I. Executive Summary

Scholars rely on library collections to support their scholarship. Out of these collections, scholars select, organize, and refine the worksets that will answer to their particular research objectives. The requirements for those worksets are becoming increasingly sophisticated and complex, both as humanities scholarship has become more interdisciplinary and as it has become more digital.

The HathiTrust is a repository that centrally collects image and text representations of library holdings digitized by the Google Books project and other mass-digitization efforts. The HathiTrust's computational infrastructure is being built to support large-scale manipulation and preservation of these representations, but it organizes them according to catalog records that were created to enable users to find books in a building or to make high-level generalizations about duplicate holdings across libraries, etc. These catalog records were never meant to support the granularity of sorting and selection or works that scholars now expect, much less page-level or chapter-level sorting and selection out of a corpus of billions of pages.

The ability to slice through a massive corpus consisting of many different library collections, and out of that to construct the precise workset required for a particular scholarly investigation, is the "game changing" potential of the HathiTrust; understanding how to do that is a research problem, and one that is keenly of interest to the HathiTrust Research Center (HTRC), since we believe that scholarship begins with the selection of appropriate resources.

Given the unprecedented size and scope of the HathiTrust corpus—in conjunction with the HTRC's unique computational access to copyrighted materials—we are proposing a project that will engage scholars in designing tools for exploration, location, and analytic grouping of materials so they can routinely conduct computational scholarship at scale, based on meaningful worksets.

"Workset Creation for Scholarly Analysis: Prototyping Project" (WCSA) seeks to address three sets of tightly intertwined research questions regarding 1) enriching the metadata in the HathiTrust corpus, 2) augmenting string-based metadata with URIs to leverage discovery and sharing through external services, and 3) formalizing the notion of collections and worksets in the context of the HathiTrust Research Center. Building upon the model of the Open Annotation Collaboration, the HTRC proposes to release an open, competitive Request for Proposals with the intent to fund four prototyping projects that will build tools for enriching and augmenting metadata for the HathiTrust corpus. Concurrently, the HTRC will work closely with the Center for Informatics Research in Science and Scholarship (CIRSS) to develop and instantiate a set of formal data models that will be used to capture and integrate the outputs of the funded prototyping projects with the larger HathiTrust corpus.

Principal Investigator:	J. Stephen Downie, University of Illinois at Urbana-Champaign
Co-Principal Investigators:	Timothy W. Cole, University of Illinois at Urbana-Champaign Beth Plale, Indiana University at Bloomington
Projected Budget:	$436,525
Time Frame:	July 1, 2013-June 30, 2015

PROPOSAL

WORKSET CREATION FOR SCHOLARLY ANALYSIS: PROTOTYPING PROJECT

SCHOLARLY COMMUNICATIONS AND INFORMATION TECHNOLOGY PROGRAM
ANDREW W. MELLON FOUNDATION
MARCH 15, 2013

Principal Investigator: **J. Stephen Downie**
Professor and Associate Dean for Research
Co-Director HathiTrust Research Center
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Co-Principal Investigators: **Timothy W. Cole**
Mathematics and Digital Content Access Librarian
Professor of Library and Information Science
Professor, University Library
University of Illinois at Urbana-Champaign

Beth Plale
Professor of Computer Science
Co-Director and Chair, HathiTrust Research Center
Director, Data to Insight Center
Managing Director, Pervasive Technology Institute (PTI)
School of Informatics and Computing
Indiana University at Bloomington

Table of Contents

1. Introduction and Motivation

1.1 Introducing the HathiTrust Research Center

The HathiTrust Research Center (HTRC) is the official research branch of the HathiTrust, which is a repository that centrally collects image and text representations of library holdings digitized by the Google Books project and other mass-digitization efforts.[1] The HathiTrust is also positioning itself to be the official digital archive of the world’s most important libraries. There are more than sixty partners in HathiTrust, and membership is open to institutions throughout the world. Table 1 below highlights the magnitude of the ever-growing HathiTrust corpus.

Description	Count
Total Volumes	10,644,397
Public Domain Volumes	3,305,946
Book Titles	5,598,627
Serial Titles	277,216
Pages	3,725,538,950
Disk Memory in Terabytes	477
Linear shelf distance in miles	126
Original material weight in tons	8,649

Table 1. HathiTrust Corpus Descriptive Statistics

The HathiTrust's computational infrastructure is being built to support large-scale manipulation and preservation of these representations, but it organizes them according to catalog records that were created to enable users to find books in a building or to make high-level generalizations about duplicate holdings across libraries, etc. These catalog records were never meant to support the granularity of sorting and selection or works that scholars now expect, much less page-level or chapter-level sorting and selection out of a corpus of billions of pages.

Roughly one-third of the items in the HathiTrust corpus are digital representations of print volumes in the public domain, and approximately two-thirds are digital representations of volumes still in copyright. Scholars using the HathiTrust corpus can currently conduct basic bibliographic searching (title, author, subject, ISBN, publisher, and year of publication) against catalog records and full-text searching across all items in the repository for both public domain works and for those in copyright. Scholars may not, however, view or download the contents of works in copyright, which greatly limits meaningful access to approximately two thirds of the corpus by scholars. In addition, Google restricts large-scale bulk access for their digital representations created by Google from public domain books (~3 million volumes), and scholars may only view and download public domain works on a page-by-page basis. While HathiTrust

[1] See http://www.hathitrust.org, http://www.hathitrust.org/about, and http://www.hathitrust.org/htrc for more information.

will provide access to custom datasets by special arrangement on a case-by-case basis, scholars must be able to identify the specific materials required,[2] which is difficult, and potentially impossible in many instances, due to the limitations of traditional library catalog records.

Over the past eighteen months, the HTRC has been developing models and tools to overcome the limitations of restricted access to the content of items in the repository and help scholars conduct interesting new analyses of works found in the HathiTrust corpus. To maximize accessibility to the entire corpus, the HTRC has been crafting tools to facilitate large-scale analyses under a “non-consumptive research” paradigm. Under this paradigm, analytic algorithms are applied to the restricted data held by HTRC. Once the analyses are run, only results are returned to researchers (occasionally with a “snippet” of contextual text). Thus, restricted material is never directly “consumed” by scholars. Again, the power of the “non-consumptive research” paradigm is limited if scholars cannot identify the works they wish to analyze.

The HTRC is a unique collaboration between the University of Illinois and Indiana University. The HTRC is co-directed by Prof. Beth Plale (Professor of Computer Science and Director of the Data to Insight Center) at Indiana University and Prof. J. Stephen Downie (Professor and Associate Dean for Research at the Graduate School of Library and Information Science) at the University of Illinois. The Indiana branch of the HTRC is administratively located at the Data to Insight Center. The Illinois branch of the HTRC is administratively located in the Graduate School of Library and Information Science (GSLIS) and has strong ongoing connections with the Illinois Informatics Institute (I3) and the University Library.

By co-locating the HTRC at two separate institutions, the center benefits from drawing upon the cyberinfrastructure expertise of the Data to Insight Center and the collection, metadata, and content expertise of the Graduate School of Library and Information Science. Projects conducted through the HTRC may be administered at Illinois or Indiana as the primary institution, depending on the nature of the research and development (see section 4.3 for more detail). For example, given its strengths in cybersecurity and large-scale computation, Professor Plale and her Indiana team are taking the lead on developing the HTRC’s secure non-consumptive computational research platform. Professors Downie and Cole, with their complementary expertise in information retrieval and bibliographic control issues, are co-leading the Illinois team on the development of prototype search tools by modifying the search capabilities of Blacklight (an open-source search tool) to better suit the unique use cases and content associated with the HathiTrust corpus. To ensure continual productive collaboration, the Indiana and Illinois development teams have an all-hands conference call every Monday. The HTRC’s executive board of Downie (Illinois), Plale (Indiana), Unsworth (Brandeis), Namachchivaya (Illinois), McDonald (Indiana) and their support staffs similarly meet each Wednesday.

[2] See http://www.hathitrust.org/data.

1.2 Introducing Workset Creation for Scholarly Analysis

The scholars who wish to utilize the HathiTrust corpus rely on digital representations of library collections to support their scholarship. Out of these collections, scholars must select, organize, and refine the worksets that will answer to their particular research objectives. The requirements for those worksets are becoming increasingly sophisticated and complex, both as humanities scholarship has become more interdisciplinary and as it has become more digital.

The ability to slice through a massive corpus consisting of many different library collections, and out of that to construct the precise workset required for a particular scholarly investigation, is the “game changing” potential of the HathiTrust; understanding how to do that is a research problem, and one that is keenly of interest to the HathiTrust Research Center, since we believe that scholarship begins with the selection of appropriate resources.

Given the unprecedented size and scope of the HathiTrust corpus—in conjunction with the HathiTrust Research Center’s unique computational access to copyrighted materials—we are proposing a project that will engage scholars in designing tools for exploration, location, and analytic grouping of materials so they can routinely conduct computational scholarship at scale, based on meaningful worksets.

In September 2012, the HTRC held its inaugural UnCamp in Bloomington, Indiana. The HTRC UnCamp attracted 130 attendees representing 44 institutions from across the United States, Canada, and Europe. The UnCamp format allowed participants many opportunities to interact directly with the HTRC organizers. These interactions included valuable feedback, suggestions and questions from the participants. During and after the UnCamp, we at HTRC have fielded a set of questions—perhaps, the most common set—that are thematically united around notions of creating worksets. For example, we have received such inquiries as:

- “What materials do you have that pertain to Japan? How many volumes are in Japanese?”
- “What materials do you have that come from, or refer to, New Zealand? Any works in Maori?”
- “How would we gather up all the works that deal with Francis Bacon? How about his contemporaries with whom he worked?”
- “Has anyone already built a definitive set of works to analyze by such authors as Dickens or Shakespeare?”
- “What musical scores are in the corpus? What works contain music notation?”
- “Which works have back of book indexes that I might analyze?”
- “How would I gather works by 16th-century women? By 19th-century men?”
- “Which works are fiction? Which are non-fiction? Which are commentaries? Essays? Poetry? Prose?”

- “How would I gather together all the images of Victorian England?”
- “Which versions of multi-copy works should I use in my experiment? Which has the best OCR?”
- “How do I merge a HathiTrust collection of works and metadata with my set of works and tags and my colleague’s annotations?”
- “How would I gather works similar to those that I currently I have in hand? Can I define different kinds of similarity?”

Much to our collective surprise, these questions continue to defy our ability to provide clear answers with any degree of confidence. Organizing resources in these ways was not anticipated by the library traditions of bibliographic control upon which the HathiTrust corpus is built, though it is perfectly logical given the forms of computational analytics now possible.

2. Concerning Worksets, Collections, and Scholarly Research

The act of bringing together related information from various kinds of collections is an essential element of the research process for humanities scholars (Brogan, 2006; Palmer, 2005). The workset is a type of collection created by scholars for their research. It is specialized to the HathiTrust context and intended to facilitate computational analysis. In many current approaches to information systems that support scholarly research, collections have not received a level of development and tool creation to match the attention given to the individual resources that collections organize. The HathiTrust corpus presents unique opportunities for the development of tools and techniques to conduct humanities research. Providing for the creation and use of worksets based on the corpus will allow a unique level of support for the practices of humanities researchers.

2.1 Scholarly Requirements

The use of electronic resources by humanities scholars has been a focus of a number of recent studies (e.g., Spiro and Segal, 2005; Warwick et al., 2008; Sukovic, 2008; Sukovic, 2011). These studies all found that the use of digitized primary source surrogates is growing in specific sub-domains of digital humanities. Conducted in part for the Bamboo Technology Project from late 2011 to early 2012, a survey of (and follow-up interviews with) a combined total of 86 English, History, and Fine Arts faculty members at 12 universities belonging to the Committee on Institutional Cooperation (CIC) confirmed that reliance on digital primary sources is now commonplace for a majority of faculty sampled (Green et al., 2013).

These studies also show that user expectations are increasingly sophisticated. As the number of digitized primary source surrogates available grows, so too do the requirements of scholarly users. Humanities scholars continue to emphasize the need for improvements in discovery and searching, but now expect sophisticated full-text searching to be integrated with more traditional bibliographic metadata-based search and discovery. They also are asking for functionality beyond simple search and discovery.

> The move from creating collections of sheer mass to considering how users access collections and what they want to be able to do with the collections is of primary importance, as users demand greater functionality and reliability from digital collections on which their research in increasingly based. Simply having access to collections of text is not enough to meet the needs of humanities scholars, and they desire functionalities that enable them to delve deeper into the material (Green et al., 2013).

Of particular interest to us is an emerging scholarly requirement in some domains to be able to gather together (e.g., in a kind of personal digital carrel)[3] subsets of texts amenable to in depth forms of analysis using advanced tools and services. In his 2006 paper, "The (Digital) Library Environment: Ten Years After", Lorcan Dempsey remarked on this natural evolution in research practice as the availability of digital resources grow.

> [R]esources need to be accessible to manipulation, to be locally managed, and to be recombined and transformed in various ways. We need to be able to pull disparate resources into custom collections. [....] We do not currently have a widely used 'service composition framework' which allows users to pull together resources easily in a work environment (Dempsey, 2006).

More recently, projects like MONK[4] have demonstrated the power of emerging text analysis tools (e.g., SEASR[5]), the importance of Dempsey's *custom collection* even when working with only a modest corpus (e.g., the MONK Workset concept), and the value of new ways to discover and cluster texts (e.g., TeksTale clustering, Flamenco faceted browsing, experimentation with a search-by-example toolset).[6]

The definitions of Dempsey's *custom collection* or Mueller's *digital carrel* as approaches to supporting emerging scholarly requirements need further clarification (we discuss the general nature of scholarly collections in relation worksets – HTRC's term for the type of collection Dempsy and Mueller describe – in section 2.2), but nonetheless the requirement appears to be real and urgent. As discussed in section 2.3, traditional library descriptive practices and the MARC (MAchine Readable Cataloging) record, originally designed around the administration and daily use of library print collections, are inadequate on their own to fulfill all the needs of

[3] Many academic libraries today still provide scholars with in-library carrels, space in proximity to library collections where the scholar can maintain and work with a subset of books needed to pursue a current research interest. In *Towards a Digital Carrel: A Report about Corpus Query Tools* (documenting the outcomes of two days of conversation among a group of humanities faculty, librarians, and information technologists, November 22-23, 2010 in Evanston, Illinois), Martin Mueller proposes by analogy *digital carrels* for scholars working with a large digital library corpus such as the HathiTrust corpus (http://panini.northwestern.edu/mmueller/corpusquerytools.pdf).

[4] See http://monkproject.org.

[5] See http://seasr.org.

[6] See http://www.monkproject.org/MONKProjectFinalReport.pdf; and http://www.bu.edu/dioa/2009/06/23/dh09-tuesday-session-3-use-cases-driving-the-tool-development-in-the-monk-project.

scholars attempting to identify, select and gather together for analysis relevant texts from a large corpus like that represented by the HathiTrust.

Full-text search goes part of the way toward compensating for insufficiently rich metadata, but simple full-text search of book-length resources tends to be imprecise and limited in a variety of ways[7] (Beall, 2008). To augment current functionality in ways that will best serve humanities scholars, richer metadata that go beyond basic bibliographic attributes and do a better job integrating resources into the Linked Data Cloud and similar Webgraphs[8] are required. While librarians are naturally inclined to take the lead on this, end-user scholarly input is also required. In his introduction to the Council on Library and Information Resources (CLIR) publication entitled, *The Idea of Order*, Chuck Henry anticipates this situation and at the same time notes the need for scholars’ input in addressing these issues. "While a greater reliance and dependency on digital resources is inevitable, the quality of the data and their organization and accessibility in service to teaching and scholarship are major concerns. Without the guiding voice of scholars, the tremendous effort now being devoted to digitizing our cultural heritage could in fact impede, not facilitate, future research." (CLIR, 2010, p. 3).

2.2 From Scholarly Collections to Worksets for Analysis

The term *collection* is used in many different ways in a variety of contexts. Scholars commonly think of collections as aggregations that contain some number of members (e.g., books, images, manuscripts, etc.) that have been brought together to serve some purpose, often to aid in the stewardship of those members, or to serve some informational purpose in the context of a scholarly activity.

However, characterizing what collections are and how they serve scholarly purposes has generated lively debate concerning their role in the design of digital library and aggregation systems. There are four contexts in which collections appear that are of particular interest for HTRC: an institutional curation context, an archival context, a referential (or virtual) context, and a thematic research context. These contexts for collections are not necessarily disjoint; a collection that participates in thematic research may also be created referentially.

> **Institutional curation context.** Museum exhibits, special collections, archives, and general library collections are produced and maintained by librarians, archivists, and curators. Some of these collections are more precisely understood as aggregations,

[7] See http://www.guild2910.org/searching.htm.

[8] There are various definitions of *Webgraph*, *Linked Data Cloud* and other popular names for graph-based models of the World Wide Web (in its entirety or in part). These definitions share in common the idea that Websites or individual Web pages can be viewed as vertices in a graph, linked one to another directionally along the edges of the graph. The graph representation of Web-accessible information resources allows for the application of graph theory techniques and methods in support of better knowledge representation and management. The better and more completely linked a resource is, the better it is represented in a Webgraph, and the better and more complete any reasoning or analysis done over the Webgraph is. There exist several Webgraph snapshots used in research. See, for example, http://web-graph.org/index.php/webgraphproperties, and http://snap.stanford.edu/data/index.html#web.

which provide a single point of access to independent, dispersed collections, allowing scholars to find and retrieve items from a wide variety of sources. HathiTrust is one such “expansive gateway” (Palmer, 2004). Institutional collections, be they digital or physical, are intentionally created wholes (Currall et al., 2004) that provide evidence for inquiry (Buckland, 1999). This is particularly true for aggregations or collections of primary source materials, as in the case of HathiTrust. For the humanities researcher, library and archival collections are highly important as coherent, dense units for exploration and study (Brogan, 2006; Palmer, 2005). Indeed, collections are often sufficiently valuable to constitute institutional capital, shown even to exert pull on scholars to visit or take positions at collecting institutions (Brockman et al., 2001).

Archival context. The practice of creating of a comprehensive record of an organization or a person’s life by systematically gathering documentary evidence of that person’s or organization’s activities is fundamental to how collections are viewed in an archival context (Hensen, 1989). Traditionally these collections contained paper records and other physical objects such as photographs. However, as the means of communication and interaction have shifted into the electronic realm, the correct means for defining and developing collections in an archival context have become an issue of debate (Yeo, 2012). Despite this uncertainty, it is clear that these collections are created around a “unifying characteristic”[9]. As digital library and aggregation systems evolve to allow the flexible creation and use of collections, it will essential to support the full representation of these unifying characteristics, either with structured descriptions or by linking with RDF (particularly for collections organized around a single person or organization).

Referential context. In the course of conducting research around a topic, a scholar may access materials that are held by a variety of archives or libraries. In order to develop a research collection that gathers together these relevant items, the scholar may create a list that specifies the locations and other salient details of those items. The creation of this kind of collection does not imply that the creator of the collection has taken over ownership or custodianship of the items gathered into the collection. Digital library and aggregation systems can provide scholars the functionality to create this kind of collection, and to feed some selected items into computational processes for analysis. We note that reference to particular items and their properties (as opposed to simply specifying a query or a retrieval set as suggested by Lagoze and Fielding in 1998) is essential for the creation of a resource with lasting scholarly value that can be sustained over time, since the composition of the underlying repository may shift over time and it is critical to know which items were included in the collection at the time of analysis. The HathiTrust website provides a “Collection Builder” that allows manual creation of referential collection[10]. This tool relies on manual search over MARC records and does not support detailed structured description of the resulting collections. Therefore the current HathiTrust collection building tool has a limited functionality that does not scale

[9] See http://www2.archivists.org/glossary/terms/c/collection.
[10] See http://babel.hathitrust.org/cgi/mb?a=listcs;colltype=pub#all.

over the entire corpus and does not support the integration of collections into analytical processes. This is the gap the workset, a particular kind of collection, aims to fill.

Thematic research context. In the course of their work, researchers create their own “digital aggregations of primary sources and related materials that support research on a theme” (Palmer, 2004). These collections serve as laboratories for humanities scholars: “thematic collections concentrating on contextual mass and activity support are coming closest to creating a laboratory environment where the day-to-day work of scholars can be performed” (Palmer, 2004). Thematic research collections demonstrate the value of coherent aggregation of heterogeneous but thematically associated content. They serve as platforms for interdisciplinary research and function as tools to support the numerous activities of scholars (from information-seeking to interpretation and analysis of sources). Thematic research collections may be considered a distinct genre of scholarly output. Scholars pull books or pieces of books from many different sections of a library, evidence from rare book rooms or special collections, and secondary sources from a variety of journals in different digital libraries. In this way, their assembly transcends institutional limits, including traditional library or archival organizational structures, in order to impose a new and personal, purposeful order on sets of resources.

Thematic research collections, of which the workset is one type, are curated subsets of a corpus or collection. Scholars may gather items together for any number of reasons.[11] For example, a collection may be based on:

- Specific authorship
- General characteristics of authors (e.g. male v. female; country of origin; era).
- Periods of time, sometimes defined in relation to historical events.
- Properties or features of the texts as wholes, or parts of texts.
- Intertextual relationships: “Allusive practices, subconscious echoes, deliberate imitation, or plain theft” (Mueller, 2010).
- Similarity, as measured in any of a number of ways.

Often, a digital research collection is (to the extent feasible) an amalgamation of heterogeneous sources. Conceptually, these sources may include primary evidence, secondary literature and annotations, data, or metadata. Technically, they comprise a vast variety of media and formats, which are consistently in flux. When gathered together, these sources function as a coherent collection of interwoven content and context. The HathiTrust corpus has the potential to serve as a basis for this kind of collection. Not just with its primary constituents (books), but with bibliographic metadata and even intra-book content, such as formal sections, captioned images, maps and charts and indexes, HathiTrust serves as an expansive aggregation of distributed sources from which related sources may be concentrated by researchers into densely thematic bodies of evidence.

[11] The following reasons are generalizations of observations in Mueller, 2010.

To be useful for computational analysis, such a collection must be expressible as movable, manipulable, eminently machine-processable data. Thus, from a research collection may be derived a workset for advanced analysis. The HTRC workset is a kind of referential, thematic research collection, gathered by researchers from the larger corpus to enable computational analysis. The demand for workset-creation facilities in the HathiTrust, specifically, is proven. At the first annual HTRC UnCamp in September 2012, users called for more ways to interface directly with the data, including ways to collect relevant sources together, prior to processing. Because scholars gather a range of things from a range of places (in fact, from all over the web), HTRC worksets should have the capacity to integrate data from external sources, including linked data sets; metadata about cultural heritage resources at archives, museums, or libraries elsewhere; bibliographic metadata elsewhere; reference resources such as gazetteers and thesauri; secondary literature; and more. Figure 1 shows the imagined relationships involved in the construction of worksets as specialized research collections.

In order for worksets created within the HTRC to act as sustainable scholarly resources over time, it is necessary to provide scholars the ability to develop descriptions of those worksets. A description includes the purpose for which the workset was created, the methods for selection and evaluation of membership in the workset, the formats and other technical aspects of the items, or links to the analytical results of processes run over worksets. These descriptions contribute substantially to usefulness of the collection for the originating scholar or for other scholars that may seek access to the collection itself as a resource.

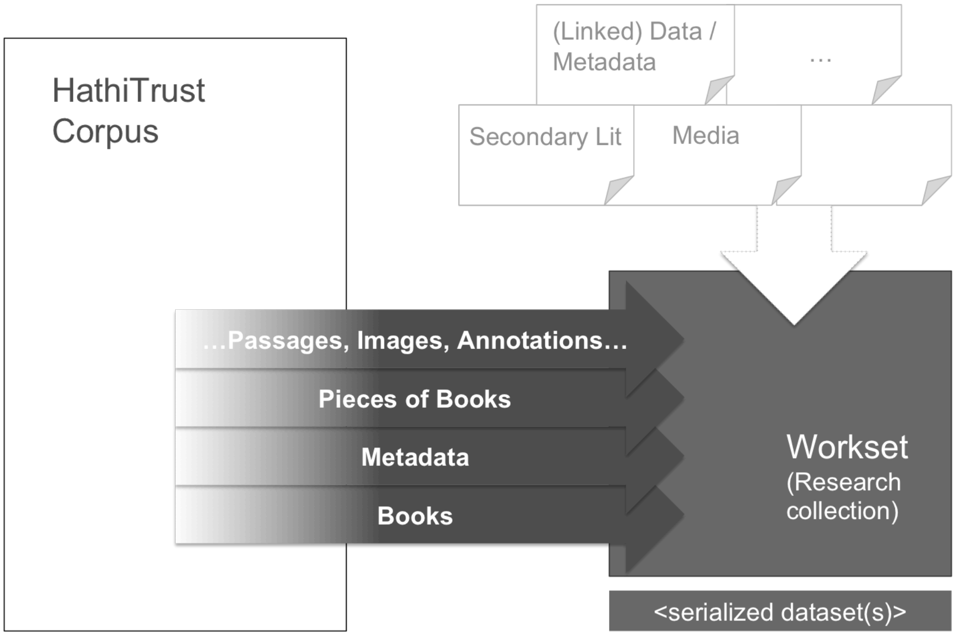

Figure 1. Imagining an HTRC Workset

2.3 Traditional Descriptive Practices & the Limitations of MARC

Libraries have a long tradition of collecting, managing and preserving information. Key to libraries being able to carry out these missions is the proper bibliographic control of the information resources they curate. Bibliographic control in the library relies on established traditions and conventions of bibliographic description. Since the 1960's the MARC (MAchine-Readable Cataloging) record has been the preferred carrier for bibliographic descriptions created and used in libraries. The MARC format in its modern serializations (e.g., MARCXML) remains a good way (in the specific context of library operations) to maintain bibliographic control over print-based library collections. However, as discussed above, scholarly requirements to define and gather items into functional worksets for analysis are challenging. MARC-based metadata alone is inadequate to fully meet scholarly requirements. This is unlikely to change. As noted more than 20 years ago by Michael Gorman, the structure of the MARC record itself, and the ways the format has come to be used by library catalogers, constrains and to some degree defines the scope and utility of library bibliographic descriptions. "The truth of the matter is that one cannot think about any aspect of cataloguing, except at the most rarified and abstract level, without taking the effects of the MARC record into account." (Gorman, 1990, p. 63).

The volumes that comprise the HathiTrust corpus are currently described exclusively by MARC records imported from library catalogs. The sparse content of these records -- most derived from original cataloging that predates MARC and in some cases even AACR2, and the inherent limitations of the MARC format, constrain a scholar's ability to discover, identify and select relevant items and to extract from a large corpus the custom scholarly research collections of digitized volumes from which to derive useful worksets to support analysis and advanced scholarly endeavors. The limitations of MARC as the sole component of item-level metadata in HathiTrust are two-fold.

First, the reliance on MARC means that there is no way to record many resource attributes and properties of interest to scholars. For instance, MARC21 records do not express a book author's gender, nationality, religion or social relationships at the time a book was published. Some mechanisms are provided in MARC, e.g., the 240 field (Uniform Title), to support limited forms of linking between editions of a work, but these mechanisms are idiosyncratic and rudimentary by today's standards. The limitations of MARC in these ways, in both traditional and now digital settings, has been a growing concern since at least the 1998 IFLA FRBR study.[12] Given the findings of the FRBR study and the growing importance of digitized and born digital resources, the limitations of MARC have spurred a number of more forward-looking initiatives, including the creation of the Metadata Object Description Schema (MODS),[13] the development of the Resource Description and Access (RDA) guidelines,[14] and the Library of Congress Bibliographic Framework Initiative (BIBFRAME).[15]

Second, given present-day economic pressures and the natural priority given to library operational needs, almost to the exclusion of all else, catalogers do not make maximum use of what MARC does have to offer. While cataloging supports core library operations and inventory management, current practice does not create records that adequately support advanced

[12] See http://www.ifla.org/publications/functional-requirements-for-bibliographic-records.
[13] See http://www.loc.gov/standards/mods.
[14] See http://www.rda-jsc.org/rda.html.
[15] See http://www.loc.gov/marc/transition.

scholarly needs. In academic libraries the dominant MARC-based applications are Integrated Library Systems (ILS). These systems, and in particular the Online Public Access Catalog (OPAC) components of these systems, have not evolved fast enough to keep up with the growing requirements of scholarly users. The MARC21 specification defines over 200 distinct 'tags,' i.e., potential data entry fields. Adding subfields, there are more than 1,700 fields and subfields available in MARC for description. However, most OPACs make use of only a handful of these fields and sub-fields. Multiple analyses of tens of millions of MARC records contained in OCLC (e.g., Moen, 2003; Moen, 2005; a 2008 OCLC Research blog entry;[16] Smith-Yoshimura, 2010) have shown empirically that catalogers typically use only a tiny fraction of the available MARC fields and subfields.

By far the most heavily used fields in cataloging books and similar materials relate to title, imprint (publication information), physical description and authorship. There is evidence that the number of fields being used on a routine basis is actually diminishing. Karen Smith-Yoshimura summarizing her 2009 analysis of 145.7 million OCLC records reports that, "Although Moen’s study showed that there were 17 fields in books, pamphlets, and printed sheets that accounted for 80% of occurrences in WorldCat in 2005, not including system-supplied fields, in our current analysis there are just four: fixed-length data elements, title, imprint statement, and physical description (008, 245, 260, 300)" (Smith-Yoshimura 2010, p. 21).

This metadata sparseness is not entirely surprising; the inclusion of only basic bibliographic information is in keeping with Library of Congress minimal cataloging best practices.[17] The good news is that MARC records do provide accurate basic bibliographic information. These data are adequate for library circulation, inventory management and most basic, known-item searching tasks. But they are not good enough for much else. Citing an earlier OCLC study (DeRosa 2005), Simth-Youshimura concludes, "Libraries rely on MARC data for library inventory control, but users do their discovery elsewhere" (Smith-Yoshimura, 2010, p. 14).

The current state of affairs in library cataloging results in inconsistent and incomplete (for many purposes) records both in the local systems and in the union cataloging systems, including in the de facto union catalog of the HathiTrust. Minimal cataloging practices, variations in cataloging record quality, and inconsistencies in the use of controlled terms lead to sparse records lacking potentially useful information, which in turn greatly impedes optimal user services (Denton & Coysh, 2011). This means that even when MARC records could contain metadata useful for workset creation, they often do not.

Table 2 shows the frequency with which specific MARC tags are used within a current sample of more than 290 million OCLC records. Fields meant to contain classification numbers, subject indexing, genre information -- fields that could potentially enable additional functionality desired by would-be users of the HathiTrust -- are too sparsely populated to be useful. The situation is typically not any better for specific sub-groups of MARC records. Table 3 shows frequency of field use for a set of 2,386 MARC records describing 19th Century British novels digitized from

[16] See http://hangingtogether.org/wp-trackback.php?p=393; an analysis of just over 96 million MARC records.
[17] See http://www.loc.gov/marc/bibliographic/bdapndxc.html#book.

the University of Illinois library collection. Again, useful information such as genre, which could have been encoded in these MARC records, simply was not with only rare exceptions.

MARC Field	**Percent of records in OCLC having instance of this field**
245 Title Statement	> 99%
260 Publication Distribution, etc.	92%
500 General Note	41%
650 Topical Term / 653 Index Term -- Uncontrolled	39% / 13%
050 LC Classification No / 082 Dewey Classification No	17% / 13%
655 Index Term -- Genre Form	12%

Table 2. Frequency of MARC fields in OCLC Records

MARC Field	**Percent of British Novel MARC records having instance of this field**
650 Topical Term	6%
050 LC Classification No / 082 Dewey Classification No	27% / 4%
655 Index Term -- Genre Form	5%

Table 3. Frequency of MARC fields used in 2,386 descriptions of 19th century British novels

3. Statement of Research Problem and Project Description

Our proposed “Workset Creation for Scholarly Analysis: Prototyping Project” (WCSA) explores three sets of tightly intertwined questions:

> **Question #1.** Can we enrich the HathiTrust corpus metadata by distilling analytics over full text? The MARC records for HathiTrust content are sparse and contain errors. Could we deploy/modify off-the-shelf tools, for example, to confirm or determine language(s) of the text, temporal coverage, spatial coverage, etc.? Perhaps, topic modeling may even be possible to add or augment subject headings, though this is a bit more speculative.
>
> **Question #2.** Can we augment string-based metadata with URIs for recognized entities – e.g., names, subjects, publication location, etc.? If so, HTRC could leverage other services to facilitate discovery and sharing. Such linkages would also create enhanced integration capability as scholars could link out of, and into, the HathiTrust universe (e.g.,

to get useful contextual information, third-party metadata, discover HathiTrust content using third-party services, etc.)

Question #3. Can we formalize the notion of collections and worksets in the HTRC context? What are the necessary elements of a “collection”? What are the necessary elements of a “workset”? How can we balance rigor with extensibility and flexibility? What roles do “data”, “metadata”, “annotations”, “tags”, “feature sets”, and so on, all play in the conception, creation, use and reuse of collections and worksets?

3.1 Prototyping Projects for Metadata Enrichment and Augmentation

To answer Questions #1 and #2, we propose an approach modeled after the successful Mellon-funded Open Annotation Collaboration (OAC).[18] Like the OAC, WCSA will be a collaborative initiative located at the Center for Informatics Research in Science and Scholarship (CIRSS)[19] and the University Library at the University of Illinois at Urbana-Champaign. The project will be led by Profs. J. Stephen Downie (PI), Timothy Cole (Co-PI), and Beth Plale (Co-PI). WCSA will select four prototyping projects through an open, competitive, RFP-based process to build tools relating to metadata enrichment and augmentation. Each prototyping project selected via the RFP (a draft of which is included as Appendix C) will be funded at $40,000 for a 9-month performance period. This approach is designed to increase awareness of issues surrounding workset creation, uncover new techniques, and deliver prototypes that will enhance the value of the HathiTrust corpus. It will also foster interactions among the HTRC, developers, and researchers. Through the RFP framework, we also hope to establish long-term collaborations among participating institutions and the HTRC. Ultimately, these interactions will enhance the value of the HathiTrust corpus and the HTRC as scholarly resources.

3.1.1 Advance Preparation for Releasing an RFP and Selecting Prototyping Projects

We expect the RFP to attract respondents developing algorithms and new techniques for in-depth text mining and topic modeling similar to the work being conducted, for example, by the Mellon-funded Proteus Project at the Center for Intelligent Information Retrieval (CIIR) at UMass Amherst.[20] CIRSS will serve as the administrative locus throughout the prototyping phase, evaluating project proposals, selecting candidates, coordinating project activities, and overseeing progress toward completion. Prior to releasing the RFP, CIRSS will revisit and expand upon lessons learned from its seminal *Google Digital Humanities Awards recipient interviews report*. This report was first conducted while establishing the HTRC to identify problems scholars encounter while conducting research in the digital humanities (Varvel & Thomer, 2011). To further understand workset-creation issues among our constituents, CIRSS will engage the digital humanities community at the annual Digital Humanities conference (July 16-19, 2013). CIRSS will also engage the digital libraries community at the Joint Conference on

[18] See http://www.openannotation.org.

[19] See http://cirss.lis.illinois.edu.

[20] See http://ciir.cs.umass.edu/index.html. As discussed in section 4.4, Prof. R. Mamnatha of the Proteus Project will be serving on the WCSA Advisory Board.

Digital Libraries (July 22-26, 2013).[21] Finally, CIRSS will lead a dedicated workset creation track at the second annual HTRC UnCamp scheduled for September 2013 at the University of Illinois.

WCSA will begin in July 2013, and the CIRSS team will launch the project by leading an initial evaluation of the HathiTrust corpus to:

- Gain a better understanding of corpus coverage (e.g., topical, geographic, temporal) and current metadata constraints to support scholars conducting research with the HathiTrust corpus and identify which aspects of the HathiTrust metadata records require further enrichment;
- Evaluate the corpus for high potential areas that might represent a match between strong coverage in the corpus and scholarly communities that have expressed interest and readiness to engage in computational research; and
- Create a representative sample of 100,000 volumes from the larger HathiTrust corpus (based on the analysis and evaluation of corpus coverage) to be maintained at the University of Illinois and used for testing the tools developed by the prototyping project awardees (see section 3.1.2 for more information).

Formalizing our understanding of the HathiTrust corpus is integral to all phases of project development. The prototyping projects solicited through the RFP process will focus primarily on improving the item-level metadata provided by catalog records. Scholars working within the context of the non-consumptive paradigm must rely on metadata descriptions to determine whether any given item fits within their determined collection criteria, and as described in Section 2, the item-level metadata currently provided for the HathiTrust corpus is insufficient for scholarly evaluation. By leveraging CIRSS’ expertise in providing access to large-scale digital library aggregations, normalizing aggregated metadata created in diverse contexts, evaluating the relationship between item-level and collection-level metadata (Wickett, 2012), and assessing how scholars use item and collection metadata to evaluate resources (Palmer, Zavalina, & Fenlon, 2010), the HTRC will be well situated to identify the types of information a scholar would need to determine the appropriateness of an item for a given collection. Applying CIRSS’ expertise in collection evaluation to the HathiTrust corpus will also allow the HTRC to identify topical strengths within the corpus, thus informing the HTRC’s ongoing strategies for targeted community outreach to scholars whose projects would successfully demonstrate the value and utility of non-consumptive computational research. The CIRSS project team will present preliminary findings at the second annual HTRC UnCamp in September 2013 and produce a final technical report in January 2014 outlining corpus strengths and identifying potential audiences for strategic community building.

The RFP will be released to the public in November 2013 with submissions due in mid-January. We will disseminate the RFP on the HathiTrust homepage, to UnCamp attendees from 2012 and 2013 and select digital library and digital humanities listervs (e.g., DigLib, JESSE, Code4Lib, DLF-Announce, ACRL-DH, ALA IGDC, Humanist, H-Net Announce, and CenterNet). In late January 2014, the project team will review proposals responding to the RFP and

[21] A tutorial proposal on collections modeling has been submitted to JCDL 2013.

generate a shortlist of prospective awardees. One representative from each project on the shortlist will be invited to present their proposal at a WCSA meeting in late February 2014. Afterward, four projects will be chosen from the shortlist, and prototyping projects will begin in April 2014. At the conclusion of the development period in January 2015, the projects will reconvene for a final Prototype Demonstration Meeting to present project prototypes, outputs, and related deliverables. This will also be an opportunity to discuss results and develop recommendations for next steps.

3.1.2 Technical Infrastructure for Prototyping Projects

The production infrastructure of the HTRC is under ongoing development and extension at Indiana University. We do not intend to directly fund any production HTRC infrastructure development at Indiana through the WCSA project. We will, however, encourage and help mediate interactions between successful respondents and the HTRC technical team throughout the prototyping phase to ensure that project deliverables can be successfully implemented within the HTRC environment. We anticipate that outcomes from the WCSA project will influence ongoing development of the HTRC production infrastructure and that there may be opportunities for subsequent focused production infrastructure that will arise from the WCSA prototyping activities.

Respondents’ day-to-day work with the production HTRC environment at Indiana University will be minimal, but the WCSA project research programmer will be responsible for assisting and facilitating respondent work with the HTRC testbed infrastructure sandbox at Illinois. Not only will this allow respondents to experience clones of core elements of the HTRC infrastructure without risk to production services, it also will allow Illinois and respondent teams to collaboratively experiment with modifications and extensions of these core infrastructure elements that may be necessary, or at least helpful, to achieving enhanced workset creation functionality. Specifically, Illinois will make available to respondent teams:[22]

1. **A representative sample of the HathiTrust corpus**[23] for downloading and use in their own environment (i.e., custom dataset to be built for this project). This will allow respondents to have ample test materials for early tool design and testing. Respondents will also have access to the complete Open Content Alliance dataset and can requests additional datasets directly from HathiTrust, but a common, project-specific dataset will facilitate comparison of results and collaboration, both between Illinois and respondents and among respondents.
2. **Access to metadata records for all public domain derived digitized volumes in HathiTrust** (~2.5 million). This includes:

[22] This list is not meant to be comprehensive since additional support requirements may arise.

[23] More than 10,000 and less than 100,000 volumes selected primarily from volumes in HathiTrust originally digitized by the Open Content Alliance (OCA), but including some Google-digitized volumes created from public domain editions as approved by HathiTrust.

a. The standard version of MARC, MODs, and Dublin Core bibliographic metadata files available from HathiTrust;[24]

b. The Metatadata Encoding and Transmission Standard (METS)[25] structural and bibliographic metadata files available one by one through the HathiTrust Data API; and

c. Alternate METS files available initially through an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[26] service exclusively for the WCSA project. These alternative METS files will include added descriptive metadata nodes (i.e., added <dmdSec> nodes) to convey MODS and TEI header bibliographic metadata (i.e., in addition to MARC metadata).

Over the course of this project, we anticipate that the WCSA METS records, or at least some subset, will be enriched by results from respondent projects (e.g., with links, additional metadata properties and values, annotations). We also anticipate exploring ResourceSync with WCSA METS records as one way to facilitate updating and enriching of HathiTrust metadata.

3. **Remote login to a WCSA-specific Unix-based development environment managed and controlled by Illinois.** This environment will allow respondent teams access to public domain derived Google digitized volumes in the Illinois sandbox, initially through a clone of the HTRC API. Through this environment, teams will be able to explore the HTRC API and collaborate with Illinois on enhancements and extensions needed to support item and collection descriptions that better support workset creation. It will also provide confidence that respondent team tools can work with the HTRC API and that enhancements and extensions of that API are feasible.

4. **Query access to an index of item-level MODS metadata records describing public domain derived HT volumes.** Built over SQL, the Illinois MODS database schema was developed initially for the DLF Aquifer project and is currently in use as a critical component of the ongoing NEH-funded Emblematica Online – Open Emblem Portal project.[27] It supports complex queries over all elements and attributes of MODS metadata records, allowing maximum discovery, identification and selection of HT volumes based on MODS encoded bibliographic metadata.

5. **Support for initial testing of the prototypes against the HathiTrust corpus using the non-consumptive framework at Indiana.** The Illinois team will co-ordinate with the Indiana team to assist the respondents in preparing their code to run against the full HathiTrust corpus (both copyright-restricted and public domain) to garner preliminary results on prototype performance at scale.

6. **Read (and potentially write) access to the UIUC HTRC triple store.** UIUC will create and maintain an RDF-based triple store for use both by the HTRC team

[24] These files will be made available in bulk directly via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) or as tab-delimited files.

[25] See http://www.loc.gov/standards/mets.

[26] See http://www.openarchives.org/pmh.

[27] See http://emblematica.grainger.illinois.edu.

members involved in this project and by successful respondents. This triple store will be used to store item-level and collection-level metadata that have been transformed into RDF. For maximum flexibility it will actually be implemented by UIUC as a quad-store -- thereby allowing an identifier for each unique RDF record processed to be associated with each triple ingested from that record into the triple store, essentially providing provenance for each triple. In other words this allows each RDF record to be treated as its own named graph.[28] RDF triples are used to express simple descriptive assertions about a resource, collection, workset, etc. For example an RDF triple can be used to convey the title of a digitized book, the identity of its author, or its relationship to another digitized book. Since RDF triples can assert relationships, e.g., can assert that a book is described by a specific Library of Congress Subject Heading, it is anticipated that the UIUC triple store component will prove key to maximizing the use of Linked Open Data to enable and facilitate workset creation.

3.2 Modeling Collections and Worksets as Scholarly Resources

To answer Question #3, we will leverage the accomplishments and expertise of CIRSS in the area of scholarly collections research. CIRSS has developed an unmatched knowledge base in this domain over the past ten years through such ongoing grant initiatives as the IMLS-funded Digital Collections and Content[29] project and the Digital Public Library of America (DPLA) Beta Sprint.[30] CIRSS has also supported doctoral research for two cutting-edge dissertations on collection-level/item-level metadata relationships and collection-level subject access for digital collections (Wickett, 2012; Zavalina, 2010). Research at CIRSS has uncovered how libraries can capitalize on the value of collections for humanities scholars in an evolving information environment (Brockman, Neumann, Palmer, & Tidline, 2001), how scholars use thematic collections in large-scale digital repositories (Palmer, Zavalina, & Fenlon, 2010), and how information professionals can extend existing data models for large-scale cultural heritage aggregations to include collections (Wickett et al., forthcoming). Working together, HTRC and CIRSS are well positioned to define and describe the notion of collections and worksets in the context of scholarly computational research against a heterogeneous, large-scale digital corpus by developing a formal HTRC Model for Collections and Worksets.

Modeling collections – with an emphasis on worksets as a type of user-created collection – within the context of HTRC will ease the transition from specific project-based implementations for improving item-level metadata to long-term solutions that ensure worksets function as scholarly resources that users can return to over time and incorporate into their research processes and workflows. An HTRC data model for workset creation needs to allow scholars to gather and describe collections of resources from HathiTrust and to integrate outside resources (e.g., a file containing author gender information, other kinds of authoritative files, secondary literature, media, references) to serve as the input to computational analytical processes (see

[28] See http://en.wikipedia.org/wiki/Named_graph.

[29] See http://imlsdcc.grainger.uiuc.edu.

[30] See http://dpla.grainger.illinois.edu/Default.aspx.

Figure 1 in Section 2.2). A successful data model will be one that is extensible and interoperable rather than tightly bound to the HTRC. This data model will facilitate the sharing of collection and workset information with tools and corpora ***both within and beyond*** HTRC.

The HTRC Model for Collections and Worksets will include two primary elements: a set of core classes for representing collections, members, and their relationships; and a set of properties for describing collections and worksets. The core classes will be designed to support workset creation over time, by modeling worksets as entities that may change over time. This will distinguish worksets in HTRC from earlier approaches to collections in digital libraries, which as seen in Gonçalves, Fox, Watson, and Kipp (2004), treat collections as (mathematical) sets of digital objects. While mathematical sets are identified strictly by their membership, scholars using collections and creating worksets for research frequently need to adjust the worksets' membership. This means that a model that treats a collection or workset simply as a set of objects with no further identifying characteristics will not fully support the development of a collection or workset as a lasting scholarly resource. A more refined underlying approach to modeling collections in digital libraries that has been proposed by Meghini and Sypratos (2010) treats a collection as having both an *extension* (the set of resources gathered into the collection) and an *intension* (a set of criteria that determine whether an individual resource should be included in a collection). Although implementing computational methods for selecting or assessing collection members is beyond the scope of the modeling efforts proposed for CIRSS, the underlying concept of a division between the set of members of a collection and the over-arching policies and criteria that reflect the scholarly, informational or aesthetic purpose of a collection will be a guiding principle for the HTRC Model for Collections and Worksets.

The core classes and relationships for the HTRC Model for Collections and Worksets will be developed around the following concepts:

> **Source corpus:** a group of resources from which resources are retrieved and evaluated in a workset creation process. In the HTRC context, the source corpus is the portion of the HathiTrust that is available to a user for workset creation. The source corpus provides metadata to support assessment of the fit of any resource to the purpose of the collection.
>
> **Collection:** a collection is a group of resources gathered together for some informational, scholarly, or aesthetic purpose. The set of members that compose a collection may change over time to fit the purpose of the collection. Collections in the HTRC context may be composed of any resource that is identifiable in HTRC.
>
> **Workset:** a set of resources that is the input to a computational process (or a series of computational processes). Membership is essential to the identity of workset. It is defined at a particular time (e.g. by specifying a list of identifiers) and its members can not change over time. Worksets may be composed of HTRC collections (or identified subsets of collections) and may include external resources. A workset may be thought of as a derivative product from a collection. Information about worksets (e.g. their

composition, the algorithms and other details of their participation in computational processes) will be attached to the collection from which the workset was derived.

The properties for collection-level description that will be specified for the HTRC Model for Collections and Worksets will be selected from established vocabularies for collection description and supplemented with additional terms to support the particular role of collections and worksets in the HTRC context. The Dublin Core Collections Application Profile[31] is based on Heaney's "Analytical Model of Collections and Their Catalogs"(2000) and provides a set of collection-level properties designed to accommodate the description of collections in a number of environments. The members of the IMLS Digital Collections and Content (DCC) project, hosted at CIRSS, have developed a schema[32] for the description and representation of collections in a large-scale aggregation that is based on the Dublin Core Collections Application Profile. Recently, CIRSS researchers have collaborated with partners from the Europeana project to develop user requirements and recommendations for the modeling of collections in digital library aggregation and exchange environments.

The schemas mentioned above and the user requirements formulated for collections in digital library aggregation environments will form the initial basis of the property set for collection description in the HTRC Model for Collections and Worksets. The CIRSS team will evaluate the extent to which the available properties support the creation and scholarly use of collections in HTRC. Since these collection-level schemas were developed with relatively stable institutionally curated collections as the primary application area, we foresee that it will be necessary to develop additional properties to support the particular needs of researchers working with HathiTrust resources to create worksets and conduct computational analysis.

In order to facilitate instantiation of the HTRC Model for Collections and Worksets for the HathiTrust, documentation of the property set for collection-level description and the core classes will be made available to prototype projects and partners. The expectation is for the model to be primarily instantiated with XML records that will be integrated into the established HathiTrust infrastructure. Therefore the CIRSS team will produce an XML Schema for the HTRC Model for Collections and Worksets that specifies required and recommended properties for the description of collections. RDFS expressions of the properties and the core classes will also be developed to allow the publication of collection-level descriptions from the HTRC as RDF and to support the integration of HTRC collection information as Linked Open Data.

4. Project Structures, Roles, and Plans

The administrative structure of the WCSA is centered in the Graduate School of Library and Information Science (GSLIS) at the University of Illinois, and the University of Illinois is the lead institution requesting funding. J. Stephen Downie is the Principal Investigator (PI), and Timothy W. Cole is the Co-Principal Investigator (Co-PI) at Illinois and Beth Plale is the Co-PI at Indiana. Responsibility for the proper administration of the grant will be assumed by GSLIS. Thus,

[31] See http://dublincore.org/groups/collections/collection-application-profile.
[32] See http://imlsdcc.grainger.illinois.edu/CDschema_elements.

Downie will be the project’s intellectual, fiduciary, and administrative leader. The majority of the day-to-day work of the project will be located at CIRSS with coordinated activities at the University Library. Downie will take the lead on the technological aspects of the project while Cole will lead on the metadata aspects. Downie and Cole will co-direct the work on formal collection modeling, with input from Plale. Plale will act as liaison between the Illinois and Indiana technical teams of the HTRC. Downie, Cole and Plale will lead on the shaping of the RFP and the selection of final prototyping projects. Plale will lead in arranging the preliminary non-consumptive test runs of the prototypes against the full HathiTrust corpus at Indiana. Support for Indiana’s non-consumptive test runs will be funded via an institutional subaward from Illinois. At project’s end, CIRSS will produce a public report with input from Downie, Cole and Plale assessing the outcomes of WCSA and providing implementation recommendations for HTRC and the HathiTrust along with recommendations for future development.

4.1 Key Personnel

4.1.1 J. Stephen Downie, PhD

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 Music Information Retrieval Evaluation eXchange (MIREX). He was the Principal Investigator on the Networked Environment for Music Analysis (NEMA) project, funded by the Andrew W. Mellon Foundation. He is Co-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 now serves as ISMIR's President. 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.

4.1.2 Timothy W. Cole

Timothy W. Cole is Mathematics and Digital Content Access Librarian, Professor of Library and Information Science, and Professor, University Library, at the University of Illinois at Urbana-Champaign. A member of the faculty at Illinois since 1989, he has held prior administrative appointments as Head of Library Digital Services and Development, Systems Librarian for Digital Projects and Assistant Engineering Librarian for Information Services. He is a Principal Investigator (PI) for the Open Annotation Collaboration (Andrew W. Mellon Foundation), a co-PI for the Emblematica Online project (National Endowment for the Humanities & the Deutsche Forschungsgemeinschaft) and a co-PI / past PI for the IMLS Digital Collections and Content project (Institute of Museum and Library Sciences). He is a member of the International Mathematical Union Committee on Electronic Information and Communication, a member of the

National Academies Committee for Planning a Global Library of the Mathematical Sciences, a member of *Library Hi Tech* Editorial Board, past chair of the National Science Digital Library Technology Standing Committee and a former member of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) Technical Committee. He has published and presented widely on metadata best practices, OAI-PMH, digital library interoperability, Open Annotation, and the use of XML and SGML for encoding metadata and digitized scholarly resources in science, mathematics and literature. For further information see: http://www.library.illinois.edu/people/bios/t-cole3/.

4.1.3 Prof. Beth Plale, PhD

Beth Plale is a Professor of Computer Science in the Indiana University School of Informatics and Computing - Bloomington (SoIC). She is Managing Director of the Indiana University Pervasive Technology Institute (PTI), Director (and founder) of the Data to Insight Center (D2I), Director (and co-founder) of the Center for Data and Search Informatics, Co-Director (and co-founder) of the HathiTrust Research Center (HTRC), and founder and steering committee member of the Research Data Alliance (RDA). She is Principal Investigator on the Data Capsule for Non-Consumptive Research project funded by the Alfred P. Sloan Foundation, Data Science Consortium - Coming Together Around Data project funded by the National Science Foundation (NSF), Collaborative Research SI2 SSE: Pipeline Framework for Ensemble Runs on Clouds funded by NSF, Proposal to Build Trident Community funded by Microsoft Research and Microsoft Exploratory Research in Workflow and Related Areas also funded by Microsoft Research. She holds a B.S. in Computer Science from University of Southern Mississippi, a MBA in Business Administration from University of La Verne, a M.S. in Computer and Information Science from Temple University, and a Ph.D. in Computer Science from State University of New York at Binghamton. Plale is well-known on the national and international scene for her research in data preservation, provenance and metadata, workflows, socio-ecological informatics, and data search and retrieval as is confirmed by her long list of publications/products and invitations to present research around the globe.

4.2 Other Roles

4.2.1 Research Programmer

The research programmer will be staffing the WCSA project on a 37.5% FTE basis. The research programmer will be responsible for assisting the respondents in the development, testing and deployment of their prototype code on the project’s HTRC testing infrastructure. The research programmer will also work with the HTRC tech and CIRSS teams to implement the appropriate bibliographic, metadata and Linked Open Data structures and technologies to prototype distribution of the enhanced metadata information. The senior programmer will also provide input on the technical aspects of the final RFP call and the selection of the four successful prototyping projects.

4.2.2 Project Coordinator

The project coordinator will be assigned to the project on a 25% FTE basis. The coordinator will be responsible for keeping the project running smoothly by facilitating the basic administration aspects of the project including research meeting planning, communications, time and effort reporting, budget monitoring, etc. The coordinator will play the lead role in planning special fact-finding meetings with the community, the RFP meeting, and the final prototype demonstration meeting. The coordinator will manage the dissemination of the RFP and then manage the submissions of the candidate respondent prototyping projects. The project coordinator will maintain ongoing communications with the respondent projects to ensure their successful and timely completions.

4.2.3 Research Assistant

The research assistant will work with the project on a 50% FTE basis. The PhD-level research assistant will be an integral intellectual contributor to the success of the project. The research assistant will be situated in CIRSS. The research assistant will be tasked primarily with the formal model work associated with the project’s Question #3. Because of his/her close ties to CIRSS, the research assistant will play a liaison role among researchers in CIRSS with an active interest in workset creation and WCSA project participants. The research assistant will guide and advise the research programmer on making the design decisions needed to instantiate the formal collection model in code.

4.2.4 Graduate Hourly

The graduate hourly workers will be providing additional co-ordination and administrative support for the WCSA project. We have budgeted for 200 hours of graduate hourly work per year at $17.50 per hour. The hourly workers will be brought on board as needed. Their duties will include helping create and maintain project websites, writing documentation materials, coordinating RFP submissions, arranging and assisting at project workshops and fact-finding meetings, dealing with reimbursements for meeting participants, and so on.

4.3 Project Partners

4.3.1 Formal Organization of HTRC

Founded in 2011, the HathiTrust Research Center is uniquely collaborative in that it is co-located within two distinct institutions: University of Illinois at Urbana-Champaign and Indiana University at Bloomington. The HTRC is constituted by three Memoranda of Understanding (MOUs) and its formation proposal. The first MOU is the agreement between Illinois and Indiana as co-equal partners in the HTRC. The second MOU establishes the official relationship between the HTRC collaboration created in the first MOU and the HathiTrust Executive Committee. The third MOU establishes terms of use with regard to the Google Books public domain data between the Illinois branch of the HTRC and Google; Indiana has an identical Google MOU in place. All three MOUs have been signed by the Illinois administration and their respective counterparts at the relevant institutions (see Appendix B). The formation proposal, jointly written by the Illinois and Indiana teams, was accepted by the HathiTrust Executive Committee in December 2010. It outlines basic framework of our HTRC collaboration with Indiana and explains how the HTRC will function as the research arm of the HathiTrust. The formation proposal also outlines the following research and development 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.

4.3.2 Data to Insight Center (Indiana University)

The Indiana University arm of the HTRC is located at the Data to Insight Center, which is a 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. Because of the Data to Insight Center's close working relationship with 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.

4.3.3 CIRSS (GSLIS, University of Illinois at Urbana-Champaign)

The University of Illinois arm of the HTRC maintains an ongoing operational relationship with the Center for Informatics Research in Science and Scholarship (CIRSS).[33] Though the HTRC is an administrative entity distinct from CIRSS, the two centers recognize and are committed to building upon synergies in three key intellectual areas: 1) digital humanities; 2) collections, curation, and metadata; and 3) socio-technical data analytics. CIRSS conducts research on information problems that impact scientific and scholarly inquiry, with a focus on the curation and integration of digital information within and across research communities. CIRSS faculty and researchers 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.

[33] See http://cirss.lis.uiuc.edu/

The center's staff includes project coordinators, data analysts, and research assistants 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 activities bridge research and education by informing the development of a new curriculum in data curation and scientific information and providing a base for student research experiences.

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

The University of Illinois arm of the HTRC includes several faculty members from the University Library who serve as key staff on the Executive Committee and Technical Team. The Library is central to the University’s mission of teaching, research, and public service. It serves the curricular and research needs of students and faculty, and is committed to maintaining the strongest collections and services possible. The Library was established in 1867 with only 644 books from $1,000 appropriated by the State of Illinois. Today it is among the preeminent research collections in the world. 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 52,000 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 is decentralized and divided into a system of departmental units located campus wide. It currently employs more than 110 academic staff and over 170 support staff, not including hourly employees and student assistants. All librarians are 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.

4.4 Advisory Board

The WCSA team has assembled an eight-member advisory board consisting of experts well versed in large-scale digital library initiatives, digital humanities, information retrieval, and discovery interfaces. We do not anticipate convening a face-to-face advisory board meeting but will convene as a whole at least once by phone or Skype and will work with Board members one-on-one and in small groups on an ongoing and ad hoc manner over the course of the project. The project team will seek consultation from Board members on development, revision and release of the RFP; evaluation of respondents’ proposals and selection of prototyping projects; the final project report; and dissemination of project outcomes. The Advisory Board with be comprised of the following members:

- **Neil Fraistat**, Director, Maryland Institute for Technology in the Humanities, University of Maryland
- **Matthew Jockers**, Assistant Professor, University of Nebraska-Lincoln

- **R. Manmatha**, Research Associate Professor, Center for Intelligent Information Retrieval, University of Massachusetts Amherst
- **Bethany Nowviskie**, Director of Digital Research & Scholarship, University of Virginia Library
- **Andreas Rauber**, Associate Professor, Department of Software Technology and Interactive Systems, Vienna University of Technology
- **Stéfan Sinclair**, Associate Professor of Digital Humanities, McGill University
- **John Unsworth**, Vice Provost for Library & Technology Services and Chief Information Officer, Brandeis University
- **John Wilkin**, Executive Director, HathiTrust

4.5 Synergistic Activities

*4.5.1 HathiTrust Research Center @ Illinois Initiative: Bridging Suppor*t

In 2012, the University of Illinois committed $606,848 in bridge funding to support the HTRC for three years while the center transitions from its initial start up period to its more established status as an initiative with its own sustainability structure and set of funded projects. These funds contribute to maintaining the core HTRC team at Illinois and building further infrastructure for the HTRC.

4.5.2 Secure Computational and Data Environments for Non-Consumptive Research (Indiana University)

Developing a secure computation and data environment for non-consumptive research for the HathiTrust Research Center is funded through a grant from the Alfred P. Sloan Foundation. Researchers at the University of Michigan and the Data to Insight Center are developing a "data capsule framework" that is founded on a principle of "trust but verify". That is, the informatics scholar is given freedom to experiment with new algorithms on a huge body of copyrighted or otherwise protected information, but technological mechanisms are in place to verify compliance with the policy of non-consumptive research. This research will develop a prototype system that can support:

1. Non-consumptive research – that is, provides safe handling of large volumes of data, and can ensure that the read restrictions of the definition hold;
2. Openness – users are not limited to using a known set of algorithms, and instead are expected to experiment with their own algorithms;
3. Efficiency – It will not be possible to analyze algorithms for conformance prior to execution;
4. Low cost and scale – Run at large-scale and low cost to users; and
5. Long term and broad value – framework will need to be designed for adoption for other purposes

4.5.3 IMLS Digital Collections and Content Project (University of Illinois)

Since 2002, the Institute for Museum and Library Services (IMLS) Digital Collections and Content (DCC) project has developed and maintained a nationally scoped aggregation that brings together cultural heritage collections and exhibits from libraries, museums, and archives from across the country. DCC provides both collection-level and item-level access to facilitate searching and browsing and to retain the institutional identities and collection contexts that are vital to how users explore and interact with cultural heritage materials. The DCC has investigated and implemented a systematic approach to developing useful, meaningful, and usable digital collections. The project team, which consisted of staff and faculty from CIRSS and the University Library at Illinois, explored how to use the relationships between collection-level and item-level metadata in federated digital repositories to preserve content and make the content more useful for scholars and the public.

Recently, the DCC project team has applied its research to other national and international aggregations. The DCC participated in the 2011 Digital Public Library of America Beta Sprint competition, which has resulted in ongoing development of the initial DPLA prototype by refining the prototype’s information retrieval algorithms and implementing additional layers of functionality that allow users to interact more directly and dynamically with the prototype’s data. The DCC team has also worked in close collaboration with researchers from Europeana to produce a white paper that provides recommendations for modeling collections in digital library aggregation and exchange environments like Europeana and The European Library.

4.5.4 Open Annotation Collaboration (University of Illinois)

Annotating is a method by which scholars across disciplines organize existing knowledge and facilitate the creation and sharing of new knowledge. It is used by individual scholars when reading as an aid to memory, to add commentary, and to classify. It can facilitate metadata enrichment, shared editing, scholarly collaboration, and pedagogy. In the context of the HTRC Workset Creation for Scholarly Analysis: Prototyping Project we anticipate that stand-off annotations will be used by tools and services to convey added metadata attributes and submit many other forms of metadata augmentations.

With the support of the Andrew W. Mellon Foundation, the Open Annotation Collaboration (2009-2013) effort has focused on annotation interoperability, the creation of a Web and resource-centric data models and ontology consistent with Linked Open Data best practices and the Semantic Web. Founded by the University of Illinois at Urbana-Champaign, JSTOR, Los Alamos National Laboratory, the University of Maryland, and the University of Queensland (Australia), the collaboration had grown by 2012 to include a total of 12 institutions worldwide. In 2012, together with the Annotation Ontology initiative (Harvard University, the University of Manchester (UK), et al.) the OAC founded the W3C Open Annotation Community Group. The OAC project is culminating in 2013 with the release of the W3C Open Annotation Community Group data model and ontology, the implementation of the Open Annotation service and tool registry, the release of a video annotation plugin for Drupal, and the release of an Open Annotation validation service and test annotation repository. The experience from OAC will inform and facilitate work with collaborating partners on this project, providing a standards-based foundation for annotation interoperability between the HTRC and tools and services developed by partners to provide metadata enrichment and augmentation.

5. Expected Outcomes and Benefits

There are six principal outcomes for the WCSA project. These outcomes will directly benefit the HathiTrust, the HTRC, and digital humanities scholarship. These are:

1. A set of prototype algorithms that could be used by the HathiTrust, HTRC and digital humanities scholars to define new collections for analysis.
2. A collection of new metadata outputs from the prototype algorithms that could be used to assist in improving access to the HathiTrust corpus and/or be used in novel analyses within or beyond the HTRC.
3. A suite of prototype Linked Open Data (or similar) resources that will expose the outputs of the prototype algorithms for use both within and beyond the HTRC context.
4. A formal model of collections and worksets that can be used to shape the development of new discovery, search and analytic resources both within and beyond the HTRC context.
5. The realization of the formal collection model in a form (or forms) that can be used to encode collections and worksets for use in actual HTRC analyses and for the subsequent publication and exchange of such collections among digital humanities scholars.
6. A set of recommendations based upon the experience of creating the first five outcomes listed above designed to guide both the HTRC and the digital scholarship community in formulating a high-impact, long-term research and development plan.

6. Intellectual Property Issues

WCSA will be subject to the Foundation’s intellectual property policy,[34] and each of the four successful respondent teams chosen from the RFP will be subject to the terms of the intellectual property agreement established between the University of Illinois and the Foundation. Respondents will be informed of the terms of this agreement as part of the RFP process and will be required to agree to its terms prior to the disbursement of awards. 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. Reports 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 license permitting non-commercial use and modification. We have modeled our RFP process (including the

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

informing respondents of the intellectual property policy and requiring a signed agreement) on the Open Annotation Collaboration.

7. Sustainability Strategies

The HTRC is a fledgling organization and is still in its growth phase. Since its inception, however, the HTRC executive board has been conscientiously putting sustainability at the top of its planning agenda. It has appointed Prof. John Unsworth, Vice-Provost at Brandeis University and HTRC Co-Founder, its Chief Sustainability Officer (CSO). Dr. Unsworth has a strong track record of finding long-term sustainability resources for the projects with which he has been involved. One avenue of long-term support that he is actively investigating is an arrangement with commercial scholarly content providers wherein HTRC would manage computational research access to their copyright-restricted materials using HTRC’s non-consumptive framework. The HTRC is also actively pursuing funding opportunities involving partners from the United States and Canada from such sources as the National Science Foundation (NSF), National Endowment for the Humanities (NEH) and the Social Science and Humanities Research Council (SSHRC). Furthermore, the HathiTrust itself has a broad base of active support from its sixty participating institutions. This broad base of support gives the HathiTrust a strong long-term sustainability foundation. Thus, should the HTRC cease to exist at some point in the future, the outputs of the WCSA will be turned over to the HathiTrust for long-term use and safekeeping.

Notwithstanding the sustainability issues pertaining to the HTRC, we believe that enabling, encouraging and supporting the continual use of a project’s outputs is the best sustainability strategy for ensuring the ongoing impact of those outputs. To this end, the open-source licensing of the WCSA’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 prototypes might be incorporated into the HathiTrust Digital Library maintained at the University of Michigan. In a similar way, HTRC will be working with the HathiTrust Board to explore how the Linked Open Data metadata resources might be integrated with the HathiTrust Digital Library. WCSA will also be developing a plan to encourage the use of its new formal collection model and tools so that digital humanities scholars might make use these tools to create and share their analytic collections as a set of new scholarly resources.

8. Reporting

Since the WCSA project will span 24 months, from July 1, 2013 to June 30, 2015, we propose the submission of two formal project reports (i.e., one interim report and one final report). The reports will include narrative commentary on the activities, successes and challenges of the project. The reports will also discuss grant expenditures in conjunction with the official budgetary accounting provided by the University of Illinois accounting office. Table 4 outlines our proposed reporting structure.

Report	Dates Covered	Due Date
Year I interim narrative and budget report	July 1, 2013 – June 30, 2014	September 30, 2014
Year II final narrative and budget report	July 1, 2014 – June 30, 2015	September, 30 2015

Table 4. Project Reporting Schedule

9. Timeline

Activity	Dates Covered	Personnel
Raise awareness of workset creation issues and gather additional user requirements from digital library and digital humanities communities at JCDL 2013 and DC 2013.	July-September 2013	PI, Co-PI (Illinois), Co-PI (Indiana), Project Coordinator
Experiment with manual and automated methods for evaluating the HathiTrust corpus.	July-November 2013	Research Assistant
Build subset of the larger HathiTrust to test the projects under development.	July-November 2013	Research Programmer (Illinois)
Host Workset Creation track at the second annual HTRC UnCamp.	September 2013	PI, Co-PI (Illinois), Project Coordinator
Report on preliminary analysis of HathiTrust corpus evaluation.	September 2013	Research Assistant
Activity (continued)	**Dates Covered**	**Personnel**
Revise RFP based on analysis of initial project and release.	November 2013	PI, Co-PI (Illinois), Co-PI (Indiana), Project Coordinator
RFP Responses Due	January 15, 2014	Prototyping Project Respondents
Review all submitted proposals	January 2014	PI, Co-PI (Illinois), Co-PI (Indiana), Project Coordinator, Research Assistant, Advisory Board
Produce technical report on corpus coverage 3of HathiTrust corpus and identify potential audiences for strategic community outreach	January 2014	Co-PI (Illinois) Research Assistant, Project Coordinator

Activity (continued)	Dates Covered	Personnel
Convene “RFP Shortlist Meeting” for proposal presentations	February 20, 2014	PI, Co-PI (Illinois), Research Programmer (Illinois), Project Coordinator, Prototyping Project Respondents
Award funding for four prototyping projects	March 2014	PI, Co-PI (Illinois), Project Coordinator
Provide data access to prototyping projects	March 2014	Research Programmer (Illinois)
Prototyping Project Period	April 1014-January 2015	Prototyping Project Respondents
Hold conference calls with teams from prototyping projects every two months to monitor progress.	April, June, August, October, December 2014	PI, Co-PI (Illinois), Research Programmer (Illinois), Project Coordinator
Consult with Indiana University regarding prototyping projects	April 2014-January 2015 as needed	PI, Co-PI (Illinois), Research Programmer (Illinois), Project Coordinator
Provide assistance running newly developed tools against HTRC infrastructure	April 2014-January 2015 as needed	Research Programmer (Illinois); Programmer (Indiana)
Gather and review related data models and user requirements for collections modeling at CIRSS	April 2015	Research Assistant
Develop and evaluate instantiated data model to support collections as scholarly resources	May 2014-November 2015	Research Assistant, Project Coordinator, Hourly Support
Continue strategic community building activities and report progress toward project completion and lessons learned at DH 2014	July 2014	PI
Produce technical report presenting data model	December 2014-January 2015	Research Assistant, Project Coordinator, Hourly Support
Conclude demonstration projects with a Prototype Demonstration Meeting	January 2015	PI, Co-PI (Illinois), Research Programmer, Project Coordinator, Research Assistant, Prototyping Project Respondents

Activity (continued)	Dates Covered	Personnel
Run tools developed through prototyping project against full corpus at Indiana University	February-March 2015	Programmer (Indiana)
Assess outcomes of demonstration projects, including feasibility of implementation at scale	March-April 2015	PI, Co-PI (Illinois), Co-PI (Indiana), Research Programmer (Illinois), Programmer (Indiana)
Identify opportunities for future development	March-April 2015	PI, Co-PI (Illinois), Co-PI (Indiana)
Produce a public report based on outcomes, assessments, and next steps	May-June 2015	PI, Co-PI (Illinois), Project Coordinator

Table 6: Timeline for July 2013-June 2015

10. Budget Commentary

No institutional overhead costs are included. No tuition remission costs are included. Inflation calculated at 3%.

Item #1. Represents one month summer salary for Downie as PI.

Item #2. Represents 5% FTE effort for Cole as Co-PI.

Item #3. Represents 37.5% FTE effort of a research programmer to support HTRC development efforts and to assist prototyping projects in working with HTRC materials and systems.

Item #4. Represents 25% FTE effort of a project coordinator to assist in the various administrative tasks of the project including report monitoring, time management, meeting and travel arrangements, etc.

Item #5. Represents 200 hours of graduate hourly support at $17.50 per hour to assist as needed with various project tasks as they arise.

Item #6. Represents the full annual stipend of a PhD student research assistant (RA) at 50% FTE effort. We intend this RA to play a leadership role in the collection modeling work at CIRSS. Does not include tuition or fee costs.

Item #7. Represents the mandatory benefits for the personnel in Items #1-4, calculated at 44.67%

Item #8. Represents the mandatory benefits for the graduate hourly support personnel calculated at 0.14%.

Item #9. Represents the mandatory benefits for PhD research assistant, calculated at 5.99%.

Item #10. Represent the costs associated with project supplies and services including, paper, cables, printing, etc. and is based upon prior project expenditures.

Item #11. Represents computer and related hardware costs. System includes a lighter-weight head node for administration, two substantial computation nodes, and substantial disk space. Set up is designed to cover the non-trivial computation and storage resources needed to support the development, testing and evaluation of the prototyping projects.

Item #11.1. HP ProLiant DL320 G6 server with 12 GB memory to act as head node. Quote obtained from the HP Public Sector Representative for the University of Illinois.

Item #11.2. Two HP ProLiant DL160 Gen8 Servers with 96 GB memory. Quote obtained from the HP Public Sector Representative for the University of Illinois.

Item #11.3. Eighteen 3TB SAS, 7200RPM, hard drives. Quote obtained from CDW-G.

Item #12. Represents the costs associated with hosting the RFP Shortlist Meeting. Intended to include travel, accommodation and meal support for participants and organizers. Chicago or similar destination likely venue. Costs calculated include reimbursement for 5 domestic participants (2 nights' lodging at $190 per night, 3 days' per diem at $28 per day, ~$500 airfare) and for 3 international participants (2 nights' lodging at $190 per night, 3 days' per diem at $28 per day ~$1,500 airfare). Cost estimates also include lodging and per diem for 5 people from the University of Illinois (2 nights' lodging at $190 per night, 3 days' per diem at $28 per day, no airfare included). The remainder of $1,968 is budgeted for meeting space for 2 days, A/V, and any food service that we provide. Domestic airfare estimates were obtained by assuming travel from Washington, DC to Chicago, IL, and international airfare estimates were obtained by assuming travel from London, UK to Chicago, IL. Airfare approximations are based on economy flight information on the American Airlines website. Reimbursement rates are the State of Illinois per diem for Chicago, IL.

Description	**People**	**Airfare**	**Nights**	**Lodging**	**Days**	**Per Diem**	**Total**
Domestic travel	5	$500	2	$190	3	$28	$4,820
International travel	3	$1,500	2	$190	3	$28	$5,892
Personnel	5	n/a	2	$190	3	$28	$2,320

Table 7: Budget Breakdown for Item #12, RFP Shortlist Meeting

Item #13. Represents the costs associated with hosting the Prototype Demonstration Meeting. Intended to include travel, accommodation and meal support for participants and organizers. Chicago or similar destination likely venue. The four prototyping awards will include travel funding for one project participant as part of the set $40,000 (See Item #17). We expect to invite more members of the advisory board for this meeting, however, which will result in directly funding the same total number of people as the RFP Shortlist meeting (13) with a total of 17 participants for the Prototype Demonstration Meeting. Costs calculated include reimbursement for 5 domestic participants (2 nights' lodging at $190 per night, 3 days' per diem at $28 per day,

~$500 airfare) and for 3 international participants (2 nights' lodging at $190 per night, 3 days' per diem at $28 per day ~$1,500 airfare). Cost estimates also include lodging and per diem for 5 people from the University of Illinois (2 nights' lodging at $190 per night, 3 days' per diem at $28 per day, no airfare included). The remainder of $1,968 is budgeted for meeting space for 2 days, A/V, and any food service that we provide. Domestic airfare estimates were obtained by assuming travel from Washington, DC to Chicago, IL, and international airfare estimates were obtained by assuming travel from London, UK to Chicago, IL. Airfare approximations are based on economy flight information on the American Airlines website. Reimbursement rates are the State of Illinois in-state per diem for Chicago, IL.

Description	**People**	**Airfare**	**Nights**	**Lodging**	**Days**	**Per Diem**	**Total**
Domestic travel	5	$500	2	$190	3	$28	$4,820
International travel	3	$1,500	2	$190	3	$28	$5,892
Personnel	5	n/a	2	$190	3	$28	$2,320

Table 8: Budget Breakdown for Item #13, Prototype Demonstration Meeting

Item #14. Represents travel costs associated with project members for meetings and conference presentations, etc. Budget provides travel to 2 domestic events for 3 people per event (4 nights' lodging at $110 per night, 5 days' per diem at $32 per day, ~$173 for car rental and mileage) and 1 international event for 1 person (5 nights' lodging at $319 per night, 6 days' per diem at $180 per day, ~$1,500). As stated in the narrative and timeline, project team members plan to attend DH 2013 (Lincoln, Nebraska, July 16-19) and JCDL 2013 (Indianapolis, Indiana, July 22-26) to raise awareness of workset creation issues and gather additional user requirements from digital library and digital humanities communities. At the project's midpoint, one project member will attend DH 2014 (Lausanne, Switzerland, dates to be determined) to continue advancing the digital humanities community building agenda, report progress toward project completion and lessons learned, and seek interim feedback from the core scholarly user group for HTRC. Cost estimated using rates for domestic out-of-state travel and Lausanne, Switzerland for international travel. Reimbursement rates are the State of Illinois per diem for out-of-state travel and the State Department Foreign Per Diem Rates for Lausanne, Switzerland. Airfare approximations are based on economy flight information on the American Airlines website. Car rental approximations are based on economy class rentals on the Enterprise website.

Description	**People**	**Travel**	**Nights**	**Lodging**	**Days**	**Per Diem**	**Total**
Domestic travel 1	3	$174	4	$110	5	$32	$2,321
Domestic travel 2	3	$174	4	$110	5	$32	$2,321
International travel	1	$1500	5	$242	6	$162	$3,358

Table 9: Budget Breakdown for Item #14, Travel Costs

Item #15. Represents communication costs associated with the project based upon prior project expenditures and includes telephone, faxing, courier and cell phone costs, etc.

Item #16. Represents costs for the Indiana HTRC group to support installing and running prototype code on the Indiana infrastructure. Year I is budgeted at 5% FTE effort of a professional programmer. Effort will increase in Year II to 10% as this will be the time when the prototyping project code will be tested against the main HTRC services at Indiana.

Item #17. Represents the cost of the four prototype projects at $40,000 each.

11. Bibliography

Beall, J. (2008). The weaknesses of full-text searching. *The Journal of Academic Librarianship 34*(5), 438–444. Retrieved February 26 from http://dx.doi.org/10.1016/j.acalib.2008.06.007

Brockman, W. S., Neumann, L., Palmer, C. L., & Tidline, T. J. (2001). Scholarly work in the humanities and the evolving information environment. Digital Library Federation/Council on Library and Information Resources. Retrieved February 13, 2013 from www.clir.org/pubs/reports/pub104/pub104.pdf

Brogan, M. (2006). Contexts and Contributions: Building the Distributed Library. Digital Library Federation/Council on Library and Information Resources. Retrieved August 2, 2010 from http://www.diglib.org/pubs/dlf106

Buckland, M. K. (1999). *Library Services in Theory and Context* (2nd ed.). New York: Pergamon Press. Retrieved August 2, 2010 from http://sunsite.berkeley.edu/Literature/Library/Services/index.html

CLIR (Council on Library and Information Resources). (2010). *The idea of order: transforming research collections for 21st century scholarship*. Washington, D.C.: CLIR.

Currall, J., Moss, M., & Stuart, S. (2004). What is a collection? *Archivaria, 58*, 131-146.

Dempsey, L. (2006, February). The (digital) library environment: Ten years after. *Ariadne, 46*. Retrieved February 13, 2013 from http://www.ariadne.ac.uk/issue46/dempsey/

Denton, W., & Coysh, S. J. (2011).Usability testing of VuFind at an academic library.*Library Hi Tech, 29*(2), 301-319.doi:10.1108/07378831111138189

De Rosa, C. (2005). *Perceptions of libraries and information resources*. A report to the OCLC membership. Dublin, Ohio: OCLC Online Computer Library Center.

Green, H., Saylor, N.,& Courtney, A. (2013). Beyond the scanned image: A needs assessment of faculty users of digital collections. Forthcoming paper to be presented at Digital Humanities 2013, Lincoln, Nebraska.

Gonçalves, M., Fox, E., Watson, L., and Kipp, N. (2004). Streams, structures, spaces, scenarios, societies (5S): A formal model for digital libraries. *ACM Transactions on Information Systems, 22*, 270-312.

Gorman, M. (1990). Descriptive cataloguing: Its past, present, and future. In M. Gorman (Ed.)

Technical Services Today and Tomorrow (63-73). Englewood, CO.: Libraries Unlimited.

Heaney, M. (2000). An analytical model of collections and their catalogs. UK Office for Library and Information Science. Retrieved on February 13, 2013 from http://www.ukoln.ac.uk/metadata/rslp/model/

Hensen, S.L. (1989). *Archives, Personal Papers, and Manuscripts: A Cataloging Manual for Archival Repositories, Historical Societies, and Manuscript Libraries* (2nd ed.). Chicago: Society of American Archivists.

Lagoze, C. & Fielding, D. (1998). Defining collections in distributed digital libraries. *D-Lib Magazine 4*(11).

Library of Congress, Network Development and MARC Standards Office. (2003). Appendix C:Minimal level record examples. MARC 21 format for bibliographic data. Retrieved on November 28, 2011 from http://www.loc.gov/marc/bibliographic/bdapndxc.html

Meghini, C. & Spyratos, N. (2010).Unifying the concept of collection in digital libraries. *Advances in Intelligent Information Systems*, 197-224.

Moen, William. (2009, October 14). Bibliographic control alphabet soup: AACR to RDA and evolution of MARC [Webinar]. Retrieved February 7, 2013 fromhttp://www.niso.org/news/events/2009/bibcontrol09/

Moen, W.E. & Benardino, P. (2003). Assessing Metadata Utilization: An Analysis of MARC Content Designation Use. In 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice — Metadata Research and Application, Seattle, Washington. Retrieved on November 28, 2011 from http://www.unt.edu/wmoen/publications/MARCPaper_Final2003.pdf

Mueller, M. (2010). Towards a digital carrel: A report about corpus query tools. Report on Mellon-funded Workshop at Northwestern University. Retrieved on February 13, 2013 from http://panini.northwestern.edu/mmueller/corpusquerytools.pdf

Palmer, C. (2004). Thematic research collections. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.) *A Companion to Digital Humanities*. Oxford: Blackwell. Retrieved on February 13, 2013 from http://www.digitalhumanities.org/companion/

Palmer, C. L. (2005). Scholarly work and the shaping of digital access. *Journal of the American Society for Information Science and Technology, 56*(11), 1140-1153.

Palmer, C. L., Zavalina, O., & Fenlon, K. (2010). Beyond size and search: Building contextual mass in digital aggregations for scholarly use. In A. Grove (Ed.) *Proceedings of the ASIS&T Annual Meeting.*

Smith-Yoshimura, K., Argus, C., Dickey, T.J., Naun, C.C., Rowlison de Ortiz, L., & Taylor, H. (2010). *Implications of MARC Tag Usage on Library Metadata Practices*. Report produced by OCLC Research in support of the RLG Partnership. Dublin, OH: OCLC Research. Retrieved on February 13, 2013 from www.oclc.org/research/publications/library/2010/2010-06.pdf

Spiro, L.& Segal, J. (2007). The impact of digital resources on humanities research.

Retrieved on February 13, 2013 from http://library.rice.edu/services/dmc/about/projects/the-impact-of-digital-resources-on-humanities-research

Sukovic, S. (2008). Convergent flows: Humanities scholars and their interactions with electronic texts. *Library Quarterly 78*(3), 263-284.

Sukovic, S.(2011). E-Texts in Research Projects in the Humanities. In A. Woodsworth & W. D. Penniman (Eds.) *Advances in Librarianship* (131-202). Bingley, UK: Emerald Group Publishing.

Varvel, V. E. Jr., & Thomer, A. (2011). *Google Digital Humanities Awards recipient interviews report.* CIRSS Report No. HTRC1101. Champaign, IL: Center for Information Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

Warwick, C., Terras, M., Huntington, P., & Pappa, N. (2008). If you build it will they come? The Lairah Study: Quantifying the use of online resources in the arts and humanities through statistical analysis of user log data. *Literary and Linguistic Computing, 23*(1), 85-102.

Wickett, K. M. (2012). Collection/item metadata relationships. Ph.D. Dissertation, University of Illinois at Urbana-Champaign.

Wickett, K. M., Isaac, A., Meghini, C., Doerr, M., Fenlon, K., Jett, J., & Palmer, C.L. (forthcoming). Modeling collections for aggregation and exchange environments. CIRSS Report. Champaign, IL: Center for Information Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

Yeo, G. (2012). The conceptual fonds and the physical collection, *Archivaria, 73*, 43-80.

Zavalina, O. (2010). Collection-level subject access in aggregations of digital collections: Metadata application and use. Ph.D. Dissertation, University of Illinois at Urbana-Champaign.