Narrative

2.1 The challenge

The HathiTrust Research Center (HTRC)[1] is the research arm of the HathiTrust (HT). The HathiTrust is best described as "...a partnership of major research institutions and libraries working to ensure that the cultural record is preserved and accessible long into the future. There are more than 100 partners in HathiTrust, and membership is open to institutions worldwide."[2] The HathiTrust corpus comprises the digitized representations of 13.68 million total volumes, 6.83 million book titles, 359,528 serial titles, and 4.78 billion pages. Roughly 39% of the items in the HathiTrust corpus are digital representations of print volumes in the public domain. Approximately 61% are digital representations of volumes still in copyright, which is HTRC's greatest challenge: to open the works to computational research while keeping the copyright content secure from misuse.

The HTRC is a collaboration between the University of Illinois and Indiana University. The HTRC is co-directed by Prof. Beth Plale (Professor of Informatics and Computing and Science Director of the Pervasive Technology Institute (PTI) at Indiana) and Prof. J. Stephen Downie (Professor and Associate Dean for Research at the Graduate School of Library and Information Science (GSLIS) at Illinois). At Illinois, HTRC is anchored in GSLIS. At Indiana University, HTRC is aligned with the university research IT organization and anchored in the PTI's Data To Insight Center. Both branches of the HTRC have strong ongoing connections with their respective University Libraries.

HTRC develops software infrastructure, models and tools to help digital humanities (DH) scholars conduct interesting new computational analyses of works of the HathiTrust corpus, with focus on analysis of larger number of works than can be done today (which we call "analysis at scale"). One of the key infrastructure components of HTRC is the Data Capsule (DC). DC is a solution to provisioning secure researcher access directly to the raw data objects of HT Computational analysis of text-based resources is a multi-step process. Data Capsules are further described in section 2.2.1, as is the way in which they are used to meet scholar needs. Scholars will often start by assembling a collection of the texts (or related digital objects such as images) that are of interest to them. This is a discovery process that can include such data cleaning tasks as optical character recognition (OCR) correction or duplicate-item removal, etc. The resulting improved collection is then fed as input to tools that manipulate the improved digital materials. Multiple tools may be used during an investigative process. Tools either come from a researcher who brings their own, or are already hosted at HTRC. Tools can combine, synthesize, extract, transform the text and image data into a form that can be used to create outputs as new text, images or graphs. The intermediary results created by one tool can in turn be used as input another tool so the resulting data can be further analyzed quantitatively or qualitatively. HTRC is guided by a security document, executed 22 July 2015, titled "HathiTrust Research Commons: HTRC Security Measures, Practices, and Policies" (abbreviated HTRC S:MPP hereafter).

The HTRC has come to call a collection of digital items brought together by a scholar for her or his computational analyses a "workset." We will discuss the notion of worksets in more detail later (Sections 2.2 and 2.3) but the workset is critical for HTRC: since the HT data cannot leave the secure commons, the workset is a transportable object that bring a scholar's research context

[1] http://www.hathitrust.org/htrc

[2] http://www.hathitrust.org/about

to the data. Our conception of worksets, their properties, manifestations, uses and implications, evolved from the Mellon-funded Workset Creation for Scholarly Analysis (WCSA): Prototyping Project conducted by HTRC from October 2013 to October 2015. WCSA undertook three main tasks:

1. User needs analyses to ascertain the current and future requirements of scholars in building and using worksets to advance their scholarship;
2. Formal modeling of worksets and their description to allow the workset to function as a persistent and custom context wherein scholars bring together their desired data with their particular computational tools; and,
3. Create a small set of prototype tools and systems, via a competitive sub-award process, that advances our understanding of worksets by demonstrating some of their potential strengths and weaknesses in meeting the needs of scholars.

2.1.1 Scholar needs

As part of WCSA's user needs analysis task, HTRC (Fenlon et al., 2014) and its WCSA prototyping partner sub-award group at Oxford (Page & Wilcox, 2015) held a series of focus groups in a variety of community venues. Our collective findings confirm that scholars have a "Wish List" that expresses their desire to:

1. Gather into a single context materials that exist both in the HT corpus and from outside sources;
2. Locate and define materials at a finer level of granularity than the volume (e.g., page, paragraph, word, letter, etc.);
3. Discover HT materials through one or more types of non-traditional metadata (e.g., machine generated genre tags, language tags, concept tags, part-of-speech annotations, extracted features, etc.);
4. Locate, analyze, and triangulate (that is, carry out research) directly on non-traditional metadata;
5. Represent and manipulate their worksets using Linked Open Data methods as proposed by the WCSA formal model (Jett, 2015), and to interpret worksets according to domain models appropriate to their investigation (Nurmikko-Fuller et al., 2015);
6. Share, publish and reuse worksets as well as intermediate analytic outputs;
7. Run their analyses against their worksets using "hand-crafted" programs derived from a wide range of popular packages that are customized to their specific research questions;
8. Use workset building and investigation tools that are customized to the domain of investigation, including views and linked materials specialized to the discipline; and,
9. Do all of the above at a scale (i.e., the number of books exceeding a few thousand) regardless of the copyright status of the source materials.

The final wish list item (#9) encapsulates simultaneously the unique opportunities created by HTRC and its greatest challenges. HTRC is uniquely positioned to bring scholars and their tools in contact with corpus of material of nearly 14 million volumes, 61% of which are under copyright. Thus we have the promise of both scale and access to copyrighted materials. To realize this promise, however, HTRC must yet meet the challenge of making large- and small-scale analyses of the HT corpus a matter of routine. Furthermore, HTRC must provide both large- and small-scale analytic opportunity at all times with absolute respect for copyright laws.

It is obvious that this makes meaningful scholarly access to the 10 million volumes under copyright rather problematic.

To surmount the conflicting challenge of affording useful analytic access to works that cannot be normally shared, the HTRC has adopted a “non-consumptive research” paradigm. The permissibility to do automated analysis on copyrighted digitized texts from university libraries has been hotly debated ever since the Authors Guild’s class action lawsuit in 2008.[3] A number of stakeholders have argued that data and text mining should be permitted, drawing on the principle of "non-expressive" use, that is, uses that do not trade on the underlying or expressive purpose of the work, meaning that text mining represents a new use and does not gain monetarily or otherwise from the original work.[4]

HTRC, with backing of HT and its university legal counsels, enables “non-expressive use" of the HT corpus. We operationalize this through the following definition of “non-consumptive" use: “no computational action or set of actions on part of users, either acting alone or in cooperation with other users over duration of one or multiple sessions can result in sufficient information gathered from the HT repository to reassemble pages from the collection for reading.” That is, we permit computational analysis of the copyrighted content but no user's text mining actions can, for instance, leak even a full page of a book to the Internet.

The time has come for HTRC to resolve the difficulties put before it to successfully facilitate the non-consumptive analyses of copyright-restricted materials. Until this is accomplished, potentially important scholarship will continue to be impeded by the notorious “1923 copyright wall.” The 1923 date denotes the commonly used shorthand to signify works in the public domain (published before 1923) and works still under copyright (published in 1923 or after). For example, Ted Underwood, an active and respected DH scholar has been using the public domain HT corpus to develop a model of character types that has begun to provide some interesting clues about changing representations of gender. Sadly, it is not safe or accurate to extrapolate past 1923, so the project has been suspended until such time as Underwood and his colleagues can leap over the 1923 wall. More information about Underwood’s project can be found in Section 2.3, Task 6. We have asked Underwood to collaborate with HTRC in breaking down the 1923 copyright wall using his research as one of our real-world, large-scale exemplars.

2.1.2 Summary of challenges

Deploying a facility in which the non-consumptive scholarly analysis of a corpus the size of the HathiTrust is possible poses no small number of challenges. The Workset Creation for Scholarly Analysis + Data Capsules (WCSA+DC) project is designed to advance a range of these inter-related challenges. A summary of the challenges and the questions they raise include:

1. DH scholars benefit today from a surfeit of existing, vetted research tools for cleaning, manipulating and analyzing digitized text,[5] but when digital material cannot be removed from its secure environment due to sensitivities such as copyright, computational analysis must come to the data instead of vice versa. That is, under the non-consumptive paradigm, computation must necessarily occur close to the data, which creates a somewhat artificial work situation for scholars. *How do we lower barriers to access?*

[3] http://en.wikipedia.org/wiki/Authors_Guild_v._Google

[4] http://www.alrc.gov.au/publications/8-non-consumptive-use/text-and-data-mining

[5] See, for example, the breadth and depth of DH tools described in the DiRT Directory of digital research tools for scholarly use, http://dirtdirectory.org/tadirah and http://dirtdirectory.org/.

Specifically, how do we represent and maintain a scholar's research context (their "workset" and tools) at the HTRC from the moment of inception of the research, through to its publication as a curated object in such a way that that helps scholars be maximally productive?

2. For many DH scholars the size of the HT corpus is both attractive and daunting. Many existing DH tools are designed to work on smaller collections of text, and many research inquiries are facilitated by the availability of more focused, homogeneous collections of texts (i.e., focused in respects relevant to the research inquiry) (Gibbs & Owens, 2012). This suggests two complementary research questions, and we will be guided by scholars in answering these questions: *How do we provide working environments and services that allow scholars to apply to larger collections the established tools and algorithms with which they are familiar (i.e., how do we help make sure HTRC can help them scale up their tools, their worksets and their next-generation research questions)? Similarly, how do we enable users to define and manage computational access to distinguished subsets of the HathiTrust corpus so as to facilitate the use of existing tools and algorithms that are practically limited in terms of scale and the construction of finer-grained worksets relevant to specific scholarly inquiries?*
3. Many text-mining tools must be trained on a representative subset of the full collection or corpus to be analyzed. *Is there a model upon which training datasets of in-copyright materials can be developed where the training dataset is not itself subject to the same restrictions as the copyrighted materials?*
4. When the size of a corpus over which analysis needs to be carried out exceeds a few thousand volumes, high performance computing (HPC) resources are needed.[6] The current security guarantees of the Data Capsule model hold for only a single virtual machine (VM). *How do we ensure safety of the copyright-restricted text and image data as it is used in analysis tools that require HPC resources? Can the threat model defined by Data Capsules be extended to analysis that requires HPC resources without loss of strength of the threat model?*
5. Text-based research requires considerable interaction between the researcher and his/her identified texts, through tagging, annotations, through derived results that are then combined with other texts (Fenlon et al., 2014, Page & Wilcox, 2015). *How do we enable deep scholar interaction with texts without compromising the safety of the texts (i.e., violating the non-consumptive research paradigm)? How do we securely manage (as appropriate to sensitivity) the intermediate results of research interactions that will be needed by other scholars, by succeeding (downstream) tools, and/or to enable the subsequent creation of derivative inquiry-specific worksets?*
6. The HT corpus contains multiple exemplars of many works and even multiple digital copies of specific editions of some works. The HT corpus includes works and editions of works included in other collections (e.g., the Early English Books Online Text Creation Partnership (EEBO-TCP) and Eighteenth Century Collections Online collections), which encompass alternative, but complementary, motivations and strategies towards curation and access. *How do we provide metadata and services to help researchers select the 'best' copy (for their inquiry) and exclude duplicative instances from their worksets for scholarly analysis? How do we leverage the strengths of complementary corpora to aid investigation of HT and vice versa?*

[6] Prior experience has shown us that a linear walk through of 1 million digitized books takes 1024 processors and 22 hours (shorthand: 1M books:1K computers:1 day).

7. As noted above, The HT corpus, while a phenomenal resource, is richer when scholars can enhance their analysis of its data by referencing, consulting, combining, and comparing with data drawn from other sources. *How do we extend the Data Capsule to allow data external to HT to be made available for computation just-in-time during a researcher's investigation without compromising the safety of the in-copyright HT materials (i.e., not allowing access to external resources to compromise non-consumptive research paradigm)?*

2.2 The Context

2.2.1 HTRC Secure Commons

The ***HTRC Secure Commons*** is the software infrastructure, data assets, computers, and storage systems that are assembled to support and ensure that researchers are able to safely carry out "non-consumptive" analysis of the sensitive text and image data of HathiTrust. At the highest level, the Secure Commons is a set of resources within a logical ring of services and computers that effectively surround and protect the sensitive data. This "circling of the wagons" effect enlists trusted software services, networks, and computers to protect the sensitive data within. The services and computers that form the layer of protection are themselves both trusted and trusted to enforce protections on the data. We refer to this set of trusted services and computers as residing within a Trust Ring that includes services, data assets, computers, networks, and HPC resources residing at both Indiana University and University of Illinois.

The architecture of the HTRC Secure Commons, as outlined in Figure 1, is made up of data products and services, knowledge products and services, tools for discovery and analysis, Data Capsule VMs, and an overarching suite of services for security and identity management. At the lowest level is the HT data itself and indexes to it. These are managed by data management services (e.g., mySQL, Cassandra noSQL store, Solr, etc.). The external data cache is a new design feature anticipated for just-in-time ingest of data from external sources as part of this proposal. Above the data management services are knowledge management services. These include data derived from the HT corpus, additional metadata, ontological information, worksets, etc. We distinguish the private workset from the published workset, a recognition that worksets can be shared. A workset is maintained in private mode for the individual scholar while being developed and for as long as the individual scholar needs access to that workset. A published workset—which is meant to be shared—has been through a curation, packaging, and licensing process and made accessible for open and public access and use (Plale et al. 2015). Above the data, knowledge products, and their respective management services are two categories of services. On the left are the discovery, analysis, cleaning, etc. tools that HTRC hosts. A researcher interacts with these tools currently through a single web interface. On the right is Data Capsules. Data Capsule, described in more detail below, through work proposed here will offer custom VMs for communities. Shown in the figure are VMs customized for Digital Humanities, Computer Science, a generic R VM, and Natural Language Processing. Security, identity management, and auditing services are cross-cutting.

We take a moment to point out that in HTRC release 2.0, the services, tools and data are accessed through a single web portal. While this solution makes security and identity management easier to do, it restricts the interfaces that a researcher can see. In other words, it throws the researcher into a new and what could be perceived as limited environment in which to work. Feedback from researchers strongly suggests HTRC needs to do more. The alternative to a single portal is a "apps" architecture where users access HTRC tools directly from their laptop or mobile device. We will pursue a more flexible architecture in WCSA+DC, with keen attention to maintaining the right balance between security, i.e. keeping the texts safe through rigorous security measures, and creating a high level of usability. This alternative is now being realized, in part, via the Data Capsule, currently under development, and described in the next section.

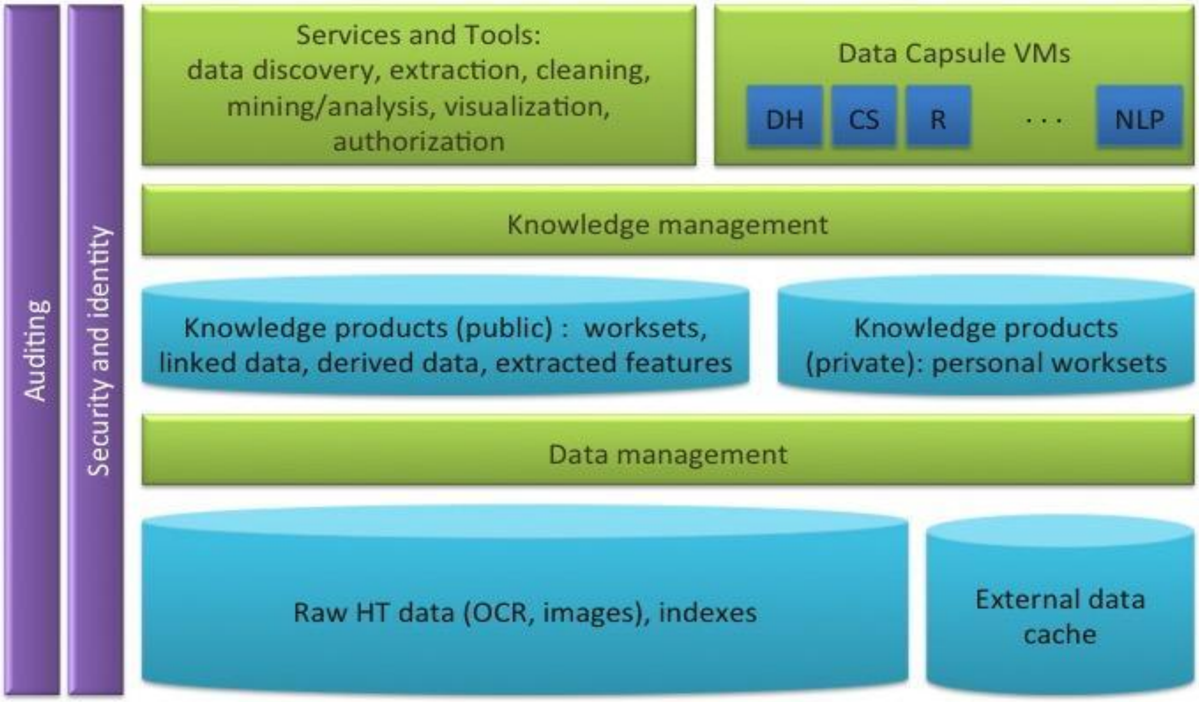
Auditing
Security and identity
Services and Tools:
data discovery, extraction, cleaning,
mining/analysis, visualization,
authorization
Data Capsule VMs
DH
CS
R
...
NLP
Knowledge management
Knowledge products (public) : worksets,
linked data, derived data, extracted features
Knowledge products
(private): personal worksets
Data management
Raw HT data (OCR, images), indexes
External data
cache

Figure 1. HTRC Secure Commons architectural components

2.2.3 Data Capsules

The HTRC Data Capsule (Zeng et al., 2014) is a solution to provisioning secure researcher access directly to the raw data objects of HT. HTRC Data Capsules are the culmination of a Sloan grant.[7] Data Capsules employ a generic running, Unix-based system built on the assumption of trust in the user—i.e., a text mining researcher is trusted to not deliberately leak repository data—reinforced with a signed formal user agreement, as discussed below on page 11. The user agreement can be found in Appendix B of the HTRC S:MPP. We also include safeguards to prevent malware acting on the user's behalf from leaking data. DCs are motivated by four constraints:

1. Non-consumptive use: can the framework provide safe handling of large volumes of protected data that undergo computational analysis?
2. Openness: can the framework support user-contributed analysis tools (that is, not limit uses to a known set of algorithms)?
3. Efficiency: can the framework support user-contributed analysis tools without resorting to code walkthroughs prior to acceptance?

[7] Final report: http://hdl.handle.net/2022/19277

4. Large-scale and low cost: can the protections be extended to utilization of large-scale national (public) computational resources?

The Data Capsule's design draws on the data capsules approach (Borders et al., 2009) to preventing misuse. Using a remote Virtual Machine (VM) model, researchers can build a VM configured with software and tooling based on their needs. Once an analysis is finished, the VM is wiped out and resources released for other users to share. VMs offer no inherent protection, however. Our approach extends the virtual machine by turning it into a "data capsule" (Borders et al., 2009) that prevents leakage of copyrighted content in the event that the VM is compromised or data analysis routines malicious.

The Data Capsule is the only form of researcher-controlled, direct access to the HathiTrust copyrighted content for the foreseeable future. It is a target environment for providing secure access to restricted datasets—where it is desirable to do so—to end-users who are generally trusted with remote access to the dataset. End-users can bring in their own software for analysis and bring in external datasets. The security risk that the system mitigates is that of software that is used for analysis being inadvertently malicious and thus leaking data to a third party (Plale et al., 2015, Zeng et al., 2014).

HTRC DC currently limits a researcher to a single VM. Under this project, HTRC DC will be extended to allow a researcher working within a Data Capsule to invoke HPC resources to run large-scale computational analysis tasks. HTRC Data Capsule was designed and prototyped through a 3-year grant from the Alfred P. Sloan Foundation that ended December 2014. The purpose of the Data Capsule grant was to prototype an environment where non-consumptive interaction with sensitive data could go on while at the same time the constraints of non-consumptive access are not violated.

Before gaining access to the HTRC DC, a researcher must agree to abide terms of the HTRC DC Service Agreement, as well as the standard HTRC user agreement. The Service Agreement will define appropriate use of HT data, emphasizing the non-consumptive use of in-copyright data. As to ***identity*** for researchers, users select their own username and password. Their identity is managed by an identity server (an instance of the WSO2 Identity Server[8]).[9] As to ***role-based access*** for the researchers, a researcher is assigned the role of HTRC Services User. This role is assigned to any person who is a member of the higher education community, who has a registered identity with HTRC Services, and who has agreed to the HTRC Services User Access and Use Agreement (see Appendix B of HTRC S:MPP). As to ***account approval***, as per Sec 4.2 of HTRC S:MPP, to obtain an HTRC Services User account, a person signs up for an account through HTRC Services. An account application must satisfy all of the following criteria:

1. Has an email address verified to be from an institution of higher education or a research library;
2. Acknowledges user responsibilities when using the HTRC Services (see HTRC S:MPP, Appendix B), which include restrictions on access to copyrighted materials;
3. Provides a reasonable purpose for using the account; and,
4. Responds to an email verification sent to their institution email address (2-step verification)

[8] http://wso2.com/products/identity-server/

[9] We are interested in moving away from running our own identity server in favor of a federated identity solution that will simplify identity management.

If all criteria are satisfied, the account is approved without human contact with the applicant. Otherwise HTRC Staff engage in a dialog with the applicant and make a determination.

The HTRC Data Capsules provide the virtual machine with two modes: a maintenance mode during which a user can access the network and install software freely, but cannot access copyrighted data; and secure mode where copyrighted texts become accessible to the user while the network access and file system access is highly constrained. In the latter mode, users are allowed access only to a predefined set of network addresses and write to a specific volume, which is only visible in secure mode. Any other change made to the system in secure mode, except for the ones made to the special volume, are lost when the mode is switched from secure to maintenance. This is to guard against the situation that copyrighted texts are saved in the VM in secure mode and copied out across the network during maintenance mode.

The prototype for non-consumptive, computational access to a restricted full-text corpus implements the following threat model:

1. Users access restricted data through remotely accessed VMs that read data from a network-accessed data service.
2. The VM that is given to the user for their use is not part of the trusted computing base. Keeping the VM outside the trusted computing base allows the user the freedom to install their own software on the VM. The remaining support is within the trusted computing base: the Virtual Machine Manager (VMM), the host that the VMM runs on, and the system services that enforce network and data access policies for the virtual machines. The HTRC data services themselves are also part of the trusted computing base.
3. We assume the possibility of malware (i.e., malicious software) being installed as well as other remotely initiated attacks on the VM. These attacks could potentially compromise the entire operating system and install a rootkit, both of which are undetectable to the end user.
4. The end users themselves are considered to act in good faith, but this does not preclude the possibility of them unwittingly allowing the system to be compromised. This is a reasonable assumption as all users are required to sign the HTRC user agreement before being allocated a VM and engaging with Data Capsule services. The user agreement is included in Appendix B of the HTRC S:MPP (finalized July, 2015). The document has been shared under separate cover.
5. Users access their Data Capsule through a Virtual Network Computing (VNC)[10] connection, giving them remote desktop access to their capsule. A user logs into a VNC session with a personal password then logs in for a second time to their Data Capsule VM (currently only Ubuntu Linux Data Capsules are supported). We are aware of the security risks in using an unencrypted VNC connection, most notably sniffing on the channel. At present, a user is trusted—i.e., we can trust that copyright content will not flow on the channel—however, this does not preclude sniffing of passwords. Addressing this vulnerability is part of our proposed effort. Employing a VNC capable of supporting encryption is our proposed fix to this vulnerability.
6. Unauthorized access of a VM is detected through the regular monitoring of active VMs. Unusual behavior , such as excessive communication port creation or usage, or excessive access to the HTRC data stores, is brought to the attention of the operations manager.

[10] See https://en.wikipedia.org/wiki/Virtual_Network_Computing for an introduction to VNC.

7. Research results are released upon review when research is complete; review includes a human review of these research results. This review is done by an HTRC staff member who is unknown to the researcher. Once this review is done, the user receives a URL in their email inbox where they can retrieve the results. In the future, released data could be encrypted and undergo automated review to detect potential abuses. Inadvertent release of results via user's email inbox requires that malware compromises exist to a user's account. This is unlikely for users who use their institution emails, which the HTRC requires. Since the user is assumed to be benign in our threat model, users are likely to report an unexpected release of result files from the system.
8. The creation and exploitation of covert channels between VMs that run on the same VM host machine are a known potential threat. For instance, a VM running in secure mode could possibly make use of such covert channels to leak data to a co-resident VM running in maintenance mode, which can in turn leak the data anywhere it pleases. We currently have a prototype solution to address the solution–it requires using two physically separated systems, one that only runs VMs in secure mode and another that runs VMs only in maintenance mode.

2.2.4 Carrying Out Computational Analysis in HTRC

Computational analysis of text and images by scholars in the digital humanities, informatics, and computational linguistics (CL) is often a complex process that involves assembling the right set of texts, running some analysis tools on the data, examining the results, running some more analysis, and so on. It is a nuanced and individual process. There is no one universal set of analysis tools and process that works for all scholars, nor often is there even a universal set that works within a discipline.

HTRC continues to learn, grow, and refine how it supports computational analysis within the Commons, and this proposal represents a next step in that maturity. HTRC began in 2011 by offering Meandre[11] (Llora et al., 2008) and the extensive suite of analysis tools built into SEASR.[12] While the SEASR analysis tools see up to 100 uses a month, they are limited in the size of a corpus over which they can run efficiently, and tend to target a narrow user group. More advanced users need access to scholar-built or scholar-customized data analysis tools, and tools that run at larger-scale than 1,000 volumes.

In response, we are currently prototyping a simple, lightweight chaining framework for analysis based on accepted phases of data analysis (Bell 2009):

1. Data extraction: data are gathered from multiple sources (databases, linked data sources, external data sources);
2. Data integration/cleaning: data are subject to harmonization, cleaning, integration, markup;
3. Data analysis: analysis algorithms are run in a training mode on a subset of the data, or are run on the integrated data;
4. Results viewing: results are viewed as text, as visualizations, graphs.

[11] http://www.seasr.org/meandre/

[12] http://www.seasr.org/

A researcher should be able to pick and run tools from each category. The chaining framework is lightweight, allowing a researcher to run only data integration/markup without having to do any other steps, for instance. Workflow systems tend to run start to finish through a set of tasks.

Tools are organized by the phase into which they best fit. This allows a scholar to contribute and utilize tools based on category. They accomplish work by assembling one tool from each category (or subset of categories); the tools then execute in a known order. The chaining framework supports complex executions: partial executions, repeated executions, and breaks of days or weeks between stages.

In the chaining framework, user-contributed data analysis tools come from multiple sources: off the shelf, contributed by community, or common open-source tools. Occasionally but rarely these tools will be developed in-house. That is, HTRC focuses its development effort on the chaining framework and information infrastructure needed to run the tools on behalf of the users.

2.2.5 Rationale for tools selected for support in the Commons

We address the question of rationale for tool selection for use in the Commons, but first explain the multifaceted nature of the term "tools" in the context of HTRC. Tools within HTRC are used in computational analysis or for broader support of research and the infrastructure. Tools vary within HTRC on their targeted scale and on our need to accommodate user-owned tools are not part of any selection process:

1. Analysis tools: tools for data mining, machine learning, text analysis, data extraction, data, data integration/markup/cleaning, and visualization.
2. Support tools: Tools that HTRC develops or uses in house in support of the data analysis tools and activities of the researcher. For example, a discovery tool for workset building, a linked data store, use of the Resource Description Framework (RDF) as a representational language, the tools that execute the data mining and analysis tools, and ferry results around.
3. Tool scale: tools separate out based on the size of the dataset the tool can handle: tools that run over <1,000 volumes are not those that run over 1M volumes. This is because the latter tools are specially programmed to run simultaneously on thousands of processors on HPC resources.
4. Level of support: tools can be brought to HTRC by a researcher for their own use. HTRC strives to support this mode of tool use as fully as possible by not constraining tool choice.

Our rationale for selecting tools for broader community use reflects the tool distinctions made above. Through early support for Meandre (Acs et al., 2011) and the SEASR suite of analysis tools, HTRC got up and going quickly. It needs to move in the direction of tools with greater variety of purpose, and tools that handle scale better. Our rationale is largely user-driven: first, we have chosen two distinguished and visible researchers in two areas (Underwood in DH and Pustejovsky in CL) to work with us on this project to create builds of HTRC Data Capsules that have community-specific tools that come pre-configured and help with adoption and spread of their use in their broader communities. These researchers know the tools in popular use in their communities and can guide us. For instance, there are common tool suites that have large uptake in communities such as scikit-learn (scikit-learn.org) for machine learning in Python. Underwood uses scikit-learn extensively and reports heavy use of this tool by DH colleagues. We will be working with Underwood and Pustejovsky to ensure that amongst the experimentation that we do, the data analysis tools selected to include will be both from those that execute over small scale data sets (<1000 volumes) and those that execute over large-scale data sets (>1,000,000 volumes). Second, we have selected an Advisory Board carefully to include a number who can represent the desires of their disciplinary communities as well. Finally, we are also developing a set of generalizable tools at HTRC that cross-cut disciplinary boundaries including our HT+Bookworm[13] n-gram searching and visualization tool (Lieberman et al., 2007; Michel et al., 2011; Auvil et al., 2015), our linked data metadata services, the HTRC Workset Builder[14] and the release of non-consumptive, extracted features datasets[15] (Capitanu et al., 2015).

2.2.6 Concerning Worksets

In many, if not most, DH research endeavors, performing a complex analytical task across the whole of the HathiTrust corpus is neither practical nor productive (Kambatla et al., 2014). For example, it would be wasteful of resources and generally not profitable to apply a tool designed and trained to identify the genre attributes of 18th century English language prose fiction to volumes containing primarily 20th century French poetry. Often the first step in a DH research inquiry is to identify the subset of materials–works, editions, volumes, chapters, pages–which are to be the fodder for the inquiry. In a corpus as large and complex as the HTDL, the actual finding of the materials and then their defining as the sought-after subset (in a form amenable to computational analysis) can be, in fact, extraordinarily difficult. It was this difficulty that motivated our the WCSA project and its investigations of worksets.

Over the course of the WCSA project, the HTRC worked with the Center for Informatics Research in Science and Scholarship (CIRSS) at Illinois. Together, after reflecting upon input and advice from the DH community, they evolved a definition of a workset as a machine-actionable research collection that has these facets (Jett, 2015):

1. An aggregation of members (e.g., volumes, pages, etc.);
2. Metadata intrinsic to the workset's essential nature (e.g., creator, selection criteria, etc.);
3. Metadata intrinsic to digital architectures (i.e., creation date & number of members in the workset);
4. Metadata supportive of human interactions (i.e., title & description);
5. Metadata derived from the metadata of the workset's members (e.g., format(s), language(s), etc.); and,
6. Metadata concerning workset provenance (e.g., derived from, used by, etc.)

Informally, worksets can be understood to consist of two parts:

1. References to the actual data that is used in a given computational analysis. Item #1 represents this part. The actual data could be a whole volume, a given page, an image, or anything other type of possible input; and,

[13] http://bookworm.htrc.illinois.edu/develop

[14] https://sharc.hathitrust.org/blacklight

[15] Non-consumptive research will be easiest when only a few parts of an analysis pipeline require direct manipulation of original texts in a secure environment. For instance, many questions can be answered simply by counting words and other low-level features of text. Such features can be extracted from copyrighted materials and provided to scholars as they are not themselves restricted by copyright. HTRC has already developed a process to extract various features securely (Capitanu, et al., 2015). See https://sharc.hathitrust.org/features and https://sharc.hathitrust.org/genre for sample extracted feature datasets.

2. Metadata elements that describe the workset itself. Items #2-6 represent this part. This metadata helps in the management of worksets all the way through the research cycle from, their conception, their various stages in the analysis process, their archiving, their citation, to their retrieval and subsequent use by later scholars.

The special relationship between a reference to a data object as mentioned above and the actual data it represents is fundamental to the notion of the workset in a non-consumptive environment like HTRC. To create compatibility with the non-consumptive research paradigm, the WCSA team deliberately developed a formal RDF-based model (see Figure 2) which broadly scopes the workset as a graph that describes research collections that are intended and designed to work wholly within the HTRC Secure Commons such as the one developed by the HTRC (Jett, 2015).

From a traditional web perspective each workset is thus an aggregation of identifiers that name entities that a DH scholar has identified as being of interest. From a semantic web perspective, each workset is an aggregation of named entities which can only be reified within the "walled-garden" context of a Trust Ring and, in the specific case of the HTRC and WCSA, that Trust Ring is the HTRC's Trust Ring. Because references to objects themselves are not restricted information, DH scholars can still export their worksets outside of the HTRC Secure Commons as the worksets are little more than a collection of links and a description of the workset itself; from a scholarly perspective this also means that the workset has an existence outside and independent of the Secure Commons, as a citable and archivable object. Furthermore, because worksets consist of these standardized references, it is a design feature of the WCSA workset model that scholars can create and manipulate worksets using future tools and resources developed and/or hosted outside of the HTRC environment; when such worksets pass back into the Secure Commons the references to the protected resources become immediately resolvable to the data itself.

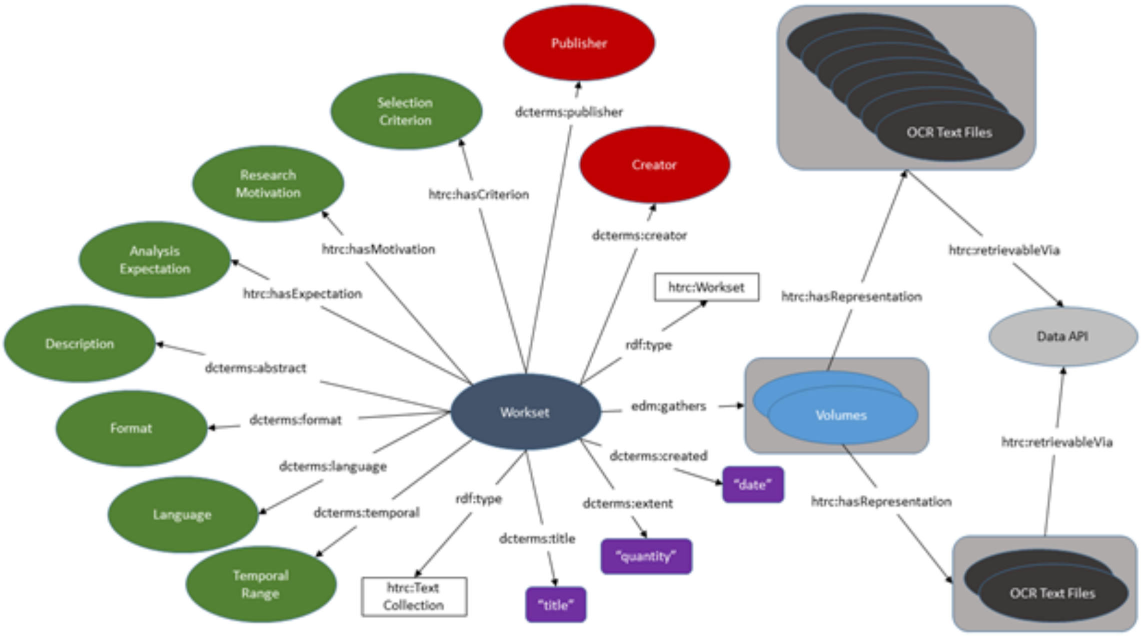

Figure 2: Combined Workset & Bibliographic Resource Data Models[16]

[16] Note that the object of edm:gathers—Volumes—is meant as an illustrative example. The data model supports the aggregation of a much broader variety of member objects. It does this by remaining agnostic with regards to *what* is being gathered.

In addition to the formal workset model, the CIRSS technical report on worksets, *Modeling Worksets in the HathiTrust Research Center*,[17] published by the WCSA team describing the model also outlined a set of recommendations (Jett, 2015). The report recommended the following set of actions be taken to realize the model in the HTRC environment. The set includes:

1. Implement the basic Workset and Bibliographic Resource models described in Section 3 of the report through a new Workset Builder[18] infrastructure.
2. Develop workflows to leverage existing HTRC MAchine-Readable Cataloging (MARC) metadata for volumes to better empower scholars to select resources for their Worksets.
3. Implement identity metadata for bibliographic granules. (Page-level relatively easy to implement, finer and more arbitrary granules will require additional development cycles.)
4. Develop and implement descriptive metadata for bibliographic granules. (Page-level relatively challenging. Other granule levels will require additional development cycles.)
5. Develop and implement means of differentiating abstract levels of content from one another. (Of relatively moderate difficulty at the Page-level. Complicated by indirection and notions like "proxies" which lead to misuse of metadata records acting in the role of avatars representing other entities.)
6. Develop and implement provenance metadata at all levels, possibly taking advantage of W3C PROV suite of standards and tools.[19] (Unless a provenance method that relies solely on infrastructure is instead identified.)

As part of its initial activities, WCSA engaged HTRC and its sub-award collaborators at Oxford University, the University of Maryland, Texas A&M University, and the University of Waikato in designing and testing prototype and proof-of-concept tools facilitating the creation of worksets suitable for computational scholarship at scale. The outcomes of the WCSA project directly inform our work on this new proposed project and provide a foundation for moving models and methods developed during the WCSA project *from prototype to production*. In describing the context for the current proposal, four outcomes are worth highlighting here (in addition to the focus group feedback gathered as part of the WCSA project and mentioned above in the scholars' wish list):

1. A RDF-compatible[20] data model (Jett et al., 2015; Jett, 2015) for describing scholarly worksets containing HathiTrust resources only or in combination with resources from other sources.
2. A model and method for creating Linked Data resources - exemplified in the creation of an RDF encoding from the University of Oxford's Early English Books Online Text Creation Partnership (EEBO-TCP[21]) resource - that is both complementary and compatible with the WCSA data model of the previous point, and thus can be used in concert.

[17] http://hdl.handle.net/2142/78149

[18] The Workset Builder mentioned here refers to HTRC's full-text bibliographic search tool found at https://sharc.hathitrust.org/blacklight. It is the currently the primary means by which scholars have been prototyping the finding of their materials and then the building of their worksets.

[19] http://www.w3.org/TR/prov-overview/

[20] http://www.w3.org/RDF/.

[21] http://www.bodleian.ox.ac.uk/eebotcp/

3. A study (Nurmikko-Fuller et al., 2015) of the challenges of aligning competing bibliographic metadata ontologies used by large digital libraries today, which currently form the foundations for HT and workset metadata; and the means by which these might be mapped to ontological structures aligned to specific scholarly investigations (for example to event-based or descriptive models) and incorporating non-bibliographic attributes and properties.
4. A proof-of-concept demonstration of how the Capisco System,[22] a tool for doing context-sensitive semantic analyses of full-text documents, making use of a Concept-in-Context (CiC) network seeded by *a priori* analysis of Wikipedia texts and identification of semantic metadata, can help better pinpoint and disambiguate the topical coverage of texts, thereby supporting the more precise building of scholarly worksets (Hinze et al., 2015).[23]
5. The ElEPHãT proof-of-concept tools[24] demonstrating the feasibility of presenting a custom workset builder and viewer tailored to include resources for study of early English literature, including both HT and EEBO-TCP volumes, while reusing common WCSA and compatible models (#1, #2) built upon Linked Data technologies (RDF and SPARQL).

These and other outcomes of the WCSA project have demonstrated the viability and potential of augmenting traditional bibliographic metadata with other attributes (including attributes developed through analysis of the full-text of a resource). These added attributes and resulting enhanced descriptions can help scholars to discover and identify more resources pertinent to their research inquiries. The challenge over the course of the new project described in this proposal will be to move from prototype to production. This will require integrating WCSA prototype implementations into HTRC production workflows. For example, in the WCSA project we demonstrated that the semantics-aware Capisco System (developed by the University of Waikato) can be used to recognize and especially to disambiguate (better than methods relying solely on simple lexical analysis alone) the topical coverage of selected texts in the HathiTrust both at the page level and (potentially at least, we believe) at the volume level. The challenges now are to establish the limits and applicability of Capisco vis-à-vis the full HT corpus and to redesign the HTRC Workset Builder application and underlying HTRC metadata and indexing infrastructure to accommodate and take full advantage of enhanced descriptions incorporating output from tools like Capisco.

As recipients of a WCSA sub-award, the Waikato team developed an innovative set of algorithms to identify semantic concepts in full-text documents, enabling targeted scholarly search. Using these algorithms as a semantic keyword tagger on the HT full-text corpus creates quality metadata to enhance existing bibliographic data of both in-copyright and public domain documents. These tags can be used to provide consumable (i.e., shareable) fine-grained (i.e., at the page level) descriptions of in-copyright materials that does not violate the HTRC's non-consumptive research paradigm. The "Concepts-in-Context" semantic tagger Capisco builds on the team's expertise with WikipediaMiner (Milne & Witten, 2013), an internationally recognized

[22] https://www.youtube.com/watch?v=2LiW_4X_6iU

[23] http://dl.acm.org/citation.cfm?id=2756920

[24] Links to screencasts of ElEPHãT creator and viewers:
https://drive.google.com/file/d/0B_9mpa6jEOQAWnVKXzdlZktNeVU/view?usp=sharing
https://drive.google.com/file/d/0B_9mpa6jEOQAckt2T2JPNi1KWFk/view?usp=sharing
https://drive.google.com/file/d/0B_9mpa6jEOQAVTNvWml6TVFiczQ/view?usp=sharing

keyword extractor using machine learning developed at Waikato. It analyses the Wikipedia link structure and the context and in which keywords for a concept occur. Capisco considers each document's context when tagging semantic concepts, and can be easily targeted for field-specific terminology. Independent evaluations have shown WikipediaMiner to deliver superior semantic accuracy in comparison to both semantic annotation tools and other keyword identification software (Jean-Louis et al., 2014). The WCSA sub-project found Capisco to yield higher quality semantic tagging than WikipediaMiner and a superior performance with the promise of excellent processing capability at scale. The accuracy of the Capisco tagging results was evaluated and compared for known subsets to results of other algorithms and in comparison to manual tagging (Hinze et al., 2015). Processing time was reduced from 12 days to a few hours for a test corpus executed in a prototype data capsule (speedup factor >50). The strength of Capisco is its support of field-specific terminology and openness to scholarly specification that goes beyond the original WikipediaMiner approach based on machine learning. Tailoring the concepts-in-context network enables high quality results for in-copyright sources and is vital to the long-term success of augmenting bibliographic metadata.

Fine-grained targeting of specific semantic areas is supported in Capisco through a feedback loop of scholarly adjustments in the Concept-in-Context network. These adjustments ensure that relevant semantic concepts are well represented and interconnected with their contexts. Connections between concepts can be inserted, prioritised or discarded for a given context. It is worth noting that scholarly adjustments to the CiC network do not depend on access to in-copyright material. Detection of missing concepts and network links is crucial for coverage of a scholar's nomenclature; this can be done via the search interface or via Capisco's CiC network editor. Prioritising and suppression of concepts or links is necessary if appropriate concepts exist in the CiC network but are not correctly identified. These cases can be triggered via three mechanisms: (1) through scholarly feedback based on bibliographic meta-data and genre (in-copyright documents), (2) through feedback on full-text (non-copyright documents), and (3) through disambiguation of scholar-provided documents.

Similarly, the prototyping work done during WCSA with the Oxford University e-Research Centre demonstrated ways to uncover relationships (of interest to DH scholars) among text resources in a large collection like the HTDL and between HT resources and resources found in other important (from a scholarly perspective) text repositories such as the those of EEBO-TCP. This motivates improvements to HTRC workset provision for both infrastructure and the models that underpin it.

For workset tooling, the challenge is to take the utility proven in the ElEPHãT demonstrator, generalize the lessons, and integrate this functionality into the next generation HTRC Workset Builder services and APIs at a scale, reliability, and maintainability beyond that of the prototype. This work will include consideration of HTRC Workset services as an extensible platform that can incorporate new corpus- and study-specific attributes and relationships, and the policies and design patterns required to enact this within the infrastructure. Together, these will enable the creation of multi-level, multi-sourced worksets collating resources from HT and multiple other text archives whilst, crucially, maintaining compatibility with the HTRC Secure Commons.

The WCSA project also generated new avenues for research and development relating to the construction and use of workset models for scholarly analysis. For example, the approaches developed in our WCSA collaboration with Oxford (Nurmikko-Fuller et al., 2015), along with Illinois' workset data modeling research (Jett et al., 2015; Jett, 2015) highlight the growing benefit to be had by integrating RDF and other Linked Open data models into the HTRC infrastructure. The Oxford collaboration also provided a proof-of-concept demonstration of workset builder functionality being hosted by an entity external to HTRC. By extending and further exploring what we learned in this regard during the WCSA project, we will be able to more precisely identify and describe relationships among HT resources, between HT resources and resources elsewhere on the Web, and between HT resources and the entities relevant their creation, i.e., the individual, events, and places relevant to resource creation. This will give scholars using HTRC more control over the inclusion or exclusion of multiple editions of a work and over the selection into a workset of the best or most representative (for the purposes of a specific research inquiry) digital copy of a work or manifestation. Building on and extending the RDF-based workset descriptive data model for worksets will also allow us to implement a more rigorous means to create worksets that gather into them exactly the resources needed for a research inquiry, including resources that or more or less granular than a volume. This will allow scholars to gather into worksets specific pages and parts of a page of a volume, or to go the other way and gather into a workset more easily all the volumes in a triple-decker novel or the examples in HT of a particular work.

2.3 The Plan

Task 1: Implement Findings from WCSA into Workset Builder

The HTRC's current workset creation tool, Workset Builder, must be redesigned and rebuilt. It was created prior to the research done as part of WCSA so it does not incorporate the WCSA formal model, the recommendations of the workset report, nor the experiences gained through the WCSA prototyping projects (including ElEPHãT and Capisco).

It also does not provide many of the features desired by the community as outlined in the "Wish List" delineated above. Simply put, we need to upgrade the current Workset Builder framework from a proof-concept to a production deployment. The re-engineering tasks that we need to undertake with regard to the current Workset Builder and our supporting infrastructure are:

1. Transform MARC bibliographic metadata describing HT volumes into Resource Description Framework (RDF)—likely using schema.org, BIBFRAME (Library of Congress Bibliographic Framework)[25] or FRBRoo[26] semantics (an object-oriented scheme based on the Functional Requirements of Bibliographic Records report[27]), as described in a recent paper by Nurmikko-Fuller et al.[28] which will include augmenting MARC with Uniform Resource Identifier (URI) and replacing text with URIs (e.g., per recommendations of Mellon-funded Linked Data 4 Libraries (LD4L) and Linked Data 4 Production (LD4P) projects).
2. Augment existing HTRC knowledge stores with a triple store to serve as an ancillary index to metadata and full text indexes.
3. Augment existing bibliographic metadata with derived and non-traditional metadata and metadata at a more granular level (e.g., page-, paragraph-, and/or word-level features derived from text and/or image analyses, etc.).

[25] http://www.loc.gov/bibframe/

[26] http://www.cidoc-crm.org/frbr_inro.html

[27] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

[28] See Nurmikko-Fuller et al., 2015 and Jett, 2015 for specifics regarding these ontologies.

4. Implement exemplars of the non-traditional metadata augmentation process by integrating the outputs from the tagging algorithms of WCSA+DC key research partners Underwood (genre tags) and Hinze (concept tags).
5. Assess quality of genre tagging outputs—along the lines discussed on page 19 and also in regard to Task 6, starting on page 26—by having project staff manually verify the appropriate relationships between a randomly sampled subset of tags and their source pages.
6. Assess quality of concept tagging outputs via randomly sampled subset of tags and their source pages. Further assessment of quality will incorporate the use of external test corpora—such as SemEval2010.[29]
7. Maintain provenance of derived metadata.
8. Through Linked Open Data approaches and other methods as appropriate, connect HTRC Workset Builder to external services such that users can easily create hybrid worksets that include by HT and non-HT resources.
9. Better accommodate the importing (and exporting) of workset descriptions, e.g., add functionality allowing users to import and transform HT collections into HTRC worksets. Export workset descriptions as RDF graphs.
10. Redesign Workset Builder searching tools (currently a Blacklight implementation) and underlying architecture to take advantage of RDF & URIs in metadata and workset graphs and to be more granular and nimbly extensible.
11. Redesign HTRC Workset Builder to support the machine-aided creation of larger worksets (e.g., on the scale of worksets containing tens of thousands or even hundreds of thousands items).
12. Enable the repurposing and refinement of the HTRC Workset Builder platform tailored to scholarly disciplines and investigations, through the export and import of worksets and resources specific to those fields of study, to and from complementary corpora and tools.
13. Solicit feedback from user communities through at the HTRC UnCamp conference[30] and through a special Advanced Collaborative Support (ACS)[31] call.

Task 2: Extending WCSA Research Ability

WCSA project outcomes suggested several extensions of the core concept of an HTRC workset that we need to explore and determine strategies for implementation in order to create an environment for HTRC users that would better meet the needs of the scholarly community as outlined in the "Wish List." Extensions that we undertake to examine in the scope of WCSA+DC include:

1. Evolve and extend the formal WCSA workset model, as described in the CIRSS technical report (Jett, 2015) and incorporating requirements from the scholarly focus groups and prototyping projects.

[29] http://stel.ub.edu/semeval2010-coref/home

[30] HTRC UnCamp is an annual workshop/conference that engages specifically with HTRC user communities, and experts in DH and CL, about HTRC services present and future.

[31] Approximately every 6 months, HTRC provides DH scholars enhanced access to its staff expertise and resources through a competitive request for proposals (RFP) process called the Advance Collaborative Support (ACS) program. The ACS program, funded by the HathiTrust, targets scholars who have specific research questions that benefit from use of HTRC services, allowing HTRC staff and leadership to better understand the questions being asked by the scholarly community and the features and services that are needed to answer them. A special ACS call will be used to advance this project: ACS proposals will be built into year 2 of this project to employ as use cases for newly-implemented HTRC services surrounding worksets and data capsules.

2. Extend current private / public workset access model to support group or instructional class-accessible worksets.
3. Examine ways to accommodate, through Workset Builder or some other mechanism, the ability to annotate the RDF-compatible description of a workset.
4. Formalize the process for integration of, or linking with, external complementary corpora (e.g. EEBO-TCP) including identification of minimal or core terms within the Workset model to enable alignment and patterns and tools for creating WCSA compatible RDF by external Digital Libraries.
5. Explore ways to define worksets that include hybrid data types and sources, e.g., a workset of volumes having page images from HT and transcripts from ECCO-TCP or EEBO-TCP.
6. Develop Workset Builder to support common DH research tasks involving common workset manipulation tasks, e.g., to support citability of worksets, the creation of sub-worksets for tool training, etc.
7. Develop Workset Builder functionality and our model of workset description to accommodate the interconnectedness of and dependencies between worksets more generally.[32]
8. Within Workset Builder, consider ways to extend the ability to present and utilize relationships that support scholarly investigation, which may be distinct from those required to construct, maintain and curate worksets themselves.[33]
9. Develop Workset Builder to support creation of worksets gathering together exemplars of specific works or manifestations, where the exemplary nature is determined by analysis.[34]
10. Further integrate Workset Builder into the fabric of HTRC. As described below, worksets created using Workset Builder will need to be available to (and understandable by) instances of Data Capsule, and have references resolve while within the Secure Commons but remain in purely reference-based while outside the Commons).
11. Enable the development, through use of SPARQL, REST or alternative APIs, of workset constructors, viewers, and data contributors that are tailored to specific investigations or fields of study, but which maintain compatibility with the WCSA model and tools.
12. Develop and explore models for curated worksets, e.g., a workset is defined, included volumes are curated in some fashion (e.g., OCR corrected, header and footers removed), and the workset becomes the set of OCR-improved texts. In this manner the workset provides assistance to the scholar by handling preparatory tasks such as data cleaning; although care must be taken (and the model must support) sufficient provenance that the scholar can review these automated processes.

[32] These can take the form of hierarchical relationships—e.g., the workset of all English language fiction between 1800 and 1849 could have as a child workset, the pages from the parent workset identified through analysis as being examples of a particular genre; he complementary workset of late 16th and early 17th century English language plays not included in the workset of Shakespeare plays.

[33] For example, the need to map from descriptive bibliographic structures to event-based ontologies used within historical study (Nurmikko-Fuller et al. 2015). While concrete extensions will be implemented from scholarly requirements, this task will also develop the framework for future extensibility driven by scholarly needs.

[34] For example, the earliest edition of a work or digital instances of specific editions (having multiple digital copies in HT) that have the best OCR quality score as determined by a specified algorithm. Provide model support to distinguish between curation-derived versioning (e.g., improved OCR of a text) and domain-intrinsic versioning (e.g. editions of a book).

13. Determine the technical requirements and develop a preliminary model of the versioning of worksets and preservation of resources gathered into worksets that are meant to be persistent, long-term by leveraging some “dynamic workset” ideas coming out of the Research Data Alliance.

Task 3: Enhancing Data Capsule Support of Researcher Environment

HTRC Data Capsules provide a strong basis for non-consumptive computing in the HTRC Secure Commons. It will be extended as part of this project in a number of ways: a) integration of the new workset model, b) support for the new software framework we are developing that allows for chaining of analysis tools, c) support the custom research environments that are being set up for digital humanities (Underwood) and computational linguistics (Pustejovsky), and d) support for researcher-driven investigations that are over corpus sizes of 1,000 volumes so require HPC resources. This effort is broken into two tasks: one that deals with the researcher environment and the other that deals with core extensions to HTRC Data Capsules itself.

In Task 3 HTRC Data Capsules will be extended in several ways: a) integration of the new workset model, b) support for the new software framework we are developing that allows for chaining of analysis tools, c) support the custom research environments that are being set up for digital humanities (Underwood) and computational linguistics (Pustejovsky).

1. When sensitive data is constrained to exist solely within the confines of a secure environment, as the in-copyright data must remain within the Secure Commons, the researcher’s context (workset) must be trustworthy and pass seamlessly, in both directions, through the “cell wall” of the Data Capsule. Extend Data Capsule so workset can be moved in and out.
2. Modify Data Capsule model to operate on (make updates to) the new Workset.
3. Given that the workset is a data object whose trustworthiness cannot be determined merely by the fact that it is a workset (by type alone), extend the security model of the Data Capsule to assess the trustworthiness of a Workset in real time.
4. Build into Data Capsule the analysis-chaining (pipeline) framework that allows chaining of analysis tasks where only parts of analysis need to be carried out in secure mode. The chaining framework must support complex executions: partial executions, repeated executions, and breaks of days or weeks between stages.
5. An analysis task may be included by reference in the Workset or may be resident in the Data Capsule because a discipline custom Data Capsule is used. Design a process whereby these tools are located and imported into the Data Capsule as needed.
6. Extend the chaining software framework so that the intermediate result of one analysis task is passed downstream to the next task in the sequence between stages in the workflow is made part of the workset. That is, workset is the context of the analysis and contains a reference to any intermediate results that are created, along with references to the tools that carried out the analysis.
7. The data products of the computational analysis can become a threat to the security of the HTRC Secure Commons because it can be a vehicle for leakage of sensitive data. Develop automated means for studying data products to determine whether its contents contain sensitive data prior to release from the Data Capsule.
8. Support the activities of Underwood and Pustejovsky in setting up HTRC Data Capsules for their communities. This support-oriented task includes testing and hardening.

Task 4: Getting to Scale Securely and Smoothly

Task 4 addresses core extensions to HTRC Data Capsules: specifically support for researcher-driven investigations that are over corpus sizes of 1,000 volumes so require HPC resources.

1. Extend Data Capsule so that from within the scholar's Data Capsule VM, a scholar can execute a million+ volume analysis task that runs in parallel on the HPC resources located within the HTRC Secure Commons. For example, using n-gram pattern matcher (see below) as exemplar system in testing.
2. Guarantee that the architecture is at least a safe as the architecture within a single VM under the existing threat model.
3. Optimize the data extraction step of processing from a Workset that references 1M+ volumes. This involves moving external data into the cache location on the HPC resource, resolving the references in the Workset in a secure manner, and mapping the workset
4. Optimize the R-P n-gram pattern matching tool so that it can consult a database of information about copyright content and quickly discard irrelevant texts when it does not have a well defined Workset to start

As foundation for accomplishing this task, we have developed a new parallel n-gram pattern matching tool/system R-P n-gram pattern matcher[35] that executes very efficiently over millions of volumes of OCR. The tool applies a set of rules written as regular expressions to each page of the OCR text that it is given. It is built on top of a popular MapReduce framework so that an analysis task runs efficiently in parallel, with scale to thousands of processors/cores possible using HPC resources at Indiana. Interestingly, the R-P n-gram tool rules proved to have far better accuracy than finding n-grams in the full text Solr index because of the way in which Solr does its lexical analysis. The framework, which takes a list of volumes as input, can be extended easily to do other kinds of processing or analysis in parallel on millions of volumes or pages. The framework has been verified to generalize across the data analysis needs of an economist and an English professor.

Task 5: Partner Contribution to Data Capsule: Computational Linguistics

Pustejovsky's contribution to development of the Data Capsule involves providing access to high-performance cloud computing Natural Language Processing (NLP) facilities for members of the research and education communities who would otherwise have no such access, or who have little background in NLP, while reducing the often prohibitive overhead now required to adapt or develop new components. In the context of the proposed work, Brandeis will focus on enabling only the most critical NLP modules within the Data Capsule.

Most of these modules were adapted and generalized as web services for the Language Application Grid (LAPPS)[36] and, in the context of the Grid, were set up to allow easy chaining of the NLP analysis tools. In addition, basic NLP analysis tools not embedded in the LAPPS Grid will be selected and tuned to allow interaction with other NLP tools. If needed, the tools will be adapted for the DH domain by retraining classifiers and adapting rule sets.

The first stage of research (steps 1-2 below) examines which NLP web services are most appropriate for the research requirements of the DC users. This includes web services that have

[35] http://www.slideshare.net/BethPlale/hathitrust-secure-commons

[36] http://lapps.anc.org/

already been wrapped and integrated into the LAPPS Grid, as well as modules that are not yet available.

The second stage (steps 3-4) involves the integration of document-level and document collection processing (genre and topic identification) modules into the Data Capsule, as well as the most basic low-level processing (finding sentence boundaries, tokens, and parts of speech).

The third stage (steps 5-6) includes more computationally intensive NLP modules, such as finding "Named Entities" such as cities, countries, people, etc., as well as performing various levels of syntactic parsing at the sentence level.

Finally, step 7 entails a detailed evaluation of the NLP services. This involves: (a) assessing the overall performance of each component service within the Data Capsule; and (b) examining the possible workflow configurations of the different services as configured in distinct pipelines. The steps required are:

1. Identify appropriate NLP web services from the LAPPS Grid, and configure for inclusion into the Data Capsule. The preliminary list of services includes sentence identification, tokenization, part-of-speech tagging, chunking, constituent parsing, dependency parsing, named entity relation extraction.
2. Select some other standard NLP modules that are not yet available on the LAPPS Grid, including coreference linking, semantic role labeling, event spotting, time stamping, document type identification, genre classification, and topic classification.
3. Adapt and train Document Structure Parser, Genre Classifier and Topic Classifier over English-language book corpus (1922-2000).
4. Adapt, train, and tune Sentence Splitter, Tokenizer, and Part-of-Speech Tagger over corpus.
5. Identify appropriate types from corpus for Named Entity Recognizer. Train and tune over corpus. Adapt event spotting and time stamping.
6. Train and tune Shallow Parsers, coreference linking and semantic role labeling over corpus.
7. Evaluation of results, revise algorithms, publish/present research.

Task 6: Partner Contribution to Data Capsule: Digital Humanities

Underwood's contribution to Data Capsule development will have two broad parts, which are connected to produce a substantive literary-historical argument about the history of character in twentieth-century fiction. The project will also deliver a set of resources that other literary scholars (or historians) could use to write similar arguments about representations of people across very large digital collections.

The first stage of research (steps 1-4 below) creates a page-level workset of English-language fiction in the twentieth century, distinguishing works of fiction from (say) nonfiction prefaces or ads at the back of the book. This builds on similar work HTRC has recently published on public-domain volumes, but extends it beyond the wall of copyright (Underwood et al., 2015). The genre classification entailed in this project only covers categories where we have found, empirically, a high level of human consensus (paratext and body text, prose and verse, fiction and nonfiction).

The second stage (steps 5-8) uses those worksets to extend research on the history of character Underwood has already undertaken with David Bamman and Noah Smith (Bamman et al., 2014). The authors of that earlier project developed a workflow (BookNLP) that uses computational linguistic tools like those described in Task 5 to trace references to a single character across an entire book. They were able to apply that workflow to 15,099 nineteenth-century novels drawn from HathiTrust, and develop a model of character types in nineteenth-century fiction. To extend that research beyond 1922, scholars will need to apply the workflow non-consumptively inside a Data Capsule.

The steps required for the whole process are:

1. Extract page-level features from non-serial post-1922 volumes, using HTRC's existing feature-extraction workflow and extending it to the copyrighted portion of the library.
2. Develop page-level genre training data. (This sounds simple, but it is labor-intensive; we coordinate multiple readers to assess levels of human agreement.)
3. Classify the pages of post-1922 English volumes by genre, using methods developed in "Understanding Genre in a Collection of a Million Volumes" (Underwood, 2015). Document our methods, reporting out-of-sample accuracy and inter-annotator agreement so other scholars know how far these results can be relied on. (We have found in practice a high level of human consensus about broad genre boundaries like those between fiction, nonfiction, and paratext: readers agree about 94.5% of pages. Algorithmic methods agree with human consensus almost as often: we achieve 93.6% accuracy. By filtering out certain problematic categories like miscellanies and school readers, we can create an even more reliable subset of volumes—precision can be over 97%.)
4. Assemble a page-level workset that covers only twentieth-century fiction (not front matter, or nonfiction prefaces, etc.). Extracted features for those pages will become a resource for other scholars, as well as a foundation for the next step of this research. Test workset accuracy on new volumes that were not part of our original training or test sets (since we're working beyond 1922, this will require consulting physical books, and/or randomly selected page scans from the HT corpus).
5. Implement Stanford CoreNLP[37] inside a Data Capsule.
6. Implement David Bamman's BookNLP inside a Data Capsule, drawing on Stanford CoreNLP as a library, and integrating recent improvements suggested in Vala, Jurgens, Piper, and Ruths 2015.
7. Apply BookNLP to the page-level workset of twentieth-century fiction mentioned in (4) above, in order to generate a dataset of information about literary characters in at least 50,000 novels 1780-2000.
8. Create a manually-corrected subset of the characterization dataset in order to assess its reliability.
9. Use our characterization data to write a substantive argument about the history of character, 1780 - 2000, focusing especially on the way characterization reveals assumptions about gender that change across time. Publish that article, and also publish the underlying data (lists of characters in particular novels, with extracted descriptions of the character, their quoted language, and verbs they govern) as a resource that can be used by other scholars.

[37] http://nlp.stanford.edu/software/corenlp.shtml

This work benefits other scholars in three ways:

1. It creates a workset of English-language fiction that they can use in their own research. This extends the literary workset 1700-1922 already published by HTRC (Underwood et al., 2015).
2. It also creates a library of utilities that other scholars can use to extract context-sensitive information about people from behind the wall of copyright. These utilities are potentially as useful for historians as they are for literary scholars; one could trace changing perceptions of historical figures, for instance, by identifying what was said about them, and what they were represented as doing, across hundreds of thousands of volumes.
3. Our goal is also to demonstrate, concretely, how tools like this can matter for cultural history—and thus motivate scholars to use the new resources we have developed in (1) and (2).

3. Data and Use Case Context for Effort

3.1 Data Description

At the time of the writing, the HTDL encompasses 13.7 million volumes of digitized content. There are 7.3 million unique titles in the corpus. The difference in quantities is due to duplicates, multi-part or serialized works. These volumes are composed from nearly 4.78 billion pages. Each page is represented by a high quality scanned image and two kinds of OCR-generated text (coordinated and uncoordinated OCR), yielding almost 15 billion file objects amounting to a total of 613 terabytes of data.

Of the 13.7 million volumes, only 5.3 million (~39%) fall within the auspices of public domain in all jurisdictions or are open access resources. The remaining 8.4 million volumes (~61%) are protected by copyright and are inaccessible to scholars.

Within the public domain data, topicality is known for about half of the volumes, 2.7 million out of 5.3 (~51%), and the most common topical areas represented (for those works whose Library of Congress classifications are known) include U.S. law, English literature, and local histories of the Americas as Table 1 shows. Similarly, as Table 2 shows, works in the public domain are most frequently in English, German, French, Spanish, and Italian.

Topicality is known for almost two-thirds, 5.4 million out of 8.4 (~65%), of the copyright-restricted portion of the corpus. The absence of topicality information for major portions of the corpus presents a significant barrier for humanities scholars to overcome when they are gathering their research materials. Among the copyright-restricted subset of the corpus, the most common topical areas (for those works whose topical areas are known) are Asian history (inclusively), Romance Language literature (i.e., literature in French, Spanish, Italian, and Portuguese), and works about industries, labor, and land use (see Table 1). Similarly, while works in English, German, and French also dominate this part of the corpus too, the next most frequent languages represented are Chinese and Japanese (see Table 2).

Topical Area (Public Domain)	Count	Topical Area (Under Copyright)	Count
KF - Law (United States)	112,907	DS - History of Asia	238,239
PR - English literature	101,662	PQ - French literature - Italian literature - Spanish literature - Portuguese literature	204,654
F - Local History of the United States and British; French; and Latin America	95,410	HD - Industries; Land use; Labor	181,683
E - History of America	87,725	PL - Languages of Eastern Asia; Africa; Oceania	164,058
AP - Periodicals	79,415	Z - Books (General). Writing. Paleography. Book industries and trade. Libraries. Bibliography	160,209
HD - Industries; Land use; Labor	71,752	HC - Economic history and conditions	120,912
Z - Books (General). Writing. Paleography. Book industries and trade. Libraries. Bibliography	68,987	PT - German; Dutch; Scandinavian; Old Norse; Old Icelandic; Old Norwegian; Modern Icelandic; Faroese; Danish; Norwegian; and Swedish literature; Flemish literature since 1830	111,075
PS - American literature	67,900	PN - Literature (General)	107,662
PQ - French literature - Italian literature - Spanish literature - Portuguese literature	63,539	F - Local History of the United States and British; French; and Latin America	104,026
BX - Christian Denominations	56,025	PG - Slavic languages; Baltic languages; Albanian language	100,626

Table 1: Top 10 topical areas of works

Language (Public Domain)	Count	Language (Under Copyright)	Count
English	3,163,099	English	4,039,323
German	642,102	German	637,657
French	565,419	French	459,755
Spanish	155,605	Chinese	457,592
Italian	134,786	Japanese	398,209
Latin	108,027	Spanish	345,289
Japanese	89,099	Russian	337,251
Russian	71,560	Italian	185,893
Chinese	60,984	Arabic	151,738
Dutch	47,716	none (Unknown)	147,232

Table 2: Top 10 most frequent languages of works

The large disparities between what can and cannot be accessed by humanities scholars present a significant barrier for research. For instance, if a researcher is interested in analyzing a large body of works in Arabic, they are at a great disadvantage as ~95.2% of HT's Arabic corpus lies beyond the reach of analytics, that is, behind the barricade of copyright restrictions. This is exactly the type of scenario that the HTRC's Secure Commons is designed to remedy. The copyright barrier also presents a significant barrier to humanities scholars working in contemporary areas of literature and history. As Figure 3 illustrates, there is a vast difference in the relative time regions covered by works in the public domain and those that remain under copyright. The graph clearly shows the 1923 copyright wall, which stands as an obstacle to researcher access to the cultural products of the last nine decades.

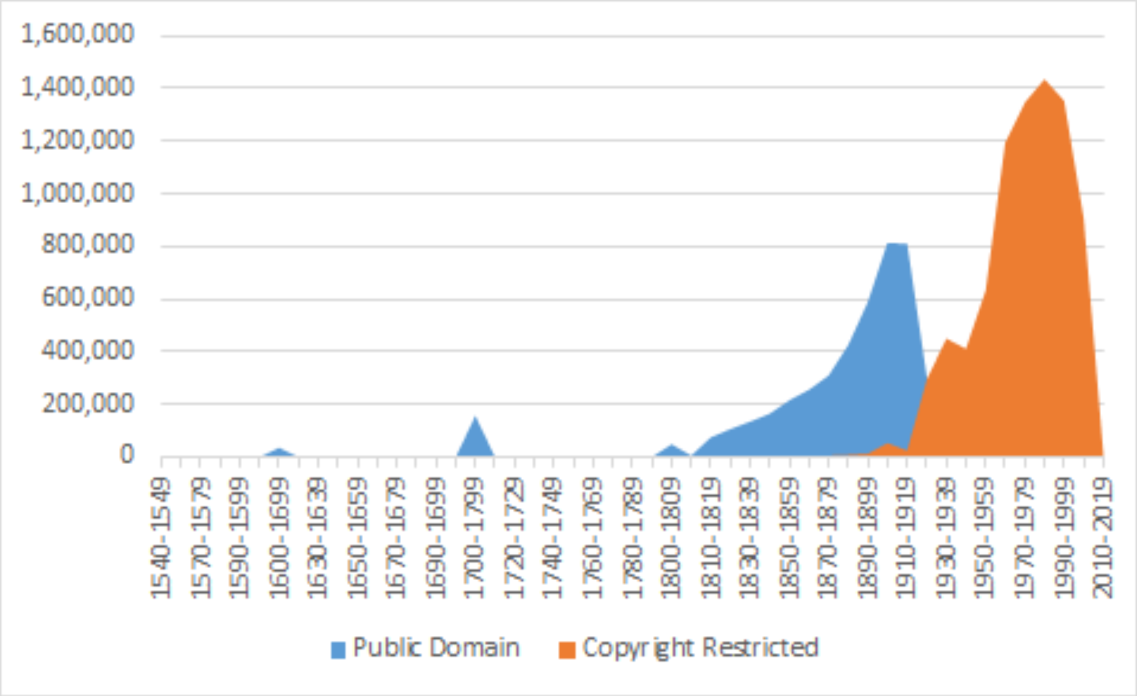

Figure 3: Volume counts in public domain (blue) and copyright-restricted (orange)

3.2 Use Cases in Context

The proposed large-scale deployment of the Capisco semantic tagging system to be undertaken in WCSA+DC constitutes a pathway for the type of fine-grained semantic tags required for in-copyright material. Using Capisco, copyright protection is ensured through utilizing semantic concepts and synonymous keywords in searching and not the underlying text itself. No elements of the copyrighted text are stored. Using traditional text-based search requires scholars to identify appropriate keywords; relevant sources remain undetected unless the right keyword is found. Large sets of unrelated documents may be included in search results when encountering identical terms referring to unrelated concepts. This problem is exacerbated for documents under copyright, as a scholar cannot manually check the results. Augmenting bibliographic metadata with fine-grained marker terms of suitable concepts could alleviate both issues and raise the quality of search results. This approach plays the same role as query expansion, but follows the reverse strategy of not enhancing queries but the document corpus. A mere extension with synonymous terms is not sufficient (and rather misleading), as appropriate keywords are those that match the semantic context of a document. An added benefit to Capisco's approach of identifying supporting context for semantic concepts is that it further affords a level of immunity to some of the inherent OCR errors contained in the HT data. Misreadings of terms that do not find semantic support in the document are discarded. Given this, the system is self-correcting for OCR pages with sparse errors and flags those needing further investigation.

In the ElEPHãT WCSA prototyping project, the selection of materials consisting of English books printed before 1700 was made not only due to local expertise of the subject matter within the University of Oxford and the Bodleian Libraries; but also because the vast majority of these texts are out of copyright, allowing development of experimental systems and interfaces without the robust access controls required for post-1922 material. Through our interactions with scholars at consultative workshops, it is clear that the ElEPHãT tools provide real value to academic investigations by enhancing the linking of resources within and between corpora, and that these advantages transcend any particular field of study to encompass scholarship using HT resources in general. Having developed and trialed these technologies in the unencumbered ElEPHãT environment, in this project we must integrate these advances with into new platforms and tools (through enhanced worksets and data capsules) that bring these benefits to all scholars using HTDL, including those working with post-1922 resources.

Since literary scholars, historians, and social scientists generally need to begin by selecting a representative sample of documents, restricting twentieth-century research to the unusual subset of texts that happen to have escaped copyright is rarely a viable option. At the moment, large-scale academic data mining of library collections effectively comes to an end in 1922. This is particularly unfortunate because the twentieth century is exactly the period when the scale of the cultural record starts to be unmanageable without data mining. Literary scholars, for instance, find it nearly impossible to generalize about twentieth-century fiction as a whole. Distant reading is beginning to trace interesting trends in the nineteenth century, but these narratives come to an abrupt stop in 1922.

For computational linguists, access to the post-1922 English language corpus is needed in order to adequately model and train the document-level and document collection processing (genre and topic identification) modules, as these are highly dependent on lexical content for creating model signatures. Further, since languages also exhibit syntactic constructional changes, it is important to have contemporary phrasal and sentence patterns for training the syntactic parsing modules. Finally, diachronic linguistic corpus analysis that includes the period after 1922 can potentially reveal interesting and significant results from the corpus that may not be found when only examining pre-1922 volumes.

4. Staff and Organization Qualifications

4.1 Organizational Strengths

The HathiTrust Research Center (HTRC)

Founded in 2011, the HathiTrust Research Center is a unique collaborative research center jointly hosted at the University of Illinois at Urbana-Champaign (Illinois) and Indiana University Bloomington (IUB). The HTRC was formed with the following goals:

- Support innovation in cyberinfrastructure to deliver optimal access and use of the HathiTrust corpus;
- Explore innovation in delivering efficient access to copyrighted material that preserves and shapes the non-trivial restriction of “non-consumptive research”;
- Identify and host existing data analysis, text mining and retrieval tools;

- Seek ways to enhance the value of the HathiTrust; and,
- Explore innovative methods for creating a sustainable research center.

In addition to drawing on the complementary strengths of IUB and Illinois, HTRC works closely with the HathiTrust and its over 80 member organizations as wide field of expertise and collaborators upon which to call. In addition, this proposal harnesses the informed and active leadership of the entire HTRC Executive Management Team (ExMgt) consisting of PIs Downie and Plale along with Beth Namachchivaya, Associate University Librarian for Research, Associate Dean of Libraries, and Professor, Illinois; Robert McDonald, Associate Dean for Library Technologies, IUB; and John Unsworth, Vice Provost, Chief Information Officer, University Librarian, Professor of English, Brandeis University. Mike Furlough, Executive Director of the HathiTrust, is also an ex-officio member of the team, and brings his expertise and perspective to the team. Beyond their integral role to the operation of the HTRC, ExMgt brings considerable collective expertise and experience in digital humanities, informatics, cyberinfrastructure, libraries, instruction, and computational linguistics. The HTRC also recently ratified an official 4-year work plan in June 2014 that was approved by the HathiTrust board for operation of the HTRC. The proposal included an award from HathiTrust of financial support for the Center, along with in-kind contributions from each of the HTRC's two host institutions.

The University of Illinois at Urbana-Champaign (Illinois)

Illinois is a nexus for digital humanities, information science and knowledge management research and development. Building on a now long history of close and successful collaboration, the proposed project will be a joint endeavor of the HTRC and the Graduate School of Library and information Science (GSLIS). We will draw on the strengths of these two entities, the institutional and collaborative strength of the UI Library as well as experience gained in past and ongoing research partnerships with the Illinois Program for Research in the Humanities, the National Center for Supercomputing Applications (NCSA) and the Center for Informatics Research in Science and Scholarship (CIRSS). Close working partnerships with these and other specialized research centers and consortia beyond the University enable the HTRC to, develop and provide multi-disciplinary, innovative and impactful services to scholars.

The Center for Informatics in Science and Scholarship (CIRSS)

CIRSS conducts research on information problems that impact scientific and scholarly inquiry with a specific focus on how digital information can advance the work of scientists and scholars, the curation of research data, and the integration of information within and across disciplines and research communities. CIRSS researchers, faculty and staff bring a range of expertise to the center's projects in areas including empirical studies of scientific information use, information modeling and representation, ontologies, data curation, and digital research collections and technologies. The center's staff includes project coordinators, research assistants and other academic staff with experience in project management, quantitative and qualitative methods, research with human subjects, and the design and conduct of multi-method research and evaluation studies in information science and cognate social sciences. CIRSS builds on synergies in four key intellectual areas: 1) digital humanities; 2) collections, curation, and metadata; 3) e-Science; and 4) socio-technical data analytics.

CIRSS is a core research center within GSLIS. Founded in 1893, GSLIS, the iSchool at Illinois, is a world leader in library and information science education, research and practice. Consistently ranked as one of the very best in the field, GSLIS has earned its reputation by creating pioneering and innovative educational opportunities, by leading groundbreaking research to advance preservation of and access to information in both traditional and digital libraries, and through its services and strong commitment to outreach and community development.

The University of Illinois Library at Urbana-Champaign (Illinois)

The Library at Illinois is one of the preeminent research libraries in the world. As the intellectual heart of the campus, the Library is committed to maintaining the strongest possible collections and services and engaging in research and development activities in pursuit of the University's mission of teaching, scholarship, and public service. The Library provides a rich range of services geared to support the curricular and research needs of students and faculty and serve the dynamic needs of scholars in the digital age both local and remote. The Library was established in 1867 with only 644 books purchased with $1,000 appropriated by the State of Illinois. Today it houses more than 22 million items, and it is known for the depth and breadth of its collections. Materials from the library are actively used, with more than 1.4 million items circulated annually and subscriptions and licenses for over 50,000 e-journals resulting in over 7 million user click-throughs per year via an e-resource registry and over 11 million full-text downloads. The Library currently employs approximately 90 faculty and 300 academic professionals, staff, and graduate assistants who work in multiple departmental libraries located across campus, as well as in an array of central public, technical, and administrative service units. The Library also encompasses a variety of virtual service points and "embedded librarian" programs that provide library services to scholars across the spectrum of research environments. Librarians are full faculty members of the University and contribute significantly to scholarly literature in their respective fields of study. The Library plays a leadership role in regional, national, and international organizations; provides services to users throughout the State of Illinois; and serves as an integral part of the worldwide scientific and scholarly community.

Indiana University Bloomington (IUB)

Founded in 1820, Indiana University Bloomington is the flagship campus of IU's eight-campus system of higher education that supports the statewide mission of providing broad access to undergraduate and graduate education for students throughout Indiana, the United States, and the world, as well as outstanding academic and cultural programs and student services. Innovation, creativity, and academic freedom are hallmarks of our world-class contributions in research, professional education, and the arts. The IU Bloomington campus is unique in bringing together a world-class technology organization (OVPIT) with a distinguished and forward thinking library system (IU Libraries) and the near 100 faculty strong School of Informatics and Computing (SoIC) in synergistic activities that advance cyberinfrastructure innovation, big data, and data science research and development. Building on such cyberinfrastructure successes as Big Red II, one of the fastest university owned high performance computing instruments (#110 in the Top 500), and the IU Data Capacitor II, our large-capacity, high-throughput, high-bandwidth file system serving all IU campuses (3.5 PB total capacity), the Research Technologies (RT) Division of our centralized University Information Technology Services (UITS) provides a solid backbone upon which to build strong cyberinfrastructure collaborations such as the HathiTrust Research Center. The HTRC leverages the collaborative engagement of our key campus cyberinfrastructure providers, data management providers, and research and development scholars to achieve success for the HathiTrust Research Center.

IU Data To Insight Center (D2I) and Pervasive Technology Institute (PTI)

The Indiana University arm of the HTRC is located at the Data To Insight Center, which is collaboration between the School of Informatics, the Indiana University Libraries, and

University Information Technology Services (UITS) at Indiana University. The center engages in interdisciplinary research and education in the preservation of scientific data, digital humanities, large-scale data management, data analytics, and visualization. The Center's current projects engage researchers in the humanities, geography, sustainability science, atmospheric science, informatics, computer science and digital libraries. A key project based within the Center – the HathiTrust Research Center (HTRC)—provides data analysis support, consulting services, data storage, results archiving, and computational resources for analysis of the HTDL, a co-developed project with the University of Illinois. Because of the Data to Insight Center's close working relationship with Indiana University's University Information Technology Services (UITS), the Center is well positioned to engage in projects that can be strengthened by IU's substantial investment in cyberinfrastructure compute and storage resources, and can in turn further strengthen these investments. The Center engages in outreach and education in service to the university and its students, the community, the State of Indiana, and the nation.

D2I is an affiliated research center of the Indiana University Pervasive Technology Institute (PTI). The mission of the PTI is to improve the quality of life in the State of Indiana and the world through novel research and innovation and service delivery in the broad domain of information technology and informatics. As a world-class organization, PTI pairs fundamental academic computational research with the widely known strengths of Indiana University through innovations and service delivery in networking and high performance computing. By means of organization into research and service centers, PTI encourages collaboration that crosses center boundaries, where practice informs the science, and science advances the practice, the results of which advance the university, state, and nation as a whole.

Indiana University Libraries

The Indiana University Libraries comprise one of the leading research library systems in North America supporting all eight campuses of the Indiana University system. The IU Libraries are committed to maintaining the research collections of Indiana University, which include more than 7.8 million books in over 900 languages. The materials support every academic discipline, with an emphasis in the humanities and social sciences, and support IU's mission of providing broad access to undergraduate and graduate education for students throughout Indiana, the United States, and the world, as well as outstanding academic and cultural programs and student services. As both a research collection repository and scholarly service provider, the IU Libraries bring a "concierge philosophy" to our support of undergraduate education via our Library Learning Commons and to our scholarly researchers and graduate students via our Scholar's Commons services. A leader in digital library development and services, the IU Libraries are an active member of the Hydra and Fedora repository communities and have brought immense leadership to creating digital content libraries of time-based media such as those delivered through our open-source applications that support the IU Variations Digital Music Library[38] and the IU Avalon Media System.[39] The IU Libraries have embarked on a new scholarly venture with the IU Press in establishing the IU Office of Scholarly Publishing, which supports a variety of open access scholarly publishing ventures for the IU and international research communities. Additionally, the IU Libraries are an important collaborator with the Office of the Vice-President for Information Technology and the Office of the Vice-Provost for Research in supporting campus and system-wide digital humanities initiatives including being founding members of the HTDL and the HathiTrust Research Center. Additionally, the IU Libraries are a leader in the

[38] http://variations.sourceforge.net

[39] http://www.avalonmediasystem.org

preservation of born-digital and converted digital content serving as founding members of the Academic Preservation Trust, and the Digital Preservation Network. Librarians at IU are faculty members of the University and contribute to the output of scholarly literature in a variety of areas covering information science, library science, and archival science.

University of Oxford e-Research Centre

The Oxford e-Research Centre at the University of Oxford is composed of diverse groups conducting digital research in and across multiple disciplines. With origins in the UK e-Science programme, it is the UK's largest e-Research Centre and has a prominent role on the international stage as well as nationally. A variety of research interests are represented in the Centre, including digital humanities, social sciences, scientific computing, biological and physical sciences, and visual computing. The Centre brings together different expertise in a multidisciplinary environment and collaborates with industrial partners and government entities to ensure maximum impact of its research and activities. The Centre is the hub of an interdisciplinary network within Oxford which includes the Faculty of Music, the Oxford Internet Institute, the Bodleian Libraries, Academic IT Services, and the Oxford Research Centre in the Humanities, working to advance digital scholarship and the transformation of research practice. The Digital Humanities team has developed significant Linked Data expertise over an eight-year period, for example in the Claros (world art), SALAMI (music), and ElEPHãT (early English texts) projects, and has recently enlarged in the music research area thanks the AHRC Transforming Musicology large grant. Researchers from the Centre play a pivotal role in bringing these new methods and technologies to the next generation of humanities scholars through their involvement in the annual Digital Humanities at Oxford Summer School. Major research activities span from Citizen Science, Cultural Heritage, Social Machines, and Smart Society to computational musicology and computational social science, Scientific Computing on novel hardware architectures and the Square Kilometre Array Radio Telescope. The Centre is closely engaged in new forms of Scholarly Communication and has an events and communications team specializing in events, workshops, and public engagement. The Centre is also a partner in the UK Software Sustainability Institute.

4.2 Principal Investigators

The overall project PI will be J. Stephen Downie (Illinois). Beth Plale (Indiana University at Bloomington) and Timothy Cole (Illinois) will serve as co-PIs. The project management structure is described below in section 4.6. Here is listed background and qualifications of PI and co-PIs.

Project PI: J. Stephen Downie (Illinois) (1.5 summer months)

J. Stephen Downie is the Associate Dean for Research and a Professor at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. Downie is the Illinois Co-Director of the HathiTrust Research Center. He is also Director of the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) and founder and ongoing director of the international Music Information Retrieval Evaluation eXchange (MIREX). He PI on the multi-institutional HathiTrust + Bookworm Project, funded by the National Endowment for the Humanities. He was the Principal Investigator (PI) on the multinational, Mellon-funded Networked Environment for Music Analysis (NEMA) project. He was the United States PI on the Structural Analysis of Large Amounts of Music Information (SALAMI) project, jointly funded by the National Science Foundation (NSF), the Canadian Social Science and Humanities Research Council (SSHRC), and the UK's Joint Information Systems Committee (JISC). He has been very active in the establishment of the Music

Information Retrieval (MIR) community through his ongoing work with the International Society for Music Information Retrieval (ISMIR) conferences and has served as ISMIR's President. He has published and lectured extensively on a wide range of digital humanities, digital libraries, and cultural informatics topics. He holds a BA (Music Theory and Composition) along with a Master's and a PhD in Library and Information Science, all earned at the University of Western Ontario, London, Canada.

Project Co-PI: Beth A. Plale (IUB) (8.25%)
Beth Plale is a founding director of the HathiTrust Research Center. A Full Professor of Informatics and Computing at Indiana University, Professor Plale directs the Data To Insight Center and is Science Director of the Pervasive Technology Institute within which D2I is housed. Professor Plale's postdoctoral studies were at Georgia Institute of Technology, and her PhD in computer science is from State University of New York Binghamton. Over the last 17 years, Professor Plale has authored 150 publications and had PI or co-PI roles in over $55,000,000 of externally funded research dollars from industry, private foundations, and federal agencies. Dr. Plale's research interests are in long-term preservation of scientific and scholarly data, data analytics, tools for metadata and provenance capture, data repositories, and data-driven cyberinfrastructure. Plale is deeply engaged in interdisciplinary research and education.

Plale chairs the Technical Advisory Board of the Research Data Alliance (RDA), and serves on the steering committee of RDA/US. She has served on the Steering Board of the Open Grid Forum (OGF), and currently sits on numerous advisory boards, including a working group in data management for NOAA, and as a member of the External Advisory Board for the NIH funded Center for Expanded Data Annotation and Retrieval (CEDAR) housed at Stanford University. She has served as general chair for two prestigious conferences that span her research interests: the prestigious ACM High Performance Distributed Computing (HPDC) and International Provenance and Annotation Workshop (IPAW). She is Department of Energy (DOE) Early Career Awardee and past Fellow of the university consortium Academic Leadership Program.

Project Co-PI: Timothy W. Cole (Illinois CIRSS/Library) (5%)
Timothy W. Cole is Mathematics Librarian (University Library) and CIRSS Coordinator for Library Applications (GSLIS). He is a co-PI for the Workset Creation for Scholarly Analysis project and for the Emblematica Online projects. He was previously the PI for the Open Annotation Collaboration projects (all phases, 2009-2013) and the Digital Collections and Content projects (phases 1 & 2, 2002–2007), as well as PI or co-PI for multiple other projects involving metadata and digital library system design, interoperability and implementation. A member of the Illinois faculty since 1989, he has held prior appointments as Interim Head of Library Digital Services and Development, Systems Librarian for Digital Projects and Assistant Engineering Librarian for Information Services. He is a member of the International Mathematical Union Committee on Electronic Information and Communication and a member of Library Hi Tech Editorial Board. He has published and presented on metadata and LOD best practices, OAI-PMH, digital library interoperability, Open Annotation, and the use of XML for encoding metadata and digitized resources in science, mathematics and literature.

4.3 Key Research Partners

Ted Underwood (Illinois) (75% of one summer month)
Ted Underwood is Professor of English and Liberal Arts and Sciences Centennial Scholar at the University of Illinois, Urbana-Champaign. He has extensive experience both in traditional

literary scholarship and in computational research on large digital libraries. Publications from the pre-computational part of his career include a book on British Romanticism (with Palgrave, 2005) and a book on the history of literary periodization (with Stanford University Press, 2013), as well as articles in *PMLA* and *Representations*. More recently, he has been exploring the interpretive leverage that can be gained by applying machine learning to literary history, especially across long timelines. He has published articles on this topic in *Representations* and *New Literary History,* and has co-authored papers presented at the Association for Computational Linguistics and the IEEE Conference on Big Data. He was recently PI of the project “Understanding Genre in a Collection of a Million Volumes,” supported by an ACLS Digital Innovation Fellowship and a NEH Digital Humanities Start-Up Grant; datasets produced by that project are publicly available on the web and have supported scholarly research as well as a popular piece on economics and the novel that appeared in Slate. On this project, Underwood will complete work in Task 6 as well as supervise and collaborate with English RA and Undergraduate Hourly Assistant in conjunction with PI and Co-PIs as well as LIS RA, and Hourly Programmer.

James Pustejovsky (Brandeis University) (8%)
Dr. James Pustejovsky holds the TJX Feldberg Chair in Computer Science at Brandeis University, where he is also Chair of the Linguistics Program, Chair of the Computational Linguistics MA Program, and Director of the Lab for Linguistics and Computation. Pustejovsky is chief architect of TimeML and ISO-TimeML, a recently adopted ISO standard for temporal information in language. He led development of a platform for temporal reasoning in language, called TARSQI (www.tarsqi.org). He has recently spearheaded the development of ISOspace, a comprehensive specification for spatial information as expressed in language, which has recently been adopted as an ISO standard, and was used in the just completed SpaceEval task for Semeval 2015. Pustejovsky is PI on a recently awarded DARPA grant, “Communicating with Computers”, which aims to develop a library of semantic primitives that enable communication between humans and intelligent artificial agents. Additionally, he is co-PI on the DARPA-funded "Big Mechanisms" effort with SIFT (Mark Burstein) and University of Colorado at Denver (Larry Hunter). This work involves recognizing mechanisms and causal inferences in biological pathways implicated in cancer. His contribution to this work involves the development of "causal event models" for linguistic expressions, in this case, within the biological literature on cancer. On this project, Pustejovksy will lead the work in Task 5, and supervise and collaborate with Verhagen, Brandeis RA in conjunction with the PI and Co-PIs.

Kevin Page (University of Oxford) (10%)
Dr. Page is a Senior Researcher at the University of Oxford, where his interests lie in the practical development and application of semantic computing to solve information gathering, structuring, and analytical problems as presented ‘in the wild’ by disciplines across the sciences and humanities. At the Oxford e-Research Centre he has recently led projects in the Digital Humanities, with a strong track record of multidisciplinary collaboration and successful application of his extensive experience of Semantic Web and Linked Data technologies. He is principal investigator of the Semantic Linking of BBC Radio (SLoBR) project with scholars of early music at Goldsmiths, University of London; and of the Early English Print in HathiTrust (ElEPHãT) WCSA sub-award, which worked on a collection of 5 billion digitized pages with colleagues from the Bodleian Libraries and the University of Illinois HathiTrust Research Center. His earlier work on web architecture and the semantic annotation and distribution of data has, through participation in several UK, EU, and international projects, been applied across a wide variety of domains including sensor networks, pervasive computing, medical education,

music information retrieval, and remote collaboration for space exploration (with NASA). He was a member of the W3C Linked Data Platform Working Group and of several W3C Community Groups.

On this project, Page will focus on work in Tasks 1 and 2, primarily centered around the enhancement of the workset through supporting hybrid data source and tailoring/curation of worksets to different scholarly settings/research.

Annika Hinze (University of Waikato) (10%, unfunded)
Dr. Annika Hinze is a Senior Lecturer in the Department of Computer Science at the University of Waikato, New Zealand, where she is head of the research Lab on Information Systems and Databases (ISDB). She was invited Professor for Context-aware Systems at the Humboldt University Berlin, Germany in 2009. She has a track record of research on semantic analysis and context-aware systems. She was PI or co-PI on a number of successful nationally and internationally funded projects on semantic document enrichment and non-expert interfaces for semantic technology—grants funded by the NZ Royal Society, German Federal Ministry of Education and Research (BMBF), and German Academic Exchange Service (DAAD). This work introduced human aspects into semantic annotation approaches and explored the quality of expert and non-expert semantic full-text annotation. She received two NZ BuildIT grants for young researcher development and was key researcher in a number of projects on context-based information presentation and event-based systems, funded by the New Zealand Ministry of Business, Innovation and Employment (MBIE), New Zealand Foundation for Research, Science and Technology (FRST) and BMBF. She was previously Principal Investigator of the Capisco project on Semantic Analysis of Documents from the HathiTrust Corpus.

On this project, Hinze and a programmer from Waikato will deploy her Capisco semantic tagging system on the HT corpus and work it into the Workset Builder.

4.4 Existing staff

Ryan Dubnicek, Project Coordinator (Illinois) (25%)
Ryan is the current Executive Assistant and Project Coordinator for the HTRC, having joined in 2013. His efforts for HTRC include meeting and event planning, budget tracking, reporting, and proposal development. Ryan has a BA in English and is currently working on his MSLIS, both at Illinois.

Janina Sarol, Project Developer (Illinois) (50%)
Janina Sarol is a Visiting Research Programmer (University Library). Her primary job assignment is to support digital library research. Sarol, who has a BS in Computer Science awarded by the University of the Philippines–Diliman in 2011 and is a member of the W3C Web Annotation Working Group and the W3C Schema.org Community Group, joined the University Library in early 2014 to take over as the lead developer for the second phase of the *Emblematica Online* project. In this role she has implemented a number of LOD features in the *OpenEmblem Portal*. In 2014 she also served as lead developer for the Library's project to create a schema.org LOD snapshot of the UIUC general collection catalog (5+ million bibliographic MARC records and 10+ million holding records) and participated in OCLC Developer House, working on extensions to the instance of VuFind that Karen Coombs of OCLC has modified to work with RDF and the WorldCat API. Sarol continues as the lead developer for *Emblematica Online* (through November 2015) and now serves also as lead developer for the *Workset Creation for Scholarly Analysis* project (through September 2015). For the proposed project, Sarol will be the

lead developer for Workset Builder redesign and the integration of new, richer worksets into WB along with the efforts to integrate WB into the fabric of the HTRC services. She will work closely with and under direction of PIs, Co-PIs, as well as alongside Ops Manager, Project Programmer and other staff.

Marc Verhagen (Brandeis University) (5%)
Marc Verhagen is a senior research scientist at the Computer Science Department at Brandeis University. He has 20+ years of experience in Natural Language Processing (NLP). His main areas of expertise are temporal and spatial processing, technology extraction from scientific documents, annotation tools and strategies, processing of health records, and NLP web services. On this project, Verhagen will be responsible, in conjunction with Pustejovsky, for executing Task 5, as well as supervising Brandeis Graduate Assistant.

4.5 Project staff to be named / hired

Research Programmer (IUB) (100%)
The Research Programmer will have at least a master's degree in Computer Science, PhD preferred. They will be responsible for the development of software that extends the data capsule and the chaining framework including integration of HPC resources and Workset. They will have working knowledge in security, operating systems, networking, high-performance computing, and big data. They will have demonstrated experience in large-scale systems development, and parallel execution frameworks such as MapReduce or Apache Spark and a firm grasp of best practice in software engineering. Additionally, the successful candidate will be able to work independently while being highly competent in a team setting, and potentially supervising graduate or undergraduate students who seek course projects with HTRC. The Research Programmer will work closely with and under direction of PIs, Co-PIs, as well as alongside Ops Manager, Project Developer and other staff.

Project Operations Manager (IUB) (20%)
Project Operations Manager position will be filled by the HTRC Operations Manager (Ops Mgr) position, which is currently open, but will be filled before project start date. Ops Mgr is responsible for the production and development services at HTRC, as well as acting as system architect. HTRC Operations Manager will be a skilled senior systems engineer responsible for overseeing daily operations, system security, and ensuring stability, availability, reliability, and safety of HTRC software and services. The candidate will be comfortable in both a research and operational setting. The candidate will possess technical vision, will be able to work with and through others to achieve the vision, will lead technical and operational efforts, will be effective in resolving conflicts. The person will operate in a matrix reporting setting, reporting ultimately to an Executive Management team of HTRC, so will be a self-starter and highly independently motivated. The candidate should be fluent with web services development, scripting languages, data management technologies, and Web authentication and authorization technologies. In addition, the candidate will participate in the grant process, start to finish, so must be able to express themselves convincingly both orally and in writing. On this project, Ops Mgr will be responsible for integration of all enhancements to HTRC services into current HTRC architecture, along with debugging and testing prior to integration.

LIS Research Assistant (Illinois) (50%)
A 50% FTE PhD Research Assistant will be assigned to this project. The PhD Research Assistant (LIS RA) will be assigned to this project on an hourly basis, i.e., on average 20 hours per week. GSLIS will provide all tuition remission; consistent with Foundation policy, no tuition

remission fees will be charged to the grant. LIS RA will be a CIRSS Affiliate PhD candidate having at least a MS in LIS (or closely related discipline), digital library expertise, some working knowledge of information modeling (e.g., RDF) and prior experience on research projects involving digital information resources. For this project, LISRA will focus on enhancing the Workset, as described in Task 1, along with the redesign and rebuild of Workset Builder, as described in Task 2. The RA will also take on other duties as assigned/deemed necessary by project leadership.

English Research Assistant (Illinois) (33%)
A 33% FTE PhD Research Assistant in the Illinois Department of English will be assigned to this project. The PhD Research Assistant will be assigned to this project on an hourly basis, i.e., on average 20 hours per week. Illinois will provide all tuition remission; consistent with Foundation policy, no tuition remission fees will be charged to the grant. English RA will be either a Master's or PhD student in English or LIS and will be trained in data management and data analysis, with a minimum of a familiarity with spreadsheets, and preferably also with Python or R. Chief duties of this RA will be to assist in Underwood's work in Task 6, chiefly by managing the creation of training data for a predictive model that identifies the page-level boundaries of genres in volumes under copyright.

Hourly Linked Data Specialist (Illinois) (800 hours/year)
An hourly Programmer and Metadata Specialist will be assigned to this project for 800 hours per project year, at a rate of $40/hour. Programmer will have ample experience with HTRC infrastructure and services, as well as a strong working knowledge of and experience with linked data semantics and infrastructures, along with development for HTRC platforms, with likely candidates. This Programmer will primarily focus on work in Task 1 and 2, including technical work on the enhancement of the workset, redesign of Workset Builder as well as help in scaling up both worksets, Workset Builder and the Data Capsule. This programmer will work closely under PIs and Co-PIs and with LIS RA, Operations Manager, IU Research Programmer and other technical staff.

Research Programmer & Systems Administrator (Illinois) (45%)
A 45% FTE Research Programmer & Systems Administrator (RPSA) at Illinois will be assigned to this project. RPSA will provide support on intensive programming efforts on reconstruction of Workset Builder and enrichment of workset metadata. In addition, RSPA will provide maintenance and monitoring of systems and infrastructure at Illinois. RSPA should have at least a Bachelor's in computer science or related field, with Master's preferred. RSPA will be familiar with HTRC systems and services. RSPA will also have multiple years experience working on project-based programming and systems administration work, including large-scale computing and database administration. RSPA will oversee all programming efforts on Workset Builder and linked data.

Systems Administrator (IUB) (30%)
A 30% Systems Administrator (SA) at IUB will be assigned to this project to provide monitoring, implementation and maintenance oversight for HTRC systems at IUB. SA will have multiple years experience as a systems administrator and be familiar with HTRC systems and services. SA will be responsible for helping to oversee operation and integration of Data Capsule enhancements at IUB systems.

Graduate Hourly support (Illinois) (150 hours/year)

Additional GSLIS staff that are candidates for a Masters or PhD in Library and Information Science will be hired/assigned to this project *on an hourly basis*, as needed. Hourly worker will be chosen from MSLIS and PhD students at GSLIS who have familiarity and expertise with HTRC systems. Graduate hourly effort will go toward supporting the work in Tasks 1, 2 and potentially 6. Graduate hourly workers will also tackle needed project tasks as they arise and are assigned by project leadership.

Undergraduate Hourly support (Illinois) (600 hours)
Additional hourly support chosen from undergraduate students in the Department of English will be hired/assigned to this project on an hourly basis, as needed. Hourly workers will need general familiarity with literary history and ability to learn creation process for page-level training data. Undergraduate hourly workers will assist in Underwood's work in Task 6, chiefly by creating page level data and manually verifying tagging outputs in source data.

Post-Doctoral Research Associate (Oxford) (50%)
The University of Oxford e-Research Centre will employ a post-doctoral Research Associate (OxRA1) to work on the implementation and modeling outputs, primarily in Tasks 1 & 2. OxRA1 will be in post for the full length of the project (2 years) at 50% FTE. The position will be filled by a candidate demonstrating knowledge and experience of Semantic Web and Linked Data technologies, their application to Digital Humanities research, and with demonstrable expertise in developing and deploying software solutions. It is expected the post will be filled by an existing member of the Oxford e-Research Centre's Digital Humanities team, who have specific recent experience working with and developing for HathiTrust data and WCSA technologies through the ElEPHãT project.

Research Assistant (Brandeis) (50%)
A 50% FTE PhD Research Assistant (RA) will be assigned to this project on an hourly basis, i.e., on average 20 hours per week. Brandeis will provide all tuition remission; consistent with Foundation policy, no tuition remission fees will be charged to the grant. We require that this student has a Master's degree in computer science or computational linguistics, and they will be chosen from current PhD students in either discipline at Brandeis. RA must have at least three years programming experience and ample, ideally also 3 years, exposure to and experience with natural language processing and web services strongly desired. For this project, Brandeis RA will work on embedding NLP modules in Data Capsule and other pieces described in Task 5, along with general technical/programming support for Pustejovsky and Verhagen.

Hourly Programmer (Waikato) (25%)
An hourly programmer at Waikato will work to integrate Hinze's Capisco semantic tagging system into the HTRC Workset Builder, under supervision from Hinze. This process will include testing and deployment of the service at scale and on the in-copyright portion of the HTDL. The developer/programmer position will have expertise in Java, Ruby on Rails and distributed programming. Prior knowledge in databases is essential, and working knowledge of semantic web techniques is helpful, but not essential. The programmer will be recruited from the CS graduates, and be available throughout the time of the project. They will be responsible for extension and adaptation of Capisco software to be run in the data capsule and disambiguation algorithms. The programmer will work closely with Hinze, who will supervise the work, as well as under direction from the PI, Co-PIs and other technical staff.

4.6 Principal Project Management Roles

This is a collaborative, multi-institutional project involving five universities across three countries, and researchers in Library and Information Science, Computer Science, Linguistics and Literature. With HTRC itself being multi-institutional, this project will utilize existing regular channels of communication and coordination along with scheduled meetings, including weekly HTRC ExMgt calls, with Downie, Plale, Operations Manager and Dubnicek participating, and location-based team meetings for IUB and Illinois HTRC staff. There will be monthly task stakeholder meetings between relevant management personnel and each task leader and Project Coordinator in order to oversee adherence to objectives and schedule. In addition, each PI and co-PI along with the HTRC Ops Manager have specific responsibilities and roles for management on this project, and will work closely with Project Coordinator.

As WCSA+DC PI, Downie will assume ultimate responsibility for the success of the project. He will be responsible for all of the project's financial, administrative and intellectual aspects. Along with Co-PI Cole, he will play a leadership role in Tasks 1 and 2. He will assist Co-PIs Plale and Cole with integrating the outcomes of Tasks 1 and 2 with Tasks 3 and 4. He will act as supervisory liaison with research partners, Underwood and Pustejovsky for Tasks 5 and 6, as well as research partners, Page and Hinze, for Tasks 1 and 2.

As project Co-PI, Cole will focus on the execution and leadership of Tasks 1 and 2, along with supervision of Illinois personnel, in conjunction with Downie. Cole will assume leadership in matters related to metadata, linked data, and open annotation standards.

As Co-PI on this project, Plale will be responsible for the execution and leadership of Tasks 3 and 4, for supervision of IUB personnel, for ensuring IUB's contributions to the other tasks are timely and appropriate, and for ensuring that the HTRC ACS program is well functioning.

The HTRC Operations Manager will serve as the Project Operations Manager and will oversee development of new infrastructure of both Workset Builder and Data Capsule to ensure proper integration and security. They will also help coordinate the development team to keep efforts on schedule and running efficiently.

As Project Coordinator, Dubnicek will coordinate staff and their efforts across institutions; help keep projects on schedule and within scope; track expenditures; help with travel arrangements; and, coordinate reporting efforts. Additionally, Dubnicek will spearhead planning and execution of UnCamp and help coordinate the ACS RFP process, both of which will be significant centers of user engagement. Dubnicek will also manage regular communications between the project team, HTRC Executive Management, and the HTRC Advisory Board. Similarly, he will create and maintain active lines of communications with all four KRPs to ensure that their efforts are fully integrated in to the project.

4.7 Advisors and soliciting user feedback

Due to the integral nature of this project with the mission of HTRC, the HTRC Advisory Board (AB) shall serve as the project advisory board, unifying project leadership and allowing the PI, Co-PIs and project collaborators to draw on the considerable wealth of knowledge present on the HTRC AB. The AB consists of between eight and ten experts across the breadth of the following disciplines: secure and large-scale computing, publishing/industry, digital humanities, computational linguistics, library, pedagogy and legal. To date, all of the below invitees have been contacted and have committed to serving on the Advisory Board:

- **Allan Lu**, Vice President of Research Tools, Services, and Platform, ProQuest
- **Wolfram Horstmann**, University Librarian, Göttingen Library, Project Lead, TextGrid
- **John Towns**, Executive Director for Science and Technology, National Center for Supercomputing Applications (NCSA)
- **Craig Stewart**, Executive Director, Pervasive Technology Institute, Indiana University
- **Stefan Sinclair,** Associate Professor, Department of Languages, Literatures, and Cultures, McGill University, Project Lead, Voyant Tools
- **Nancy Ide**, Professor, Department of Computer Science, Vassar
- **Jennifer Vinopal**, Librarian for Digital Scholarship Initiatives, New York University
- **Claire Stewart**, Associate University Librarian for Research and Learning, University of Minnesota Libraries
- **Matthew Sag**, Professor of Law, Loyola University Chicago
- **Greg Raschke,** HathiTrust Program Steering Committee member, Associate Director for Collections and Scholarly Communication, North Carolina State University

Over the course of the project we will keep Board members apprised of progress and consult with members one-on-one as appropriate to each member's expertise. We will convene a meeting of the Board at the end of project year 1 to present preliminary work and solicit feedback from Board members. To save travel costs, this meeting will be held in conjunction with a major conference or event at which a majority of board members will be in attendance. We also plan to utilize videoconferencing for members who cannot, cost-effectively, make the trip to meet face-to-face.

We anticipate drawing on the wide and varied expertise of board members with regard to both their unique perspectives stemming from their subject areas as well as their placement as leaders in their fields who can gauge trends within their research areas and promote HTRC developments to other researchers in said fields.

In addition to guidance from the Advisory Board, we plan to engage our strong and diverse user community to solicit feedback on enhancements and upgrades to HTRC services. This will happen in two ways: through a special, focused Request for Proposals (RFP) for Advanced Collaborative Support and through HTRC UnCamp. These two activities will be funded from the budget of HTRC.

At 12 months into the project, the PI and Co-PIs will work with HTRC staff and KRPs to craft a targeted ACS RFP for research questions and proposals that can be answered through use of the enhanced workset, Workset Builder and Data Capsule. In this process, we will receive preliminary testing of the services as well as feedback of tools and services and recommendations for changes/inclusions.[40] As was the case with the ACS projects discussed in Section 6, the RFP for the WCSA+DC ACS iteration will be broadly disseminated via mailing lists of cognate disciplinary interest groups, conference handouts, and posting on the HathiTrust

[40] The short-form 2014-15 ACS RFP can be found at: https://www.hathitrust.org/htrc/acs-rfp. The complete RFP PDF is located at: https://goo.gl/zTilG2.

site. WCSA+DC ACS proposals will be competitively reviewed by a committee drawn from members of the HTRC Advisory Board, HTRC ExMgt and the HTRC ACS Group (i.e., the HTRC team that runs the ACS program). This committee will make recommendations to the HTRC co-directors, Downie and Plale, who will make the final determinations. We estimate that 3-5 projects will be supported. HTRC ACS awards are modeled in the form of HTRC staff time. ACS awardees receive access to dedicated HTRC staff to collaborate on the proposed project during the award period. The HTRC ACS staff consists of specialists in computer science and information science fields. They also receive access to computational resources at HTRC.

Additionally, the HTRC convenes UnCamp—a workshop/unconference—every 12-18 months, and will do so within the first 18 project months. At UnCamp, HTRC brings in speakers and presenters who use HTRC tools as well as solicits feedback from the user base on current services, future services to add, and direction of HTRC services based on their needs. Further, UnCamp presents an opportunity for HTRC leadership, staff and partners to train users on HTRC services, and would present an excellent chance to unveil new tools and services in a setting where our users could get support for not only understanding their use personally, but to be able to train their colleagues, as well.

5. Work Plan and Expected Outcomes

5.1 Summary of expected outcomes and benefits

The principal benefit of project will be the ability of scholars to discover, select and analyze materials from both the HathiTrust and external sources regardless of the copyright status of those materials and to do so at scale. By the end of the project, HTRC will be in a position to finally support the user base of scholars that have analytic research questions that can only be answered by exploring large-scale data collections that have been, until now, denied them because of copyright restrictions.

The key outcomes and benefits of the WCSA+DC project are:

1. The deployment of a new Workset Builder tool that enhances search and discovery across the entire HTDL by complementing traditional volume-level bibliographic metadata with new metadata derived from a variety of sources at various levels granularity.
2. The integration of exemplar non-traditional metadata into the worksets model to create worksets that are more useful to DH scholars.
3. The publication of a new formal workset model that can be used by others to build useful workset manipulation tools and to assist developers of new analytic tools.
4. The establishment of access affordances to such external repositories as TCP-EEBO to allow for ingestion of non-HT material into the analytic process.
5. The creation of Linked Open Data resources to help scholars find, select, integrate and disseminate a wider range of data as part of their scholarly analysis life cycle.
6. A new Data Capsule framework that integrates worksets, runs at scale, and does both in a secure, non-consumptive, manner.
7. A set of standard tools that support the scholarly analysis life-cycle that have been selected, adapted and then shown to run safely at scale in the HTRC non-consumptive research environment.

8. A set of exemplar pre-built Data Capsules that incorporate tools commonly used by both the DH and CL communities that scholars can then customize to their specific needs.
9. A set of features extracted from both public domain and in-copyright data that scholars can use for exploration and analysis.
10. Scholarly publications disseminating project results, on both the technical and scholarly aspects of the project, to the digital libraries, DH and CL communities.

5.2 Detailed Task Breakdown

For a detailed summary of chronology, time estimates and personnel allocations for each task, please refer to the Gantt chart in section 5.3. For below listed tasks, institutions in brackets represent the lead institution on the major tasks with personnel in parentheses denoting the lead on the subtask. Detailed information for project personnel can be found in Section 4.

Task 1: Implement Findings from WCSA into Workset Builder [Illinois, Oxford, Waikato]

1. Transform MARC bibliographic data into RDF (Timelines) **(Cole)**
    a. Augment MARC with URIs
    b. Replace text with URIs
2. Augment HTRC knowledge stores **(Sarol)**
    a. Create triple store
    b. Connect triple store to HTRC
3. Incorporate derived and non-traditional metadata at more granular levels **(Downie, Cole, Underwood, Page, Hinze)**
    a. Page-, paragraph-, and/or word-level features
    b. Implement genre tagging algorithm
    c. Implement Capisco concept tagging algorithm, including finer grained tagging, such as paragraph- and section-level tags
    d. Assess quality of tagging outputs through manual verification of source pages for sampled subset of tags and comparing against standard external test corpora
4. Enable creation of worksets with hybrid data types **(Cole, Page)**
    a. Connect HTRC Workset Builder (WB) to external resources
    b. Deploy ability to add non-HT resources to worksets
5. Improve and Redesign WB **(Downie, Cole, Sarol)**
    a. Redesign underlying architecture of WB
    b. Enable workset description export as RDF graphs, import/transformation of HT collections to HTRC worksets, import of worksets and resources in field-specific, external corpora and tools
    c. Adapt Workset Builder to take advantage of RDF & URIs in metadata
    d. Add support for machine-created worksets > 10,000 volumes
6. Addition of semantic search in WB **(Hinze, Downie)**
    a. Adaptation, transfer and deployment of Capisco semantic tagging on HT
    b. Analysis and enhancements for performance and scalability
    c. Software packaging, release and reporting of tagging
7. Solicit feedback from user community **(Downie, Plale, Cole, Dubnicek)**
    a. Release special call for Advanced Collaborative Support (ACS) proposals based on the above additions/enhancements to workset creation
    b. Gather user feedback and suggestions at HTRC UnCamp

Task 2: Extending WCSA Research Ability [Illinois, Oxford]

1. Evolve and extend formal WCSA workset model, as described in CIRSS technical report (Jett, 2015) **(Cole)**
2. Extend current private/public workset access model **(Cole, Plale, Ops Mgr)**
3. Examine ways to accommodate ability to annotate RDF-compatible workset descriptions **(Cole, Sarol)**
4. Develop WB to support common DH research tasks in workset construct **(Downie, Underwood)**
    a. Support citability and interconnectedness of worksets and sub-workset creation
5. Support ability to present and utilize relationships that support scholarly investigation, which may be distinct from those required to construct, maintain and curate worksets themselves **(Page, Cole)**
    a. Model ability to map from descriptive bibliographic structures to event-based ontologies used within historical study (Nurmikko-Fuller et al., 2015)
    b. Model framework for future extensibility driven by scholarly needs
6. Develop WB to support workset creation through volume analysis **(Cole, Page)**
    a. Allow for de-duping and OCR evaluation of volumes in worksets, and intermediate data products from creation and analysis of worksets
    b. Model support to distinguish between curation-derived versioning (e.g. improved OCR of a text) and domain-intrinsic versioning (e.g. editions of a book)
7. Enable the development of Workset constructors, viewers, and data contributors that are tailored to specific investigations or fields of study, but maintain compatibility with the WCSA model and tools **(Page)**
    a. Formalize the process for integration of, or linking with, external complementary corpora (e.g. EEBO-TCP)
    b. Allow for identification of minimal or core terms within the Workset model to enable alignment and patterns
    c. Incorporate tools for creating workset-compatible RDF by external digital libraries
8. Model support of worksets with hybrid data sources, e.g., non-HT sources **(Sarol)**
9. Model versioning and preservation of worksets **(Plale, Downie)**
10. Model curated worksets **(Page, Cole, Sarol)**
    a. Including: OCR correction, header/footer removal, retention of provenance.
11. Integrate WB into the fabric of HTRC services **(Downie, Plale, Cole)**
    a. Enable worksets to be available and understood on Data Capsule (DC)
    b. Allow creation of worksets from DC results

Task 3: Enhancing Data Capsule Support of Researcher Environment [Indiana]

1. Extend Data Capsule to support workset moving in and out **(Plale)**
2. Extend the security model of the Data Capsule to assess trustworthiness of a Workset prior to ingest **(Plale, Ops Mgr)**
    a. Workset is a data object whose trustworthiness cannot be determined only by its classification as a workset, necessitating new security model
3. Build multi-stage analysis in Data Capsule **(Plale, Ops Mgr)**
    a. Develop the process for researchers to define multi-stage analysis outside of a Data Capsule then import the specification with its associated workset and tools
    b. Develop the mechanism by which the multi-stage analysis is invoked from within Data Capsule
    c. Extend the multi-stage analysis framework so that intermediate data outputs can be used by subsequent steps or to create a new workset

    d. Extend the workset model so that it can represent result sets
4. Implement automated results analysis to ensure no leaking of restricted data that is a result from computational analysis **(Plale)**
5. Enhance DC support for custom research environments **(Plale, Ops Mgr)**

Task 4: Getting to Scale Securely and Smoothly [Indiana]

1. Enable secure, user accessible computational analysis of HT corpus at large-scale **(Plale, Ops Mgr)**
    a. Deploy a new tool, R-P n-gram pattern matching tool, for computational analysis of millions of volumes, which allows parallel analysis to a higher level of accuracy than previously seen
    b. Optimize R-P n-gram pattern matching tool so that it can quickly discard irrelevant texts when it does not have a well defined Workset to start
2. Allow for more relatively free-form research investigation process to take place over large-scale sensitive textual data **(Plale)**
    a. Optimize data extraction processing of a Workset that references 1M+ volumes
2. Extend Data Capsules so that from within a scholar's own Data Capsule, computational analysis can be invoked that requires parallel execution on HPC resources located within the HTRC Secure Commons **(Plale, Ops Mgr)**

Task 5: Partner Contribution to Data Capsule, Computational Linguistics [Brandeis]

1. Identify appropriate NLP tools from the LAPPS Grid, and configure for inclusion into DC **(Pustejovsky)**
2. Adapt and train Document Structure Parser and Genre Classifier over English-language book corpus for dates 1922-2000 **(Verhagen, Pustejovsky)**
3. Adapt, train, and tune Sentence Splitter, Tokenizer, and Part-of-Speech Tagger over entire corpus **(Verhagen, Pustejovsky)**
4. Identify appropriate types from corpus for Named Entity Recognizer. Train and tune over corpus **(Verhagen, Pustejovsky)**
5. Train and tune Shallow Parsers over corpus **(Pustejovsky)**
6. Evaluation of results, revise algorithms as needed/desired, and publish findings **(Pustejovsky, Verhagen)**

Task 6: Partner Contribution to Data Capsule, Digital Humanities [Illinois]

1. Extending extracted features to in-copyright content **(Underwood)**
    a. Extract features from English-language books 1922-2000
    b. Develop a page-level training set for the 1922-2000 period
    c. Create a 20th Century fiction workset
2. Integrate page-level predictions into HTRC workset builder **(Underwood, Cole, Downie)**
3. Tune existing NLP workflow for extracting characterization, and incorporate it into a Data Capsule. **(Underwood, Plale, Ops Mgr)**
    a. Improve NLP workflow in Bamman, et. al, 2014 by incorporating insights in Vala, et al. 2015.
    b. Integrate this workflow in a Data Capsule.
4. Run analysis of characterization as use case of new tools **(Underwood, Downie)**
    a. Run DC with NLP workflow in the 20th century fiction workset.
    b. Assemble a dataset of characters and characterizations 1780-2000.
    c. Manually correct a subset of the characterization data, so we have ground truth to assess reliability.

5.3 Schedule of Completion

WCSA+DC Proposed Schedule of Completion

			Year 1												Year 2											
			1 Jan. 2016 - 31 Dec. 2016												1 Jan. 2017 - 31 Dec. 2017											
			ja	fe	mr	ap	my	je	jl	ag	se	oc	no	de	ja	fe	mr	ap	my	je	jl	ag	se	oc	no	de
Category	***Activity***	***Personnel***	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1. Implement Findings from WCSA into Workset Builder	Transform MARC bibliographic data into RDF (Bibframe, Schema, FRBRoo)	TC, JS, KP, ORA	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■						
	Supplement HTRC indexes with triple stores	LDA, JS	■	■	■	■	■	■																		
	Incorporate derived and non-traditional metadata at more granular levels, e.g. genre and concept tagging, page-level and more granular features	TC, KP, TU, AH, WP, JD, LDA, LRA, ORA	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■						
	Enable creation of worksets with hybrid, external data types	LRA, LDA							■	■	■	■	■	■												
	Improve and Redesign WB	TC, JS, LDA, LRA, DH	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■									
	Redesign underlying architecture of WB	LRA				▒	▒	▒	▒	▒	▒	▒	▒	▒												
	Enable workset description export as RDF graphs	LRA				▒	▒	▒	▒																	
	Enable import/transformation of HT collections to HTRC worksets	JS, TC, LDA, LRA								▒	▒	▒	▒	▒												
	Enable import of worksets external repositories	KP, TC, JD, LRA								▒	▒	▒	▒	▒												
	Adapt Workset Builder to take advantage of RDF & URIs in metadata	JS, TC, LDA, LRA				▒	▒	▒	▒	▒	▒	▒	▒	▒												
	Add support for machine-created worksets > 10,000 volumes	LDA								▒	▒	▒	▒	▒	▒	▒	▒									
	Deploy Capsico semantic search in WB	AH, WP, JD	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	Solicit feedback from user community via Advanced Collaborative Support (ACS) and UnCamp	BP, JD, TC, RD															■	■	■	■	■	■	■	■	■	■
2. Extending WCSA Research Ability	Refine, further evolve and publicize formal WCSA workset model	LRA, LDA	■	■	■	■	■	■																		
	Extend current private/public workset access model	LDA, DH, BP				■	■	■	■	■	■	■	■	■												
	Accommodate ability to annotate RDF-compatible workset descriptions	TC, JS, LRA							■	■	■	■	■	■	■	■	■	■	■	■						
	Develop WB to support common DH research tasks in workset construction, e.g., citability, interconnectedness, etc.	JD, TC, TU, LRA, LDA													■	■	■	■	■	■	■	■	■	■	■	■
	Develop WB to support workset creation through volume analysis, e.g. de-duping, intermediate data products, multiple versioning models	TC, LDA, KP, ORA, LRA, DH													■	■	■	■	■	■	■	■	■	■	■	■
	Support relationships via crosswalking models and framework for future extensibility	KP, TC, LDA, ORA, LRA													■	■	■	■	■	■	■	■	■	■	■	■
	Enable Workset constructors, viewers and data contributors tailored to specific investigations/fields of study	KP, TC, LRA, LDA, DH																			■	■	■	■	■	■
	Extend model to support of worksets with hybrid data sources	TC, JS, LRA	■	■	■	■	■	■																		
	Extend model to support versioning and preservation of worksets	BP, JD, LDA							■	■	■	■	■	■	■	■	■									
	Extend model to support curated worksets	KP, TC, LRA, LDA																■	■	■	■	■	■			
	Integrate WB into the fabric of HTRC services	JD, BP, TC, JS, DH, LDA, LRA													■	■	■	■	■	■	■	■	■	■	■	■
3. Enhancing Data Capsule Support of Research Environment	Extend architecture to enable workset to pass secure cell wall of DC	BP, RP, LDA	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■						
	Extend security model of DC to assess trustworthiness of workset	BP, RP	■	■	■	■	■	■	■	■	■	■	■	■												
	Build multi-stage analysis in Data Capsule	BP, RP, LDA	■	■	■	■	■	■	■	■	■	■	■	■												
	Extend DC to import and invoke multi-stage analysis	BP, RP, DH, LDA													▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
	Extend architecture system so that intermediate data outputs can be used to create new worksets	BP, RP, DH, LDA	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒												
	Develop data product model	BP, RP, DH, LDA																			▒	▒	▒	▒	▒	▒
	Build in chaining framework for analysis tasks	BP, RP, DH, LDA																			▒	▒	▒	▒	▒	▒
	Automated results analysis	BP, RP									■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	Enhance DC support for custom research environments	BP, RP, DH									■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	Embed analysis task/tools in workset/DC, as specified by CRE	BP, RP, DH																			■	■	■	■	■	■
	Support Underwood and Pustejovsky in setting up their CREs	BP, RP, DH																			■	■	■	■	■	■
	Implement automated results analysis	BP, RP																			■	■	■	■	■	■
4. Getting to Scale Securely and Smoothly	Scale analytics operations up, e.g., to millions volumes	RP, BP, DH, LDA	■	■	■	■	■	■	■	■	■	■	■	■												
	Deploy R-P n-gram pattern matcher to allows parallel analysis	RP, BP, LDA	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒												
	Ensure safety of current model is maintained within a single VM at scale	RP, BP, LDA	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
	Allow freer-form research investigation process over large-scale restricted data	RP, BP, DH, LDA										■	■	■	■	■	■	■	■	■						
	Optimize data extraction processing of a Workset that references 1M+ volumes	RP, BP, DH, LDA													▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
	Extend DC to allow 1M+ volume executions, from within a scholar's Virtual Machine (VM), to run in parallel on High Performance Computing (HPC) resources within the HTRC Secure Commons	RP, BP, LDA													▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
	Optimize R-P n-gram pattern matching tool to quickly discard irrelevant texts	RP, BP, DH, LDA													▒	▒	▒	▒	▒	▒						
5. Partner Contribution to Data Capsule, Computational Linguistics	Identify appropriate NLP web services from the LAPPS Grid, and configure for inclusion into DC	JP, MV, BGA	■	■	■	■																				
	Adapt and train Document Structure Parser and Genre Classifier over English-language book corpus for dates 1922-2000	MV, BGA					■	■	■	■																
	Adapt, train, and tune Sentence Splitter, Tokenizer, and Part-of-Speech Tagger over entire corpus	MV, BGA									■	■	■	■												
	Identify appropriate types from corpus for Named Entity Recognizer. Train and tune over corpus	MV, BGA													■	■	■	■								
	Train and tune Shallow Parsers over corpus	JP, BGA																	■	■	■					
	Evaluation of results, revise algorithms as needed/desired, and publish findings	JP, MV																				■	■	■	■	■
6. Partner Contribution to Data Capsule, Digital Humanities	Extend extracted features to in-copyright content	TU, ERA, HUG	■	■	■	■																				
	Construct page-level training sets for in-copyright volumes	TU, ERA, HUG	■	■	■	■																				
	Create a 20th Century fiction workset	TU, ERA					■	■																		
	Integrate page-level predictions into WB	TU, TC, JD, JS, LDA, LRA					■	■																		
	Incorporate existing NLP workflow (from Bamman et al.) into DC	TU, BP, DH							■	■	■	■														
	Run analysis in DC with NLP workflow, clean and assess results	TU, ERA											■	■	■	■	■	■	■	■						
7. Administration	Project coordination	RD	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	Reporting	JD, BP, TC, RD												■												■
	Engagement with Advsiory Board	JD, BP, TC, RD	■					■						■						■						■
	Promote improvements in HTRC services	RD, JD, BP, TC										■	■	■									■	■	■	■

JD = J. Stephen Downie, **BP** = Beth Plale, **TC** = Tim Cole, **TU** = Ted Underwood, **JS** = Janina Sarol, **DH** = Dirk Herr-Hoyman, **RP** = Research Programmer (IUB), **RD** = Ryan Dubnicek, **KP** = Kevin Page, **JP**= James Pustejovsky, **MV** = Marc Verhagen, **AH** = Annika Hinze, **WP** = Waikato Programmer, **LDA** = Programmer/Linked Data Architect (Illinois,) **LRA** = LIS Research Assistant (Illinois), **ORA** = Oxford Research Assistant, **ERA** = English Research Assistant (Illinois), **BRA** = Brandeis Research Assistant, **HGA** = Hourly Graduate Assistant (Illinois), **HUG** = Hourly Undergraduate Assistant (Illinois)

6. Synergistic Work

A growing community of scholars in the humanities and social sciences are grappling with questions that require access to large digital libraries. The Canadian Social Sciences and Humanities Research Council has recently funded a seven-year, $1.8 million grant, NovelTM,[41] that unites seventeen public- and private-sector partners to produce a literary history of the novel on a larger scale than hitherto attempted. Both HTRC, led by Downie, and Underwood, as a leading DH scholar, are partners and HTRC's tools and worksets have become a central resource for the project. Non-consumptive research is a high priority for all of these projects, and they will all profit from the expansion of the HTRC Data Capsule outlined in this proposal.

The past two decades have seen a surge in the emergence of new analytics tools that digital humanists can yoke to the aims of their research agendas. Complete tool suites such as Voyant-Tools[42] (Sinclair & Rockwell, 2012; Sinclair & Rockwell, 2015) and WEKA[43] (Holmes et al., 1994; Witten et al., 1999) showcase the variety of analytics processes that the HTRC aims to provide humanists in the future. Directories such as The DiRT Directory[44] provide not only a wealth of information regarding different tools that correspond to functionalities that HTRC's community of humanities scholars desire (Fenlon et al., 2014) but also suggest additional kinds of data contained within the HTDL corpus (e.g., music, maps, etc.) that analytics tools can be applied to.

While this proposal describes integrating the outcomes of two of the WCSA sub-projects (Hinze et al., 2015; Nurmikko-Fuller et al., 2015), lessons learned from the other WCSA sub-projects (Page & Wilcox, 2015; Hinze et al., 2015; Muñoz, 2015; Biggers et al., 2015) are also relevant to the work outlined above. In addition to capitalizing on advancements made during the course of the initial WCSA and DC projects discussed at length above, WCSA+DC is well positioned to leverage related HTRC efforts such as the aforementioned NEH-funded HT+Bookworm project and ACS projects,[45] among others. Current HTRC ACS awards:

1. *"The Trace of Theory" (Geoffrey Rockwell, University of Alberta, Laura Mandell, Texas A&M University, Stefan Sinclair, McGill University, Matthew Wilkens, University of Notre Dame, Susan Brown, University of Guelph)*. The main research question being asked as part of this project is: can we find and track theory, especially literary theory, in texts using computers? This project uses subsetting of the HT corpus and text mining to track theory through its textual traces, and develop tools and computational methods for tracking the concept of "theory."
2. *"Detecting Literary Plagiarisms: The case of Oliver Goldsmith" (Douglas Duhaime, University of Notre Dame)*. A number of recent studies have demonstrated that computational approaches to the study of plagiarism can significantly improve our understanding of literary history. The chief goal of this proposal is to add to the work of researchers studying plagiarism and its identification by developing tools scholars can use to identify other instances of plagiarism and textual reuse within the HTRC's data. To that end, the study seeks to focus on detecting the literary thefts of Oliver Goldsmith, a historian, playwright, scientist, and poet who was one of the most celebrated authors of

[41] http://novel-tm.ca/

[42] http://voyant-tools.org/

[43] http://www.cs.waikato.ac.nz/ml/weka/

[44] http://dirtdirectory.org/

[45] https://www.hathitrust.org/htrc_acs_awards_spring2015

the eighteenth century. Goldsmith serves as a useful test case for the development of a plagiarism detection platform because he famously stole much of his material from other writers.

3. *"Taxonomizing the Texts: Towards Cultural-Scale Models of Full Text" (Colin Allen and Jamie Murdock, Indiana University)*. This project will deploy an improved infrastructure for robust corpus building and modeling tools within the HTRC Data Capsule framework that will enable us to answer research questions requiring large-scale computational experiments on the HTDL. Our research questions depend on the capacity to randomly sample from full text data to train semantic models from large worksets extracted from the HTDL. This project will prototype a system for testing and visualizing topic models using worksets selected according to the Library of Congress Subject Headings (LCSH) hierarchy, and use the same approach to build a showcase project highlighting Thomas Jefferson's historically important "Trist catalog" of 6,487 titles that he sold to the United States Congress in 1815 to re-establish the library after Congress had been destroyed in a fire set by British troops.
4. *"Tracking Technology Diffusion Through Time in the HathiTrust Corpus" (Michelle Alexopoulos, University of Toronto)*. Dr. Alexopoulos' computational economics-based research focuses on analyzing the diffusion of technologies as evidenced in the published record of material of the HathiTrust. Technology diffusion studied in this way could overturn accepted theories about when a technology stopped having an economic and societal impact. She is studying 1000 technologies ("steam engine") is one example, and examining all relevant content in HT where the technologies appear in the print.

The HTRC team is also involved in a research project with Matt Wilkens at Notre Dame University entitled "NER Geolocation for Literary Geography at Scale." Dr. Wilkens has an ACLS Fellowship Award to look at "Geography at Scale" in the HT digital collection. This research will use the Stanford NER (Natural Entity Extraction) software to extract location from the HT full text. The NER software has already been trained to find locations in English, Spanish, German, and Chinese and these represent over half of the HT collection. In one of the first uses of the HT in-copyright works, processing will run the NER using HTRC's HPC (high performance computing) infrastructure in a map-reduce style of processing.

Secure Commons

The Census Bureau has established Census Research Data Centers[46] at 18 locations across the country. Potential researchers can apply to gain access, and are then allowed to conduct research only on machines within the physical research data center. Upon completion of their approved research, the results are manually reviewed before release. Although this procedure arguably has very strong security against data leaks, it is also very restrictive. A researcher must physically travel to one of the 18 locations. Secure Medical Workspace (Shoffner et al., 2013) and CMS Virtual Research Data Center[47] provide virtual workspaces to researchers for work with clinical data. A virtual workspace is backed by a virtual machine equipped with secure software that prevents data leak over some channels. Cloud Terminal runs a thin terminal application on an untrusted operating system to display information while the actual computation logic against data is deployed to a secure cloud environment (Martignoni et al., 2012).

Scholarly Commons

[46] https://www.census.gov/ces/rdcresearch/

[47] http://www.resdac.org/cms-data/request/cms-virtual-research-data-center/

HTRC has developed the Scholarly Commons group—a joint effort between the Universities of Illinois and Indiana to promote training and educational programs in the use of the HTRC text data mining services to HathiTrust institutions—including libraries, digital scholarship centers, and among faculty and students in the classroom. The 2015-2016 academic year is a critical juncture for HTRC as we shift from experimental prototyping toward hardening production services and increasing user engagement and outreach. The HTRC has just hired a Visiting Digital Humanities Specialist with support from Illinois, whose roles and responsibilities include programmatic outreach and instruction. In addition, the HTRC has an award pending with the Institute of Museum and Library Services (IMLS) Laura Bush 21st Century Librarian program that will extend the scope of projected center activities to ensure deep engagement with selected user communities and expand outreach activities to a national scale. The project team will develop a curriculum to be piloted at five academic institutions. Once fully developed, the curriculum will be shared widely through a Train the Trainer program comprising a minimum of 10 events during the third year of the grant. Program attendees will then disseminate the curriculum to their home institution, embedding new services and initiatives at some 200 institutions and providing training opportunities for roughly four million students. By the conclusion of the grant period, the team will also release an instructional webinar and will package the curriculum as an open education resource for use by some 120,000 libraries in the United States. The anticipated start date of the IMLS grant is October 2015. Partner institutions include Northwestern University, Lafayette College, the University of North Carolina, Indiana University and the University of Illinois.

7. Intellectual Property

This project will be subject to the Foundation's intellectual property policy.[48] All software deliverables will be made available to the non-profit educational, scholarly and charitable communities on a royalty-free basis under an open source license allowing free redistribution, derived works, etc.; all pre-existing software that will be embedded in or used to derive deliverables is already made available under appropriate open source license.

All software developed in this project will be licensed under one of the Open Software Initiative (OSI) approved open source licenses.[49] OSI approved licenses have gone through the OSI license review process and are verified to comply with the open source definition.[50]

Reports, presentations and web-posted deliverables will be made freely and openly available to the non-profit educational, scholarly and charitable communities on a royalty-free basis, under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported license[51] permitting non-commercial use and modification.

8. Sustainability

For this project, improvements will be made to HTRC services, which will be utilized by user community, preservation of said services is a given, and essential. HTRC envisions these

[48] http://www.mellon.org/about_foundation/policies/AWMF-IP-October-2011.pdf/at_download/file

[49] http://opensource.org/licenses

[50] http://opensource.org/osd

[51] http://creativecommons.org/licenses/by-nc-sa/3.0/

services to be substantial enough and significantly impactful in research that they will be used for the foreseeable future. Additionally, the intention is to improve the framework and structure beneath the enhanced research capabilities of the services, which will allow them to persist even as they are enhanced, or newer, more varied tools are added.

In addition all reports, scripts, codes and data developed by this project will be maintained on a project website and/or in a project-specific GitHub repository linked from that Website. The project Website will remain operational and publicly accessible through at least 2020.

The open-source licensing of WCSA+DC's products is a key part of our sustainability strategy. Project code and documentation will be made available to the world via the HTRC's web-based code repository. The HTRC (and the digital humanities community) truly need the kinds of processes promised by the prototype projects, and because of this, it is our intention to use and/or further develop the code from the successful prototypes for use in the day-to-day operations of the HTRC. We will also explore with the HathiTrust Board which services might be incorporated into the HTDL maintained at the University of Michigan. Similarly, HTRC will be working with the HathiTrust Board to explore how the Linked Open Data metadata resources might be integrated with the HTDL.

9. Reporting

Since the proposed project will span 24 months, from January 1, 2016 to December 31, 2017, we anticipate the submission of two formal project reports (i.e., one Interim Report to be submitted before the end of March 2017, i.e., within 90 days after the grant project start date 1 year anniversary, and one Project Final Report to be submitted within 90 days of the grant project's end date (31 December 2017). With much of the tasks overlapping project years, the reports will include narrative commentary on the activities, successes and challenges of the project.

The Project Final Report will detail the results of the project tasks and the outcomes of each. Our metric for success will be completion of the deliverables detailed in section 5.1 of this proposal, along with the feedback gathered in our ACS call and from the HTRC user community at UnCamp. Both reports will also discuss grant expenditures for the period covered in conjunction with the official budgetary accounting provided by the University of Illinois grants and contracts accounting office. J. Stephen Downie will prepare the reports in collaboration with the Project Coordinator, co-PIs and Research Assistants. Downie will have the ultimate responsibility for timely completion and submission of these reports.

10. Budget

10.1 Budget Spreadsheet

Please see the attached project budget, in the provided template, in Excel file. Note that there is one consolidated project budget in USD and supplemental budget spreadsheets for both Oxford and Waikato in the provided template for non-US institutions.

Investment Income
The University of Illinois will invest funds in accordance with the investment strategies outlined in Section 14 of our Business Financial Policies and Procedures.[61] Per section 16.1.5 of the same policy, “Interest Earned on Non-Federal Cash Advances - Interest earned on advances made by other sponsors is included in the University's temporary investment pool. If there is an agreement with the sponsor, interest income is added to the project account. Otherwise, this income is distributed on the same basis as other earnings on temporary investments.” Interest generated over the course of this project will be added to the project account and allocated to the project’s budget line for supplies.

12. Bibliography and Related Papers

12.1 Bibliography

Acs, B. et al. (2011). Meandre data-intensive application infrastructure: Extreme scalability for cloud and/or grid computing. In Proceedings of *2011 Int’l Conf on New Frontiers in Artificial Intelligence*. Berlin: Springer-Verlag, pp 233–242.

Auvil, L., Aiden, E. L., Downie, J. S., Schmidt, B., Bhattacharyya, S. & Organisciak, P. (2015). Exploration of Billions of Words of the HathiTrust Corpus with Bookworm: HathiTrust + Bookworm Project. Poster presented at *Digital Humanities 2015 (DH 2015) Conference*, Sydney, Australia. 29 June - 3 July 2015

Bamman, D., Underwood, T. & Smith, N. A. (2014). A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, MD, pp 370-379.

Bell, G. (2009). *Forward to The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research, pp. xiii-xvii.

Biggers, K., Audenaert, N., & Houston, N. M. (2015). VisualPage: Workset Creation through Image Analysis of Document Pages. Final Report for Workset Creation for Scholarly Analysis: Prototyping Project. Champaign, IL: University of Illinois at Urbana-Champaign. Accessible via: http://hdl.handle.net/2142/79020

Borders, K., Zhao, X. & Prakash, A. (2009). Securing sensitive content in a view-only file system. *ACM Workshop on Digital Rights Management*. New York: ACM, pp 27-36.

Fenlon, K., Senseney, M., Green, H., Bhattacharyya, S., Willis, C. & Downie, J. S. (2014). Scholar-built collections: A study of user requirements for research in large-scale digital libraries. In *Proceedings of the 77th ASIS&T Annual Meeting*. Seattle, WA.

Capitanu, B., Underwood, T., Organisciak, P., Bhattacharyya, S., Auvil, L. & Downie, J. S. (2015). Extracted feature dataset from 4.8 million HTDL public domain volumes (0.2). HathiTrust Research Center. doi:10.13012/j8td9v7m.

Gibbs, F. & Owens, T. (2012). Building better digital humanities tools: Toward broader audiences and user-centered designs. *Digital Humanities Quarterly 6*(2). Accessible via: http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html

Hinze, A., Taube-Schock, C., Bainbridge, D., Matamua, R. & Downie, J. S. (2015). Improving Access to Large-scale Digital Libraries Through Semantic-enhanced Search and Disambiguation. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '15)*. New York: ACM, pp 147-156. DOI=10.1145/2756406.2756920

Hinze, A., Taube-Schock, C., Cunningham, S. J. & Bainbridge, D. (2015). Capisco: Semantic Analysis of Documents from the HathiTrust Corpus. Final Report for Workset Creation for Scholarly Analysis: Prototyping Project. Champaign, IL: University of Illinois at Urbana-Champaign. Accessible via: http://hdl.handle.net/2142/79023

Holmes, G., Donkin, A. & Witten, I. H. (1994). Weka: A machine learning workbench. In *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems*. Brisbane, Australia.

Jett, J., Maden, C., Fallaw, C. & Downie, J. S. (2015). Conceptualizing worksets for non-consumptive research. Poster presented at *iConference 2015*, Newport Beach, CA, 24-27 March 2015.

Jett, J. (2015). Modeling worksets in the HathiTrust Research Center: CIRSS Technical Report WCSA0715. Champaign, IL: University of Illinois at Urbana-Champaign. Available via: http://hdl.handle.net/2142/78149

Jean-Louis, L., Zouaq, A., Gagnon, M. & Ensan, F. (2014). An assessment of online semantic annotators for the keyword extraction task. In *PRICAI 2014: Trends in Artificial Intelligence*. Berlin: Springer, pp 548-560.

Kambatla, K., Kollias, G., Kumar, V. & Grama, A. (2014). Trends in big data analysis. *Journal of Parallel & Distributed Computing 74*(7), pp 2561-2573.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T. & Nowak, M. (2007). Quantifying the evolutionary dynamics of language. *Nature 449*, pp 713-716. doi:10.1038/nature06137

Llora, X., Acs, B, Auvil, L., Capitanu, B., Welge, M. E. & Goldberg, D. E. (2008). Meandre: Semantic-driven data-intensive flows in the clouds. In *Proceedings of eScience 2008*, pp 238-245. DOI: 10.1109/eScience.2008.172

Martingnoni, L., Poosankam, P., Zaharia, M., Han, J., McCamant, S., Song, D., Paxson, V., Perrig, A., Shenker, S. & Stoica, I. (2012). Cloud Terminal: Secure Access to Sensitive Applications from Untrusted Systems, In *Proceedings of USENIX ATC '12*, pp 165-176.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. & Lieberman, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science 331*(6014), pp 176-182. DOI: 10.1126/science.1199644

Milne, D. & Witten, I. H. (2013). An open-source toolkit for mining Wikipedia. *Artificial Intelligence 194*, pp 222-239.

Muñoz, T. (2015). Distributed Metadata Correction and Annotation. Final Report for Workset Creation for Scholarly Analysis: Prototyping Project. University of Illinois at Urbana-Champaign. Accessible via: http://hdl.handle.net/2142/79021

Nurmikko-Fuller, T., Page, K., Willcox, P., Jett, J., Maden, C., Cole, T., Fallaw, C., Senseney, M. & Downie, J. S. (2015). Building complex research collections in digital libraries: A survey of ontology implications. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '15)*. New York: ACM.

Page, K. & Willcox, P. (2015). ElEPHãT: Early English Print in the HathiTrust, a Linked Semantic Worksets Prototype. Final Report for Workset Creation for Scholarly Analysis: Prototyping Project. Champaign, IL: University of Illinois at Urbana-Champaign. Accessible via: http://hdl.handle.net/2142/79017

Plale, B., Kouper, I., Goodwell, A., and Suriarachchi, I. (2015). "Trust Threads: Minimal Provenance for Data Publishing and Reuse," in *Big Data is Not a Monolith: Policies, Practices and Problems*, Sugimoto, C., Ekbia, H. and Mattioli, M. Eds., Cambridge, MA: The MIT Press (forthcoming)

Shoffner, M., Owen, P., Mostafa, J., Lamm, B., Wang, X., Schmitt, C. P. & Ahalt, S. C. (2013). The Secure Medical Research Workspace: An IT Infrastructure to Enable Secure Research on Clinical Data. *Clinical and Translational Science 6*(4), pp. 222-225.

Sinclair, S. & Rockwell, G. (2012). Introduction to distant reading techniques with Voyant Tools, multilingual edition. Workshop held at *DH 2012*. 3 February 2012, Hamburg, Germany.

Sinclair, S. & Rockwell, G. (2015). *A Practical Guide to Text Analysis with Voyant Tools*. Accessible via: http://docs.voyant-tools.org/

Underwood, T. (2015). Understanding Genre in a Collection of a Million Volumes, Interim Report. figshare. Available via: http://dx.doi.org/10.6084/m9.figshare.1281251

Underwood, T., Capitanu, B., Organisciak, P., Bhattacharyya, S., Auvil, L., Fallaw, C. & Downie, J. S. (2015). Word Frequencies in English-Language Literature, 1700-1922 (0.2) [Dataset]. HathiTrust Research Center. DOI:10.13012/J8JW8BSJ.

Vala, H., Jurgens, D., Piper, A. & Ruths, D (2015). "Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts." *Proceedings of 2015 Conference for Empirical Methods in Natural Language Processing (EMNLP 2015)*, pp 769-774.
Available: http://www.emnlp2015.org/proceedings/EMNLP/pdf/EMNLP088.pdf

Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G. & Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with Java implementations. In *Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems*, pp. 192–196.

12.2 WCSA Papers Not Cited Above

Fenlon, K., Fallaw, C., Cole, T. & Han, M. J. (2014). A preliminary evaluation of HathiTrust metadata: Assessing the sufficiency of legacy records. Short paper presented at Digital Libraries 2014, London, UK, 8-12 September 2014.

Fenlon, K., Cole, T., Han, M. J., Willis, C. & Fallaw, C. (2014). Rethinking HathiTrust metadata to support workset creation for scholarly analysis. Poster presented at Digital Humanities 2014, Lausanne, Switzerland, 7-12 July 2014.

Green, H. E., Fenlon, K., Senseney, M., Bhattcharyya, S., Willis, C., Organisciak, P. & Downie, J. S. (2014). Using collections and worksets in large-scale corpora: Preliminary findings from the Workset Creation for Scholarly Analysis project. Poster presented at iConference 2014. Berlin, Germany, 4-7 March 2014.

Downie, J. S., Cole, T., Plale, B., Fenlon, K., Wickett, K. & Senseney, M. (2013). The Workset Creation for Scholarly Analysis (WCSA) prototyping project: Background and goals. Chicago Colloquium on Digital Humanities and Computer Science, Chicago, IL, 5-7 December, 2013.

Downie, J. S., Cole, T., Plale, B. & Unsworth, J. (2013). Workset Creation for Scholarly Analysis: Preliminary Research at the HathiTrust Research Center. Poster presented at Japanese Association for Digital Humanities. Kyoto, Japan, 19-21 September 2013.

12.3 Sloan Data Capsule Papers not cited above

Plale, B., Prakash, A. & McDonald, R. (2015). The Data Capsule for Non-Consumptive Research: Final Report. Indiana University: Bloomington, IN. http://hdl.handle.net/2022/19277

Zeng, J., Ruan, G., Crowell, A., Prakash, A., Plale, B. (2014). Cloud computing data capsules for non-consumptive use of texts. *Proceedings of 5th Workshop on Scientific Cloud Computing (ScienceCloud)*, ACM, pp. 9-16, DOI: 10.1145/2608029.2608031

13. PI, Co-PI and Key Research Partner CVs

Brief Curriculum Vitae are attached for the PI, each co-PI and KRPs listed below and in the following order:

- J. Stephen Downie
- Beth A. Plale
- Timothy W. Cole

- Ted Underwood
- Kevin Page
- James Pustejovsky
- Annika Hinze

14. Letters of Support

Letters of support from the following are attached:

- Mike Furlough, HathiTrust Executive Director
- HTRC Advisory Board members:
  - Stefan Sinclair, Professor, Department of Languages, Literatures, and Cultures & Director, McGill Centre for Digital Humanities, McGill University
  - Craig Stewart, Executive Director, Pervasive Technology Institute (PTI) & Associate Dean for Research Technologies, Indiana University
- Rick Van Kooten, Indiana University Vice Provost for Research

15. Appendices

Appendix A – HTRC Security: Measures, Practices and Policies

This document, an agreement between the host institutions of HathiTrust and HTRC, Illinois, Indiana University Bloomington and University of Michigan, details the security measures and policies surrounding the services and data utilized and held by HTRC. This document is currently only internal to three Universities, and we request that it remain private.

Appendix B - "Budget and Financial Report for Non-US Institutions" spreadsheets for University of Oxford and University of Waikato

Appendix C – Documentation for Brandeis University personnel benefits rate changes during project span