MODELING WORKSETS IN THE HATHITRUST RESEARCH CENTER

FOR THE
WORKSET CREATION FOR SCHOLARLY ANALYSIS PROJECT

PREPARED BY:

JACOB JETT

CENTER FOR INFORMATICS RESEARCH IN SCIENCE AND SCHOLARSHIP
GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CIRSS TECHNICAL REPORT, WCSA0715

REVISED SEPTEMBER 2015

AKNOWLEDGEMENTS

The author of this report gratefully acknowledges the contributions of fellow WCSA and HTRC project members Christopher Maden, Colleen Fallaw, and Timothy Cole without whose efforts and insights, the prototyping infrastructure necessary for the emergence of the following report would not have been possible. Feedback from the GSLIS-based Conceptual Foundations Group, especially with regards to David Dubin and Allen Renear was very helpful in shaping the distinction between formalizing what Worksets are and developing a model through which they can be realized through infrastructure. Likewise, feedback from Europeana experts Valentine Charles and Antoine Isaac led directly to the employment of the edm:isGatheredInto predicate in the model. Convergence among collection modeling efforts is a good thing. Finally, this report was made possible through the generous funding of The Andrew W. Mellon Foundation, Grant Ref # 21300666.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The HathiTrust Digital Library (HTDL) represents the combined output of several large scale digitization efforts and its corpus contains over 14.2 million volumes, made up of billions of pages of digitized content. The HathiTrust Research Center (HTRC), is a collaborative research effort jointly based at Indiana University and the University of Illinois at Urbana-Champaign, that is engaged in the development of tools and services that afford digital humanities scholars new opportunities for engaging with the HTDL corpus. Since almost two-thirds of the HTDL corpus remains within the purview of copyright, the HTRC has initiated a series of research projects assigned to explore how scholars can use works that remain protected by copyright as objects of research.

Each of these research projects is fully aligned with notions of *non-consumptive use* that emerged with the 2009 proposed settlement for the Authors Guild v. Google case[1] (and its 2011 amended version). The primary tenets of *non-consumptive use* or *non-consumptive research* are:[2]

1. The mandate that the research methodologies to be used conform to those used in typical computation analysis.
2. Human researchers cannot interact with large portions of data that comprise works within the purview of copyright protections.
3. The data products cannot contain data that could later be reassembled in such a way as to reproduce the works from which it was derived.

The Workset Creation for Scholarly Analysis (WCSA) project, one of the HTRC's research initiatives, represents one of the approaches taken to develop a web service that would enable researchers to examine the HTDL's corpus as research objects while adhering to the notion of *non-consumptive research*.[3] A cornerstone the HTRC's research paradigm is the notion of a *workset*, a collection entity analogous to a scholar's research collection. The need to conform to the *non-consumptive research* notion's expectations places additional constraints on the *workset* notion, resulting in an entity that is specialized for automated computational analytics environments. The workset then is an entity that contains all of a scholar's gathered research materials and is consumed by the HTRC's automated analytical workflows which in turn produce a set of data results for the scholar to remark upon and report.

The goal of this report is to formalize the notion of collections and worksets in the context of the HTRC and the WCSA project. A conceptual analysis is carried out using formal methods to tease apart the relationship between collections and worksets as kinds. It arrives at the following axiom to formally define worksets as entities.

[1] Eventually dismissed in 2013 (http://arstechnica.com/tech-policy/2013/11/google-books-ruled-legal-in-massive-win-for-fair-use/)

[2] Cf https://lib.stanford.edu/files/GBS-StanfordAmicus-9-8-09-2.pdf for additional details on the definition of *non-consumptive research*

[3] The Sloan funded Data Capsule project is another example of an HTRC initiative developing tools that conform to emerging *non-consumptive research* standard.

$$\text{Aw: } \forall y \Big(\Big(\Big(\big(\exists x \exists c \exists w \, (gathered(x,y) \wedge meetsCriteria(x,c) \wedge definedBy(c,w)) \big) \wedge \exists z \, hasResearchMotivation(y,z) \Big) \wedge intendedForUse(y,a) \Big) \leftrightarrow Workset(y) \Big)$$, where a is the expectation that the contents of y will be consumed by an automated process for in accordance with the non-consumptive research paradigm.

Using this definition, which describes worksets as collections whose items are gathered together according to a set of criteria and which has a both a research motivation and the particular expectation that it will be exploited by automated analytics workflows, in conjunction with the following use cases (see Table 1), a series of technical requirements are developed (see Table 2). A basic conceptual model, fitting worksets into the non-consumptive paradigm's workflow, is developed. A general data model (Figure 1) for both Worksets and Bibliographic Resources are developed with an RDF-based linked data infrastructure approach in mind.

A preliminary set of extensions to the data models is described so that any infrastructure resulting from their implementation can be positioned so as to easily begin to evolve with regards to its scope and basic functionality. These evolutions include: developing accommodations for exploiting page-level entities as bibliographic resources, leveraging existing data outputs (such as extracted page features) as metadata, developing formal vocabularies to describe and exploit bibliographic resources at both finer and more arbitrary granularities (e.g., paragraphs, sentences, poems, chapters, etc.), among others.

Table 1: Collected Use Cases

Code	Use Case	Source
UC1	Contents must be restricted to exploitation by automated analytics agents.	HTRC
UC2	Must be a citable research product.	Scholars
UC3	Must afford flexibility to assemble desired units of analysis.	Scholars
UC4	Must conform to linked data standards.	Technical

Table 2: Collected Technical Requirements

Code	Requirement	Supports
TR1	Worksets are containers.	UC1, UC2, UC3, UC4
TR2	Worksets are unique, globally persistent entities.	UC1, UC2, UC4
TR3	Worksets are immutable.	UC1, UC2, UC4
TR4	Worksets are agnostic with respect to item granularity and source.	UC3
TR5	Workset properties are informed by item properties.	UC1
TR6	Bibliographic resources are unique, globally persistent entities.	UC1, UC2, UC4, TR2, TR3
TR7	Bibliographic resources are immutable.	UC1, UC2, UC4, TR2, TR3
TR8	Bibliographic resource properties must be enumerated.	UC2, UC3, UC4, TR3, TR4, TR5

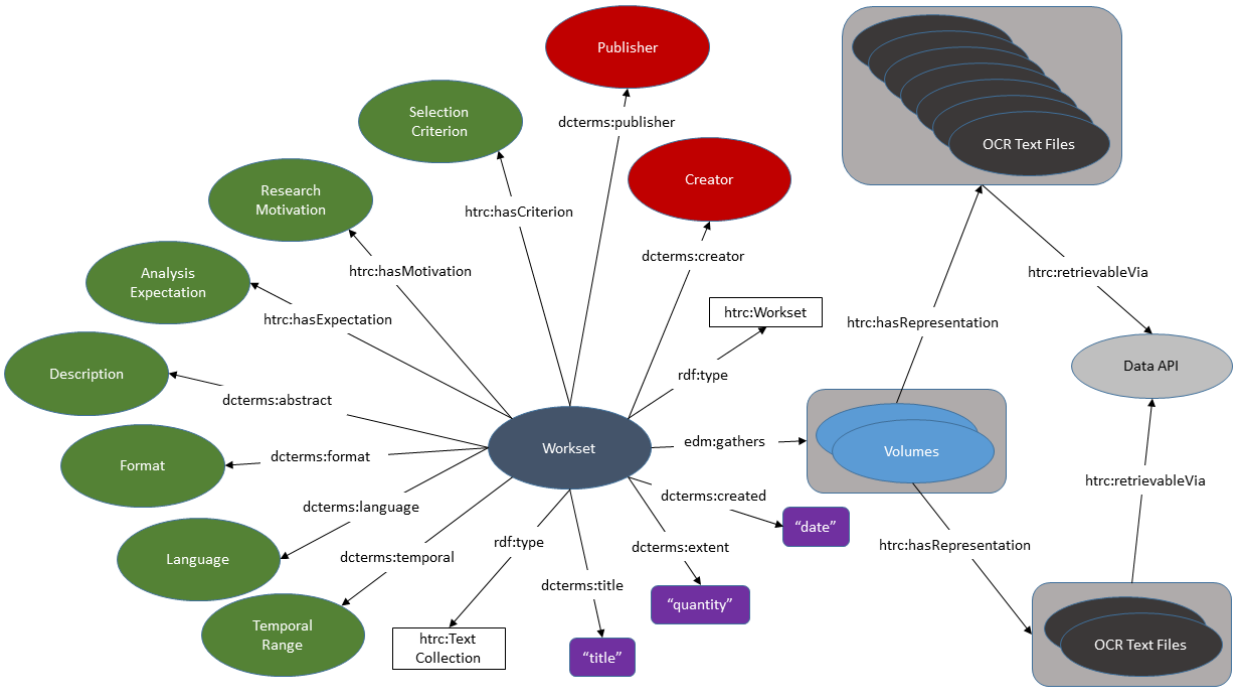

Figure 1: Combined Workset & Bibliographic Resource Data Models

The report closes with the recommendation that a new HTRC Workset Builder be implemented applying the data models illustrated in Figure 1. A series of additional recommendations to expand the models and the functionalities of the infrastructure are also detailed. In conjunction with the illustrated models, this report recommends the development of an accompanying data model for analytics modules. This accompanying data model will greatly facilitate the ability for developers to craft analytics tools for non-consumptive paradigms as the expectations for inputs and outputs can be clearly defined and coherently managed on a well-defined basis in relation to the HTRC's overall technical infrastructure.

1. INTRODUCTION & BACKGROUND

The HathiTrust Research Center (HTRC) is a research initiative of the HathiTrust Digital Library (HTDL). The library contains over 14.2 million volumes, comprising several billion pages of digitized text. It is the HTRC's goal to expand upon the suite of tools and services being offered to support scholarly access to the HathiTrust corpus. Enabled by the HTRC's tools and services, scholars can select subsets of the HathiTrust corpus for computational analysis in accordance to their particular research objectives. The HTRC refers to these researcher curated subsets, and any external data sources that scholars associate with them, as "worksets".

The Workset Creation for Scholarly Analysis (WCSA) project is a research initiative directly affiliated with the HTRC. This two-year project has been funded by the Andrew W. Mellon Foundation with three goals in mind:

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 HTRC.

This white paper is specifically focused upon addressing the third WCSA goal, formalizing the notion of collections and worksets in the context of the HTRC. To that end, its narrative is laid out into three narrowly scoped but overlapping sections. This first section describes the primary institutional use case that drives its development and a series of scholarly use cases derived from past studies of digital humanities scholars. The second section considers the precise nature of worksets both as a kind of entity and in relation to similar entities and proposes a formal definition for worksets in first order logic. The third and final section puts forth an initial specification designed to propel the Workset concept beyond its nascent existence within existing HTRC infrastructure and lays out a step-by-step plan by which it can be evolved to better meet the scholarly needs described in the use cases.

1.1 INSTITUTIONAL USE CASE (UC1)

As about two thirds of the HTDL's corpus remains under copyright, HTRC web services are being built so that the full contents of the copyright-restricted materials are never exposed to the end users. Under this paradigm, scholars must rely upon descriptive metadata about the volumes within the corpora in order to assemble their worksets. In turn, they must also rely on the descriptive metadata about their worksets to make claims and fashion citations regarding their research data.

The overall, workbench-like workflow (Figure 2) being fashioned from the HTRC's web services will then enable scholars to submit their worksets to a number of analytics tools that operate within the HTRC context. Since it is expected that the various analytics tools will be both, provided by the HTRC and developed by the scholars themselves, it is necessary to develop a formal definition of what kind of entity a workset is within the confines of the HTRC's web services. As flexibility in what can be used to create worksets is one of the most desirable features, a workset's member items cannot be constrained to just the volumes comprising the HathiTrust corpus, but must also be able to

include corpora from outside of the HathiTrust and the research products that result from the HTRC's workflow.

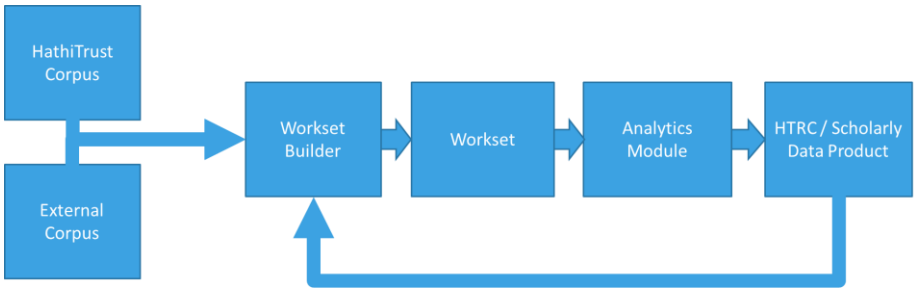

Figure 2: HTRC's Scholarly Workflow

Each workset is the outcome of a scholarly curatorial process and as such, it is a research product in its own right. From the scholar's perspective, each workset comprises a series of links which can be leveraged by various HTRC analytics processes to gather together all of the resources that the scholar intends to analyze and discuss. They provide what Palmer et al. (2010) call the "contextual mass" through which humanities scholarship is possible. In every way, a workset fulfills the same workbench role that a scholar's carefully curated collection of research materials does and so, we define worksets as a kind of research collection.

Beyond its essential nature as a kind of research collection, each workset is designed to play a crucial role in support of the HTRC's non-consumptive research paradigm. The non-consumptive research paradigm restricts a scholar's direct access to their research materials. The expectation is that all of the artifacts that they have gathered together into their workset will be analyzed by one of the analytics modules with the HTRC context. This is a necessary and assumed expectation, without which much of the potential of exploiting the HTDL's corpus as a dataset will not be realized. This is because some two-thirds of the corpus remains within the bounds of copyright protections.

In this context the workset plays the role of an intermediary structure between the scholarly user and the analytics processes. It provides a convenient structure which gathers together all of the scholar's research materials in one place, which a particular analytics module then ingests, retrieves all of the named artifacts, carries out its analysis, and reports back its findings.

1.2 USE CASES DERIVED FROM USER STUDIES

In addition to the primary institutional use case, a number of use cases have been derived from a series of interviews with scholars using HTRC resources to analyze portions of the HTDL corpus that WCSA researchers carried out in the Summer of 2013 (Fenlon et al., 2014). In conjunction with an earlier study (Varvel & Thomer, 2011) conducted at the University of Illinois at Urbana-Champaign, the user study findings provide a number of details that allow the development of additional use cases specific to various desired functionalities that worksets should be able to facilitate. The following listing provides details on user expectations and the capabilities of technology.

1.2.1 WORKSETS AS RESEARCH PRODUCTS (UC2)

As noted above, collection building is an important aspect of the scholarly process. Scholars are frequently interested in obtaining the oldest possible instances of editions for analysis. Prior to the late 1990s and early 2000s, this frequently meant that it took months or years of painstaking effort to acquire copies of works. When copies could not be procured, then scholars would need to either travel to institutions that had a copy in their stewardship or had to look for equivalent works to serve as exemplars. In the end, a scholar's research collection was the output of a great deal of curatorial effort on her or his part.

The advent of digital libraries helped to alleviate some of the stresses of accessibility and availability but brought with them a problem endemic in the catalogs of cultural heritage institutions – differing accounts of identical bibliographic resources. An example of this problem appears in Nurmikko-Fuller et al. (2015), which notes:

> "R. L. Stevenson's Weir of Hermiston illustrates this problem. HT[DL] metadata returns 13 distinct catalog records, six dating from 1896. Of these, four have New York as place of publication, one has London and one, Leipzig, with minor differences in page count and title." – (excerpted from Nurmikko-Fuller et al., 2015).

The increased ease of access and overall availability of resources has not significantly lessened the curatorial effort through which scholars go to assemble their research collections. The problem is not limited to the bibliographic resources themselves. Many digitized representations have their own metadata descriptions, frequently written by the entities responsible for the digitization.

After going to such great lengths, it stands to reason that scholars see their research collections as research products in and of themselves. As a product each workset should be a citable, immutable data entity. Since scholars gather their research objects from many different sources, worksets also need to able to accommodate bibliographic resources from outside of the HathiTrust context. Further, the architecture for assembling worksets needs to equip scholars with tools to facilitate reconciling different digital objects and differing accounts of those digital objects with one another so as to ensure the scholar's ability to find and use the objects most appropriate for her or his workset.

1.2.2 WORKSET MEMBER GRANULARITY (UC3)

With the advent of applying computer-mediated statistical analysis to large corpuses of text (Companion, 2004; Underwood, 2012), scholars can now carry out the kinds of large-scale analyses that their forbears only dreamed about. These new analytics have brought with them their own suite of challenges.

Before, when a scholar wanted to carry out an analysis of the works that they had gathered together, they could easily focus on just the parts of the text they had selected, ignoring those portions not pertinent to their work, e.g., front and end matter, table of contents, headers and footers, etc. Computational algorithms do not process text in a manner that is comparable to humans. They do not, for example, easily distinguish page headers and footers from the rest of the text on the page. They have no intrinsic knowledge of what comprises front matter, end matter, or a table of contents.

It takes a great deal of additional computational preparation, by experts writing complex algorithms, to zero in on the small granules that are the true units of a scholar's analysis or, as Fenlon et al. point out:

> "Units of analysis are the actual targets of scholars' analytic work: what kinds of things they aim to study, which correspond directly to the kinds of things they aim to collect. … For example, one respondent noted: 'It is very essential to work at the level of a particular chapter, with the actual text…We cannot talk so meaningfully about the work of a writer as a whole, in the abstract. The interpretation is based on actual text, at smaller units of analysis' (P7)." – (excerpted from Fenlon et al., 2014).

The system architecture and its underlying data model need to be able to support the scholar's ability to assemble worksets from a myriad of arbitrary bibliographic resources of differing levels of granularity. In addition to the ability for the scholar to pick out individual pages of books to analyze, the architecture must support the scholar's ability to pick out chapters, sections, paragraphs, and sentences, which might span multiple pages. Even more arbitrary divisions need to be supported, such as the ability to chunk text into 500 word blocks.

1.2.3 FACILITATING REPRESENTATION OF DATA IN GRAPH FORM (UC4)

One of the advantages of using semantic technology to record and represent data structures is to leverage those structures through named graphs. A graph representation allows data to escape from the silo-like confines of typical document structures that are serialized onto the web. Assertions that metadata records are typically used to make about bibliographic entities can instead be linked directly to those entities. Troublesome abstract entities that are packed into such documents, such as FRBR's work, expression, and manifestation, can be teased apart.

Graph representations also allow data from HTRC's outputs to be reused as metadata in new contexts. One example of such use is repurposing extracted page features (Organisciak et al., 2015), linking them as metadata to the pages they describe. This is a level of granularity that traditional library cataloging methods could never support under ordinary circumstances. Supplying metadata about finer granules helps to empower scholars with sufficient information to exploit them as objects to be curated in ensuing worksets.

Graph representations also reveal interesting new characteristics about objects within the HathiTrust corpus. Such representations frequently reveal relationships between entities that are not immediately apparent from the accounts of the metadata records about them. Such is the case in Figure 2 (below), where, in a recent article (Nurmikko-Fuller, 2015), it was found quite by accident that a book in the HathiTrust corpus was a copy of one that also appears in the EEBO-TCP corpus. What is most striking is through the process of microfilming, reprinting, and digitizing, two distinct metadata accounts emerged describing the exact same textual entity.

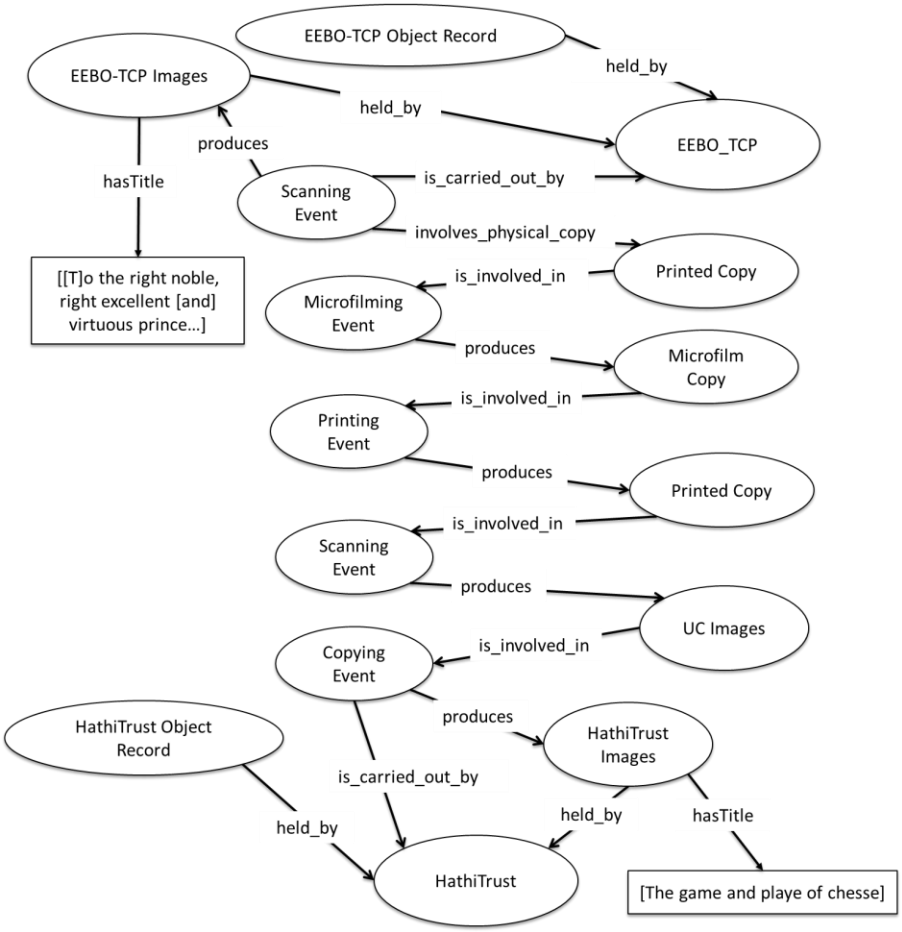

Figure 3: Provenance of Distinct Digitizations of a Book Found in Two Separate Repositories

The graph of copying events clearly illustrates that the two digitized copies hail from a single physical ancestor, which incidentally lives in the collections of the British Library. While the graph in the figure doesn't represent all of the relationships and entities that a full graph representation would capture, it does showcase how the metadata accounts of the book have diverged to the point that the two digitized copies are called by different titles.

2. THE REFERENTIAL PERSPECTIVE

In their 2013 paper, "What it is vs. how we shall: complementary agendas for data models and architectures," Dubin et al. argue that data models can be interpreted in two ways. On the one hand, a data model acts as a representation for a particular knowledge domain. On the other hand, it acts as a sweeping plan by which the day-to-day development activities of a Web service must conform. In this section we discuss the former account, reviewing previous work on formalizing collection definitions before arriving at a particular formalization appropriate for the HTRC's specific context. In the subsequent section (Section 3) we lay out a formal data model and suggest a series of extensions for it, which if implemented, begin to approach realization of the formalized *Workset* definition.

2.1 CHARACTERIZING WORKSETS AS A KIND OF COLLECTION

Despite the fact that a consistent and singularly authoritative definition for the notion of "collection" has yet to emerge (see Hill et al., 1999; Currall, Moss, & Stuart, 2004; Wickett et al., 2013a; Palmer et al., 2015; among others), the act of collecting resources is a key scholarly activity. A large amount of work contemplating the nature and usage of scholarly research collections has already been completed (Lynch, 2002; Currall, Moss, & Stuart, 2004; Palmer, 2004; Palmer & Knutson, 2004; Palmer et al., 2006). Recent research has revealed that, with regards to digital libraries at all scales, researchers need the means to bring heterogeneous digital objects together into one mass of research materials if they are to be able to engage in scholarly processes within today's digital landscape (Varvel & Thomer, 2011).

In addition to collecting disparate resources together for their personal research agendas, there is increasing evidence that indicates that scholars actively search for and exploit digital collections as distinct resources (Zavalina, 2010). It has been noted that the research collections themselves are becoming research products to be cited and reused in additional contexts (Palmer, 2004), as we noted in section one above. A recent survey of the HTRC's existing user base confirms that scholars see HTRC worksets in a manner that is consistent to prevailing views on digital collections. They specifically view them as citable research products in their own right and desire sufficient tools to relate their publications back to the sources from which research results are derived (Fenlon et al., 2014).

Worksets are gathered together in much the same way that other scholarly research collections are, which is to say that each one is the result of curatorial effort on the scholar's part. As is the case in many digital libraries, scholars using the HTRC's services employ a variety of database queries to generate the materials that they gather into their worksets. Each member of a workset is identified through a unique URI, which is frequently in the form of a URL and most typically a Handle.[4]

The following set of formalizations take the descriptivist point of view, describing a collection as a resource at a particular point of time. As such, the resulting formalizations will be best employed in developing description-based metadata vocabularies and data models as the authors have done in

[4] http://www.handle.net/

section 2. It would be remiss of us though not to point out that a narrativist point of view could have just as easily been taken, resulting in formalizations that are better suited to the development of event-based vocabularies and infrastructure. In practice, when developing robust infrastructure, both points of view need to be accounted for and their vocabularies and infrastructures employed to best effect for the sake of the end users' utility. It is not always clear how this can be done efficiently and the third section of this white paper grapples with those practical aspects.

2.1.1 FORMALLY DEFINING COLLECTIONS

Accepting the notion that worksets must be some kind of research collection, it stands to reason that what must first be produced is an adequate definition that serves as an identity condition for something to be a collection. One definition promulgated by the Dublin Core Metadata Initiative (DCMI) is as follows[5]:

> D1: If something, *x*, has been gathered into some other thing, *y*, then *y* is a collection.

In first order predicate logic this definition can be interpreted into the following axiom:

> A1: $\forall y\big(\exists x\ isGatheredInto(x,y) \leftrightarrow\ Collection(y)\big)$[6]

This formalization seems to satisfy the need to define collections as a kind of entity generally and conforms to some pre-existing notions of collections in digital libraries. However, if this definition is employed directly, then completely random collections might have to be admitted into any nascent Web service that is designed using the definition.

Recall that the ultimate intention is to define *worksets* such that they are a kind of *research collection*[7]. It seems likely that research collections are themselves, a kind of collection and not synonymous with the set of all things that are collections. D1 is not, in and of itself, sufficient for this task as it lacks any means for remarking on how *research collections* differ from *collections* in general. Some additional constraints are called for.

2.1.2 FORMALLY DEFINING RESEARCH COLLECTIONS

As noted above, a great deal of work on the nature of research collections, especially scholarly, research collections has already been accomplished. There are many relevant themes that interweave throughout the various accounts listed. Two particularly pertinent themes that emerge in

[5] http://dublincore.org/groups/collections/collection-application-profile/

[6] See, Renear et al. (2008) and Wickett (2009) who employ the property "isGatheredInto(x,y)" as an assumed identity characteristic for Collections in their work with collection / item relationships. See also Wickett et al. (2011) who provide a thorough explication of the kind of property that *isGatheredInto* is and from whose work Axiom A1 is directly derived.

[7] Note also that we have not remarked on the nature of *x*. It can easily be the case that *x* is, itself, a collection. This necessarily (and purposely) entails that collections are the kind of entity that can be gathered into one another, i.e., collections can be gathered together into bigger collections.

scholarly works discussing the nature of scholarly research collections (Lynch, 2002; Currall, Moss, & Stuart, 2004; Palmer, 2004; Palmer & Knutson, 2004; Palmer et al., 2006) are that:

- research collections are the products of curatorial effort, i.e., they are created by an entity through some means of selection, and
- research collections serve a specific role within a scholarly research workflow, i.e., they have some motivated purpose.

Of course, research collections are not the only kind of collection that result from the curatorial process of selection according to specific criteria. It seems likely that the requirements expressed in the first bullet above can be rephrased, narrowing the focus to collections that are the products of selection according to specific criteria. It can further be stated that such collections are a specific kind of collection, which, for want of better terms, we will refer to as curated collections.

Generating a new definition for this refinement of the general collection, produces the following:

> D2: 'If something, *x*, has been gathered into some collection, *y*, according to some set of criteria, $\mathbb{C}$, as defined by some agent, *w*, then that collection, *y*, is a curated collection.'

An axiom for D2 cannot be easily expressed in first-order predicate calculus. Since the criteria are best expressed as a set of things to which some function (e.g., some property, attribute, or factor) of *x* correlates, the expressivity of first-order logic is not sufficient to faithfully represent to complex relationship between functions and sets. As our ultimate goal is to use these axioms and definitions to suggest properties particular to *worksets* that any resulting metadata ontology will need to be able to record, we will set aside definition D2 and re-examine whether or not an approach using second-order logic will better articulate it at a future date.

We still need to provide a definition that is sufficiently simple enough that it can be expressed as an axiom in first-order predicate calculus and still be a sufficiently adequate representation of what a *curated collection* is. To fulfill this need we produce definition D2', which we freely admit is really more of a gloss than a proper definition.

> D2': 'If something, *x*, meets some criterion, *c*, and that criterion, *c*, has been defined by some agent, *w*, and it is also the case that that *x* has been gathered into some collection, *y*, then that collection, *y*, is a *curated collection*.

We can interpret this definition into the following, somewhat cumbersome axiom:

$$\text{A2': } \forall y \left(\exists x \exists c \exists w \left(gathered(x, y) \wedge meetsCriteria(x, c) \wedge definedBy(c, w) \right) \leftrightarrow CuratedCollection(y) \right)$$

One possible objection to A2' is that the curatorial process of selection seems to have been lost from the definition. From a narrativist's point of view this is a valid complaint. The theoretical model seems to be missing evidence that the actual selecting event took place. This can be ameliorated by further developing the axiom with descriptions of evidence for the selection event.

A2"a: $\forall c \left(\exists x \exists w \left(\left(meetsCriterion(x,c) \wedge definedBy(c,w)\right) \leftrightarrow selectedAccordingTo(x,w)\right)\right)$

A2"b: $\forall y \left(\exists x \exists w \left(gatheredInto(x,y) \wedge selectedAccordingTo(x,w)\right) \leftrightarrow CuratedCollection(y)\right)$

Now the descriptivist has a complaint. If we can substitute a single predicate that abstracts the two predicates concerning criterion *c*, then doesn't that imply that criteria themselves don't play a direct role in the formation of *curated collections*? We think that this *is* a valid complaint. The goal was to capture the criteria used in the curation of the *workset* in the workset metadata so that they might be shared with and assessed by other scholars (who may wish to reuse the workset in their analytic context). Axioms A2"a and b defeat this purpose and so it is our preference to use Axiom A2'.

Having arrived at a formalization that adequately describes curated collections as a kind of collection, the next step is to describe research collections as a kind of curated collection. Rephrasing the text in the second bullet (above) allows the creation of the following definition:

> D3: 'If something, *x*, has been gathered into some curated collection, *y*, for the purposes of some research motivation, *z*, then that curated collection, *y*, is a research collection.'

This produces the next axiom:

A3: $\forall y \left(\left(CuratedCollection(y) \wedge \exists z\, hasResearchMotivation(y,z)\right) \leftrightarrow ResearchCollection(y)\right)$

This produces an adequate definition for research collections in general but, here the astute reader may complain, what about scholars? The astute reader will have noticed how we have been writing about research collections as the products of scholars and for the purposes of scholars but, we have neglected to define what a scholar is.

We will continue that neglect. This is for three reasons:

1. Since being a scholar is a non-rigid property of agents, we are not certain if there are limitations with regards to which agents can or cannot be a scholar.
2. We are not convinced that scholars are the only kinds of agents that can validly create research collections, and do not want to preclude their creation by other kinds of agents, like students.
3. We are engaged in developing a formal model of worksets, not scholars; developing a formal model of scholars is simply out of scope.

Since it has already been established that worksets are a kind of research collection, we have very nearly arrived at a sufficient definition for worksets already. All that remains is the addition of some

constraint or constraints that are necessary to differentiate worksets specifically from other kinds of research collections in general.

2.1.3 FORMALLY DEFINING WORKSETS

Recall from the beginning that there is a primary use case around which all of the HTRC's putative and nascent infrastructure has evolved – that of the non-consumptive research paradigm. It is the expectation of the non-consumptive research paradigm that the workset will be consumed and its gathered contents operated upon by one of the HTRC's many analytics modules. The key differentiation of worksets from other kinds of research collections revolves around this paradigm-mandated expectation and our formal account of worksets must incorporate it. As a consequence, worksets are a kind of collection whose full utility can only be realized within the HTRC context. This is the price one pays for analytical access to billions of pages of materials that remain bound by copyright.

The resulting definition is as follows:

> D4: 'If something, x, has been gathered into a research collection, y, with the intention, a, that that y's contents be consumed by an automated process for analysis according to the non-consumptive research paradigm, then y is a workset.'

This produces the following axiom:

> A4: $\forall y \left(\left(ResearchCollection(y) \ \wedge \ intendedForUse(y, a) \right) \leftrightarrow Workset(y) \right)$, where a is the expectation that the contents of y will be consumed by an automated process for analysis in accordance with the non-consumptive research paradigm.

By substitution we can arrive at this more thorough and precise axiom.

> A4': $\forall y \Bigg(\Bigg(\Big(\big(\exists x \exists c \exists w \left(gathered(x, y) \ \wedge \ meetsCriteria(x, c) \ \wedge \ definedBy(c, w) \right) \big) \ \wedge \ \exists z \ hasResearchMotivation(y, z) \Big) \wedge intendedForUse(y, a) \Bigg) \leftrightarrow Workset(y) \Bigg)$, where a is the expectation that the contents of y will be consumed by an automated process for in accordance with the non-consumptive research paradigm.

Of course, the astute reader might have one final complaint with regards to context. The definition does not specify the HTRC context specifically, nor is it meant to. The primary constraint is the expectation for automated consumption of the workset's contents that the HTRC's non-consumptive paradigm mandates. The definition delivers that through expectation a. If the expectation seems

generic it is because we expect that many other contexts will have need of both the non-consumptive paradigm and will employ automated methods for analyzing collection content. There is no reason that definition cannot be portable beyond the HTRC's specific context.

2.2 OTHER KINDS OF COLLECTIONS

Constraining the essential nature of collections themselves to define specialized kinds of collections is not the only method by which specialization of collections can occur. We can also add various constraints to the things that are gathered into the collections, specializing collections according to their contents.

One such specialization that has been suggested in other ongoing collection formalization work is that of referential collections (Wickett et al., 2013a). In their report Wickett et al. distinguish between two types of collections – referential collections[8] and holdings collections. The former are loosely defined as collections of links, while the latter are distinguished by the existence of stewardship relationships between the entities gathered into the collections and the institutions responsible for them.

In the grant proposal submitted to the Andrew W. Mellon foundation, some effort was made to align the initial workset definition to the notion of referential collection. The above definition does not contradict this (it operates orthogonally to such constraints); however, an interesting dichotomy begins to reveal itself if worksets are also a kind of referential collection.

From the perspective of the scholar, each workset is a collection of links to collections of digital representations of pages in specific books. This is because of the nature of holdings collections, which Wickett et al. define around terms of stewardship of actual artifacts (digital or otherwise). Since, under the HTRC's existing workset architecture and infrastructure, the digital artifacts are physically stored within the HTRC's data architecture and each analytics module forges a copy (a "new" artifact) which is quickly ingested into its internal workflows.

From the perspective of the analytics modules, worksets are holdings collections over which they have sole stewardship of during the process of analysis. This in an important distinction because of the HTRC's non-consumptive research paradigm. The paradigm can be seen to be working when each workset appears to the scholar as a referential collection and to the analytics module as a holdings collection over which it has total stewardship. To use a popular analogy, the non-consumptive paradigm works when the scholar can see the cake and the analytics tools can eat the cake and then tell everyone how it tasted.

It is important to remember that the file artifacts being analyzed are extremely ephemeral and when the analytics module's work is completed, the workset is, again, for all intents and purposes, essentially a referential collection. The strange dichotomy though does indicate that either the notion of referential collection does not work as intended (as the essential nature of the objects within the

[8] Note that this use of the term "referential collection" is not the same as its use in music theory where it typically names a group or set of scales from which melodies and harmonies are drawn. See, for example, Pearsall, 2012, for additional information on the term's use in the music domain.

collection shift back and forth over time) or that the referential collection definition is not a good fit for worksets.

3. THE COHORTATIVE PERSPECTIVE

Having established the theoretical underpinnings of the workset concept, we can now consider the practical implications and begin developing a series of models that will facilitate the creation of infrastructure that supports the use cases expressed in Section 1. This is accomplished by first discussing the relationship between the use cases and the basic functionalities that they require infrastructure to accommodate. The specifics of the workset conceptual and representational models are then developed. A staged plan is then presented for the additional infrastructure and entity models necessary to realize all of the functionalities described in the technical requirements. Finally choice of implementation technologies, unresolved issues, and future avenues for expansion of the underlying data models are all discussed.

3.1. TECHNICAL REQUIREMENTS

From the use cases detailed in Section 1, we have derived a number of requirements that the resulting workset model and its supporting architectures must fulfill. These requirements are that:

- A workset is a container for a scholar's aggregated units of analysis – analogous to a scholar's research collection and constrained by the non-consumptive paradigm's expectation for usage.
- A workset is a persistent, globally unique entity that can be directly cited.
- A workset possesses provenance properties supporting change awareness within the HTRC context so that a description of its nature at the time of analysis persists over time.
- A workset's membership requirements must be flexible enough to allow for the arbitrary aggregation of heterogeneous resources, with regard to:
    - Granularity of resources that will be considered a unit of analysis and
    - Source from which a particular member entity is retrieved.
- A workset possesses a number of properties which, when expressed in the form of metadata, facilitate its identification, selection, citation, and use. These properties are informed by three sources:
    - The formalization of the workset notion, e.g., that it is formed by things being gathered together.
    - The properties that are intrinsic to the workset at a particular point of time, e.g., the number of things gathered into it.
    - The properties that describe the things gathered into the workset, e.g., the languages of the things gathered into it.

3.1.1. WORKSETS AS CONTAINERS (TR1)

HTRC analytics modules are expected to accept a listing of resources to be analyzed, retrieve them, carry out the desired analysis, and then report back the results to the scholar who assembled the workset. They are not merely a list though. Their creation is the purposeful result of curatorial efforts on the part of the scholar and in every way, a workset acts as a container for resources of interest to the scholar.

Whether the scholar used a database query to gather together her or his materials, or each exemplar was painstakingly acquired over a long period of searching, the resources that are contained within each workset are not a random collection of materials. Not only does each workset contain the bibliographic resources that the scholar has gathered together but it also records important metadata regarding when the scholar gathered them together and for which analytics tool package they are intended, capturing important contextualizing information about the aggregated whole.

The workset entity allows both the scholar and the HTRC architecture to treat the aggregated whole of the scholar's efforts as a single entity. The scholar can annotate the whole, cite it as a dataset, and even repurpose it through submission to additional analytics processes or selection as part of a new workset.

This requirement supports all four use cases, UC1, UC2, UC3, and UC4.

3.1.2. WORKSETS AS GLOBALLY UNIQUE, PERSISTENT ENTITIES (TR2)

As noted in the use cases above, scholars perceive of their research collections as research products in and of themselves. To instantiate this in the HTRC milieu, each workset needs a globally unique, persistent identifier. This identifier directly supports the ability of scholars and other researchers to cite and otherwise publish references to the underlying dataset.

To ensure persistence of the entity over time and to match the expectation that worksets are immutable entities, some form of versioning will need to be applied at both the workset and bibliographic resource levels. Such measures will ensure the stability of the workset as a citable data resource. There are further implications for the persistence of the entities comprised by the workset which are discussed in further detail below.

This requirement supports UC1, UC2, and UC4.

3.1.3. WORKSET PROVENANCE PROPERTIES (TR3)

Aligned with the above requirement, our expectation is that worksets will evolve and change over time as scholars select new materials for inclusion and old materials for exclusion. The basic bibliographic resources, OCR text files and scanned image files, are also known to experience periodic changes as new OCR and new scans become available.

It is important that workset provenance metadata records significant events in the workset's lifecycle. In addition to ordinary events, such as the workset's creation, the provenance aspect of the workset model must adequately support the capture of events that are silent and invisible with respect to end users. Specifically, since the basic workset is expected to gather together HTRC volumes, workset provenance metadata must adequately capture changes in any of the workset's constituent members so that, at a minimum, those who cite the workset as a dataset are aware that the underlying data is different from the dataset that previous citations referred to.

This has clear implications for the provenance metadata requirements of the entities contained within the workset, which are described in further detail below.

This requirement supports UC1, UC2, and UC4.

3.1.4. WORKSET MEMBER GRANULARITY AND SOURCE (TR4)

As has already been described, scholars using HTRC resources do not want to be limited in their selection of analysis worthy materials to just those things that the HathiTrust Digital Library possesses. Generally speaking, this should be possible so long as the resources from outside entities are identified via identifiers that are resolvable resources and that can be ingested into and processed by the analytics tools that the scholar means to examine. There may be additional requirements made with regards to the descriptive and provenance metadata possessed by such resources. As such they will need to be either already conformant to the requirements listed below or be submitted through an HTRC workflow which can add sufficient detail that they then conform.

In addition, there are clear needs for better granularity measures. At the time of the writing of this report the existing functionality within the HTRC's extant infrastructure is such that worksets can only accommodate whole volumes. Which is to say that a workset only accommodates abstract, aggregate entities that serve as containers for ordered sets of page entities.

For a start, it would be advantageous to be able to gather together individual page entities so that analytics tools can be made more efficient. Beyond this, the use case detailed in UC2 clearly demonstrates that there is a desire among scholars for more control over what can be used as a unit of analysis.

This requirement operates in direct support of UC2.

3.1.5. PROPERTIES SHARED BY WORKSET AND BIBLIOGRAPHIC RESOURCES (TR5)

Among the scenarios that are expected within the HTRC milieu is that scholars will reuse one another's worksets or portions of those worksets for a variety of research scenarios. To facilitate workset findability on the basis of workset contents, workset metadata must adequately capture properties that propagate to it from its constituent bibliographic resources. As noted above, much of the work formalizing the rules by which propagation of properties obtains has been described in Renear et al. (2008a, 2008b), Wickett et al. (2010), and Wickett (2012).

This requirement supports UC1.

3.2. WORKSET REPRESENTATION & DESCRIPTION

The following conceptual model (Figure 4) has been derived with the requirements discussed in the previous section in mind. The essence of worksets is simple enough. They are containers which gather together resources and serve as inputs to analytics processes. Strictly speaking the conceptual model can be extremely succinct – an entity, with some properties and a pair of relationships to other entities.

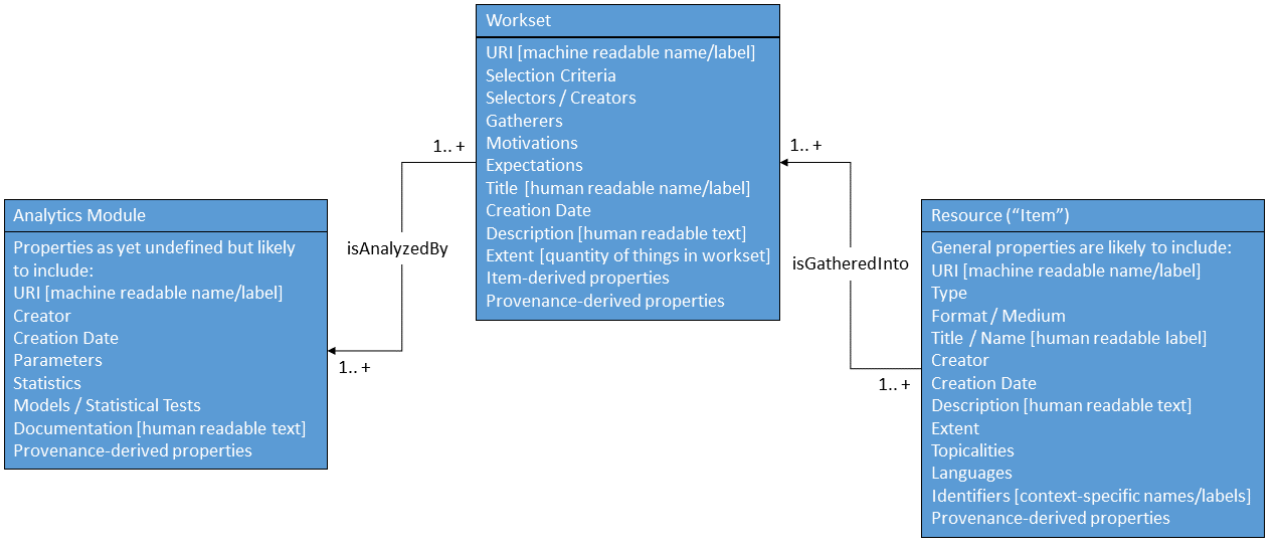

Figure 4: Simple Conceptual Model for Worksets

The key feature of worksets is their relationship to their constituent items. The relationship "isGatheredInto" is the core around which worksets are built. This aligns well with the HTRC's existing notion of worksets, which is that they are containers for volumes with some additional metadata.[9] The primary difference between the conceptual model in Figure 4 and the one that underlies the xml schema document is the domain of what can be gathered into the workset.

In the case of the existing schema, an assumption was made that scholars would only want to gather HTRC volumes (i.e., books) into worksets for analysis. The conceptual model derived from the use cases makes no such assumptions. The domain of things that can be gathered into worksets is any resource on the Web. For purely practical reasons, the metadata model described below constrains the domain to just volumes. This is because much of the infrastructure to leverage finer grained entities and entities of different kinds still needs to be developed. Some of this infrastructure will be suggested and outlined in the following pages but the quintessential reality of the model's application is that the existing infrastructure can only cope with volume-level entities.

Work on modeling collections within various contexts is an ongoing process and a number of "isGatheredInto"-type predicates have emerged. The oldest of these appears in the guise of the DCMI's Collection Application Profile which uses the more generic predicate, dc:isPartOf, to gather a collection's items together. In an attempt to expand on the DCMI-CAP a group of researchers have been developing a collection model for the Europeana digital library (Wickett et al., 2014). As part of this work, they have suggested an edm:isGatheredInto predicate which seems to meet the HTRC's needs rather nicely.

In fact much the Europeana Data Model (EDM)'s seems as though it can easily be employed within the HTRC context. However, as there are a great number of expectations for the behavior of data

[9] See Appendix A for the HTRC's existing workset xml schema document.

within the HTRC context and because it is not a metadata aggregator like Europeana, many of the following recommendations use broader-scoped, more generic predicates whenever possible. The issue of alignment with other large scale digital library initiatives should be re-examined at such a time as the HTRC's infrastructure has finally realized the vision laid out here. For this reason, the only EDM-specific predicate that will be employed by the subsequent data model is edm:isGatheredInto.

The diagram in Figure 4 also suggests several attributes that are important features for describing worksets. The most critical of these are the workset's HTRC-based name – its identifier. This label is intended to uniquely identify the workset both within and without the HTRC's context.

Table 3: Comparison of Workset Attributes

Workset Elements (from existing workset.xsd)	**Workset Properties (from new conceptual model)**
identifier	*identifer*
version	various provenance-derived properties
	creation date
name	title / name
description	description
author	creator / selector
	gatherer
rating	
average rating	
last modified date	various provenance-derived properties
last modified by agent	various provenance-derived properties
volume count	extent
public	accessibility
	language(s)
	motivation(s)
	expectation(s)

The above table compares the HTRC's existing workset schema to a newly proposed model. Several of the properties (identifier, creation date, and extent) are implied by the intrinsic nature of the technical infrastructure that worksets are a part of. Several of the properties (creator/selector, gatherer, motivation, and expectation) are suggested by the formal definition of worksets proposed by axiom (A4'). Several of the properties (accessibility and language) are derived from properties possessed by the workset's member items. Several properties (title/name and description) are necessary to meet the needs of human users. Finally, properties having to do with the workset's long-term provenance (e.g., date last modified) can be derived from its relationships to the various provenance infrastructures of the resulting technical infrastructure.

The following sub-sections detail the specifics of the workset data model to which any resulting implementation must conform. While it has been the author's intention that an implementation using RDF-based technologies be the ultimate result of the model that is subsequently described, it can easily be instantiated using more traditional digital library technologies.

3.2.1. WORKSETS

As described above, a *Workset* is a kind of research collection specific to the HTRC context. A Workset is an information resource and SHOULD have a URI. All collections produced within the HTRC's workset builder architecture MUST be instances of the class htrc:Workset. As a collection, a Workset gathers together a group of other entities which stand in relation to it as *items* though the edm:isGatheredInto predicate.

The domain of edm:isGatheredInto is normally rdfs:Resource (i.e., any RDF resource) but for practical considerations we anticipate most working implementations will constrain this according to the needs of their local contexts. In the HTRC case, an initial implementation constrained to Volumes (described more fully below) will suffice to demonstrate the model's viability. The range of edm:isGatheredInto is dcmi:Collection and htrc:Workset is understood to be a sub-class of dcmi:Collection.

Vocabulary

Entity / Property	Type	Definition
htrc:Workset	Class	A sub-type of dcmi:Collection with an additional Expectation constraint. The htrc:Workset class MUST be associated with a Workset.
edm:isGatheredInto (reciprocal edm:gathers)	Relationship	The relationship between a Collection and an *item* that has been gathered into it. There MUST be 1 or more edm:isGatheredInto relationships associated with a Workset.

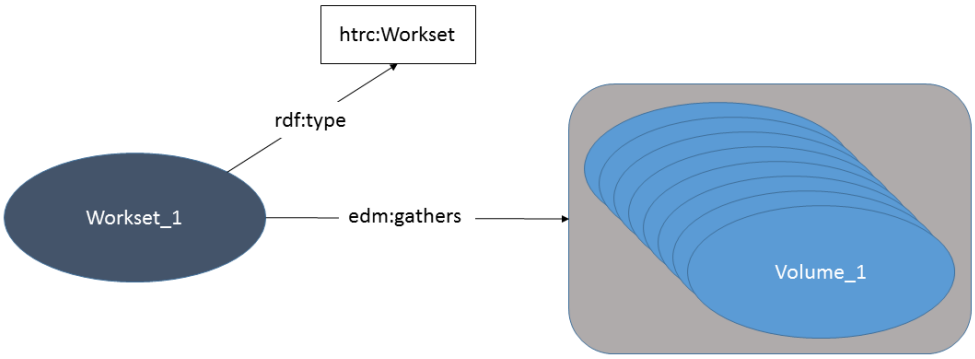

Figure 5: Basic Workset Model

3.2.2. WORKSET DESCRIPTIVE METADATA

To fulfill the various technical requirements and use cases described above, a variety of descriptive metadata is necessary. Some of this metadata is intrinsic to the very nature of Worksets as they have been defined above, some is intrinsic to the digital architectures that Worksets are expected to play a role in, and some metadata explicitly supports various kinds of exploitations made by humans.

3.2.2.1. WORKSET METADATA INTRINSIC TO WORKSETS

The first kind of metadata, that intrinsic to the nature of worksets themselves include the Criteria by which *items* are selected for inclusion in the Workset, the Agents which determine the Criteria and the Agents which do the gathering of the *items*, which may or may not be the same, and finally the Motivations and Expectations that detail the Workset's intended role as a research product.

With the realization that much of this metadata requires data entry on the part of the scholarly user making the Workset, with only two exceptions, all of these properties are expressed in terms of SHOULD rather than MUST.

A Workset MUST be related to 1 or more Agents who are solely responsible for defining the curatorial criteria according to which the *items* within the Workset have been gathered. These Agents are related to the Workset through the dcterms:creator predicate.

A Workset SHOULD be related to 1 or more Criteria that adequately capture the curatorial criteria that the Agent responsible for the Workset's creation has defined. There are no constraints with regards to how the criteria may be expressed. They may, for example, be expressed as skos:Concepts, as text descriptions, or as programmatic queries (e.g., a SQL Query). This relationship is expressed using the htrc:hasCriterion predicate.

A Workset MAY be related to exactly 1 agent responsible for the physical act of gathering together the *items* comprised by the Workset. It may be the case that this entity is the same as the one defining the Workset's selection criteria or it may be, among other possibilities, some form automated agent that programmatically applies the selection criteria and returns results (e.g., it could be a data API). This relationship is expressed using the dcterms:publisher predicate.

A Workset SHOULD be related to 1 or more Motivations that describe its research context. Similar to Criteria, these Motivations may be expressed in several forms, including skos:Concepts, text descriptions, etc. The relationship between a Workset and its Motivation is expressed using the htrc:hasMotivation predicate.

A Workset ALWAYS HAS at least 1 Expectation to which it is related. This Expectation is that the Workset's items are such that they can be analyzed by some automated analytics workflow. Since this Expectation is homogeneous across all Worksets, it may be inferred by the fact that an entity is an instance of a Workset (and so no manual entry on the part of an Agent is necessary). Within the context of the HTRC this Expectation is the set of all Analytics Modules.

A Workset MAY be related to additional Expectations. These additional Expectations all express additional constraints regarding which kinds of analytics modules that a Workset is intended to work with. This helps the HTRC analytics workflows avoid collisions in cases of data / algorithm mismatches. For instance, a Workset that gathers together image files will be inappropriate for an algorithm designed to analyze text data content. Scholars using the Workset Builder tools should be encouraged to select one or more analytics modules that are most appropriate for the analyses they desire to take place. The Expectation relationship is expressed through the htrc:hasExpectation predicate.

Vocabulary

Predicate	Domain	Range	Cardinality
dcterms:creator	htrc:Workset	dcterms:Agent	1+
htrc:hasCriterion	htrc:Workset	rdfs:Resource or rdfs:Literal	0+
htrc:hasResearchMotivation	htrc:Workset	rdfs:Resource or rdfs:Literal	0+
htrc:intendedForUse	htrc:Workset	rdfs:Resource[10]	1+[11]

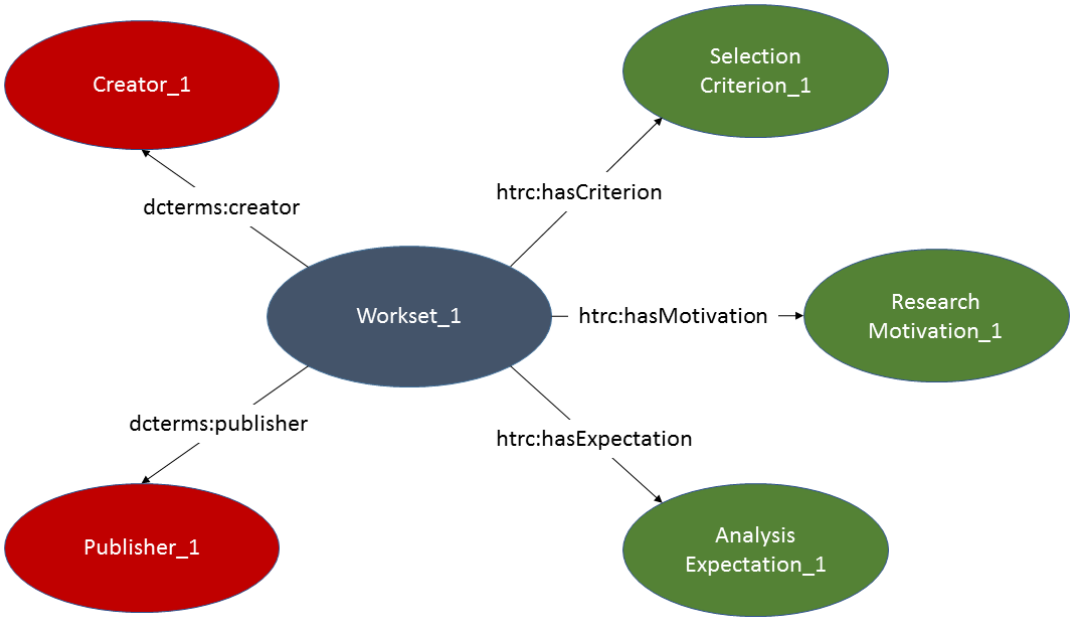

Figure 6: Metadata Intrinsic to Worksets

3.2.2.2. WORKSET METADATA INTRINSIC TO DIGITAL ARCHITECTURES

The second kind of Workset metadata encompasses the kinds of things that computers are very good at. Specifically these are counting how many things have been gathered into a Workset and recording when it was first created or published. As such the following two metadata relationships do not express things that the scholarly end user necessarily needs to be aware of. Instead they describe more specific technical requirements that an implementer must be conscious of during the development cycle.

[10] It is not the case that any RDF resource can be used as an Expectation; however, it is not the purpose of this technical report to formalize a definition for the entities evinced by analytics modules, pipelines, processes, or workflows. For practical purposes, any resulting implementation will need to mint identifiers for these things in such a way that the range of htrc:hasExpectation can be constrained without reducing its validity to only those analytics tools within the HTRC context. To this end a general model for analytics modules, pipelines, processes, tools, and workflows needs to be developed.

[11] As noted, the one mandatory Expectation can be inferred from the instantiation of the Workset itself and no physical implementation of anything representing this tacit fact need make its way into any resulting infrastructure.

A Workset MUST be related to an xsd:integer that expresses the number of *items* that have been gathered into it. This integer SHOULD be produced programmatically whenever a metadata description of a Workset is required by an Agent. The relationship between the Workset and the integer is expressed using the dcterms:extent predicate.

A Workset MUST be related to an xsd:date that expresses the exact time at which an Agent (which may be either the Agent defining the selection Criteria or the Agent that gathers the *items*) first begins to assemble the Workset. The relationship between the Workset and the date on which it was created is the dcterms:created predicate.

Vocabulary

Predicate	Domain	Range	Cardinality
dcterms:extent	htrc:Workset	xsd:integer	1
dcterms:created	htrc:Workset	xsd:date	1

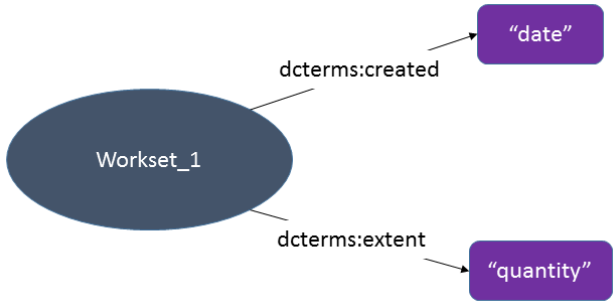

Figure 7: Metadata Intrinsic to Digital Architecture

3.2.2.3. METADATA FOR HUMAN-CENTRIC INTERACTIONS

Among the use cases that this data model answers to are requirements that Worksets be the kind of entities that are citable. To some extent, citation is a very human oriented activity and comes with some expectations that the Worksets might be repurposed for various reuse scenarios. The implication is that they must be findable, preferably in the ordinary manners that humans employ. Again, there must be an admission of the limits of patience that the typical scholar will have for the entry of otherwise helpful metadata, entailing that not all of the following properties be mandatory.

A Workset MUST be related to a Name / Label beyond the URI that the digital infrastructure will be referring to it by. This name is to be expressed as an xsd:string and is related to the Workset through the use of the dcterms:title predicate.

A Workset SHOULD be related to a description (e.g., an xsd:string, a web-page, etc.) that human beings can exploit to gain a better sense of what the Workset contains and for what purposes it was brought into being. Among other things, free-text descriptions are helpful for expressing descriptive metadata that often goes unexpressed due to want of defined spaces within data models (Zavalina et al., 2008). The relationship between the description and the Workset is expressed through the dcterms:abstract predicate.

Vocabulary

Predicate	Domain	Range	Cardinality
dcterms:title	htrc:Workset	xsd:string	1
dcterms:abstract	htrc:Workset	rdfs:Resource or rdfs:Literal	0 or 1

Figure 8: Human-centric Metadata

3.2.2.4. WORKSET METADATA DERIVED FROM ITEMS

A series of formalized rules that describe the conditions under which attributes and attribute values propagate between collections and the items in them have already been described (Wickett, Renear, & Urban, 2010). These rules should be leveraged during the implementation phase to further reduce the amount of data entry labor expected of the scholars building the worksets. The rules set forth by Wickett, Renear, and Urban suggest that all Collections, including Worksets, possess the following properties, whose values can be derived from related properties possessed by their *items*.

A Workset MUST be related to the language or languages of the *items* gathered into it. The relationship is expressed through the dcterms:language predicate and conforms to the following rule, expressed here in first-order predicate calculus.

A5: $\forall y \forall z \left(\left(language(y,z) \ \wedge \ Collection(y) \right) \rightarrow \ \exists x \left(isGatheredInto(x,y) \wedge language\ (x,z) \right) \right)$ – excerpted from Wickett, Renear, & Urban (2010), p 6.

A Workset MUST be related to a temporal range that indicates its temporal scope. The relationship is expressed through the dcterms:temporal predicate and conforms to the following rule expressed here in first-order predicate calculus.

A6: $\forall y \forall z \left(\left(temporalCoverage(y,z) \ \wedge Collection(y) \right) \ \rightarrow \ \exists x \left(isGatheredInto(x,y) \ \wedge \ \exists w \left(date(x,w) \wedge temporalWithin(w,z) \right) \right) \right)$ – excerpted from Wickett, Renear, & Urban (2010), p 7.

A Workset MAY be related to the kinds of the materials that have been gathered into it. This relationship is expressed through the presence of an additional Class typing of the Workset, e.g.:

:workset1 rdf:type htrc:ImageCollection .

In terms of the rules that Wickett, Renear, and Urban suggest, this typing is the end product of generalizations made about the types of the *items*. It conforms to the following rule expressed here in first-order predicate calculus.

A7: $\forall y \forall z \left(\left(itemType(y,z) \wedge Collection(\mathrm{y}) \right) \rightarrow \exists x \left(isGatheredInto(x,y) \wedge \exists w \left(type(x,w) \wedge generalizes(w,z) \right) \right) \right)$ – excerpted from Wickett, Renear, & Urban (2010), p 7.

Table 2 (below) sets forth some of the commonly expected type values for *z* in the axiom above.

A Workset MAY be related to the format or formats of the *items* gathered into it. The relationship is expressed through the dcterms:format predicate and conforms to the following rule, expressed here in first-order predicate calculus.

A8: $\forall y \forall z \left(\left(itemFormat(y,z) \wedge Collection(y) \right) \rightarrow \exists x \left(isGatheredInto(x,y) \wedge format(x,z) \right) \right)$ – excerpted from Wickett, Renear, & Urban (2010), p 8.

Vocabulary

Predicate	Domain	Range	Cardinality
dcterms:language	htrc:Workset	rdfs:Resource or rdfs:Literal	1+
dcterms:temporal	htrc:Workset	rdfs:Resource or rdfs:Literal	1
dcterms:format	htrc:Workset	rdfs:Resource or rdfs:Literal	0+

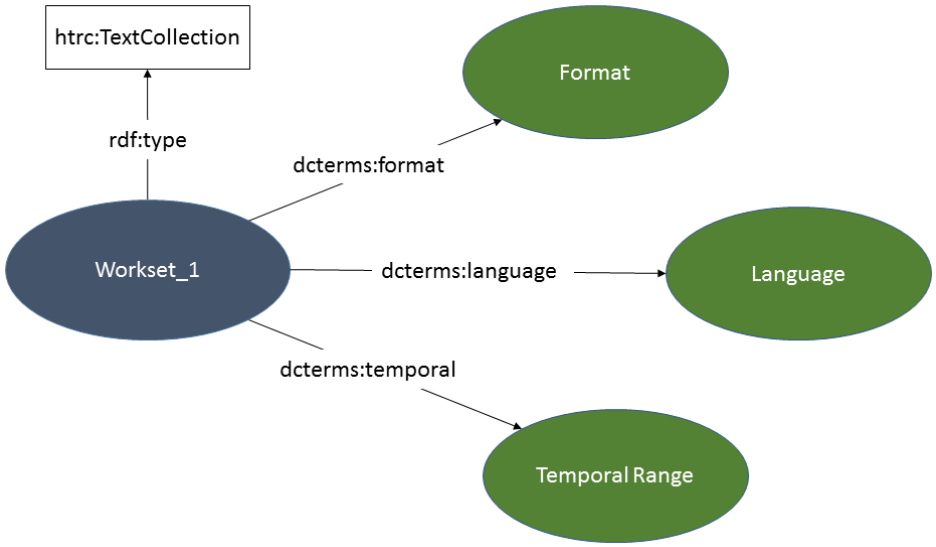

Figure 9: Metadata Derived from Workset Members

Table 4: Some common types of collections according to content[12]

Entity	Type	Definition
htrc:TextCollection	Class	A collection of works expressed by representations of text.
htrc:ImageCollection	Class	A collection of works expressed by representations of images.
htrc:AudioCollection	Class	A collection of works expressed by representations of audio.
htrc:MediaCollection	Class	A heterogeneous collection of works expressed by representations in two or more different kinds of media.
htrc:VideoCollection	Class	A collection of works expressed by representations of moving images.
htrc:GameCollection	Class	A collection of works expressed by representations of games.

3.2.3. WORKSET PROVENANCE METADATA

Throughout much of the WCSA grant proposal is the notion that Worksets are immutable things. This idea stands in direct opposition to our notion of collections which are things that gain and lose members over time. The best way to keep track of this is through some versioning apparatus which allows older versions of Workset graphs to be accessed and cited at later times. The specifics of such apparatuses may be grounded in either the architecture's underlying data model, in the technology that is employed to implement it, or in some combination of the two.

Rather than develop an additional data model specification, it is the recommendation of this report that the Workset model be extended with an existing provenance model. An event-based model will likely prove to be the most effective. Of these there are three existing vocabularies which may be best fits: FRBRoo, the PROV Ontology (PROV-O),[13] and the Systematic Assertion Model (SAM).[14] As discussed further below with regards to *item*-level metadata, FRBRoo,[15] is a specialized extension of CIDOC-CRM. [16] In the HTRC context it would provide vocabulary to both preserve various events in the lifecycles of bibliographic resources and represent the higher-level abstract entities of FRBR, such as Work, Expression, and Manifestation. However, FRBRoo also brings with it CIDOC-CRM's entire suite of vocabulary for describing entities. This may not be appropriate as descriptive metadata is likely to be captured through vocabularies that are specific to each kind of bibliographic resource, and despite it thoroughness, there are several levels of granularity that CIDOC-CRM does not capture details about entailing the need for more specific metadata models. Thus a wholesale application of FRBRoo would likely create a large amount of redundant data.

PROV-O is a recommendation for a provenance specific vocabulary developed at the World Wide Web Consortium (W3C)[17] and is specifically designed to capture the kind of silent events that occur in the HathiTrust Digital Library when page-level file objects are replaced. Since it is specialized for

[12] This table is not meant to be an exhaustive listing of content types, which is outside of the scope of this report. The larger HTRC community will need to carefully examine this issue and develop a more nuanced listing with such additional content collection types as it deems useful to the activities of the whole.
[13] http://www.w3.org/TR/prov-o/
[14] Cf Wickett et al. (2012a) and Wickett et al. (2013b) for details on SAM.
[15] http://www.cidoc-crm.org/frbr_inro.html
[16] http://cidoc-crm.org/
[17] http://www.w3.org/

web architectures and it is relatively "light-weight" (see Figure 11 below), as ontologies go, it may be the most suitable for capturing both versioning information with regards to worksets and bibliographic resources, and chaining together similar or familial bibliographic resources through shared events in their lifecycles as illustrated far above in Figure 3.

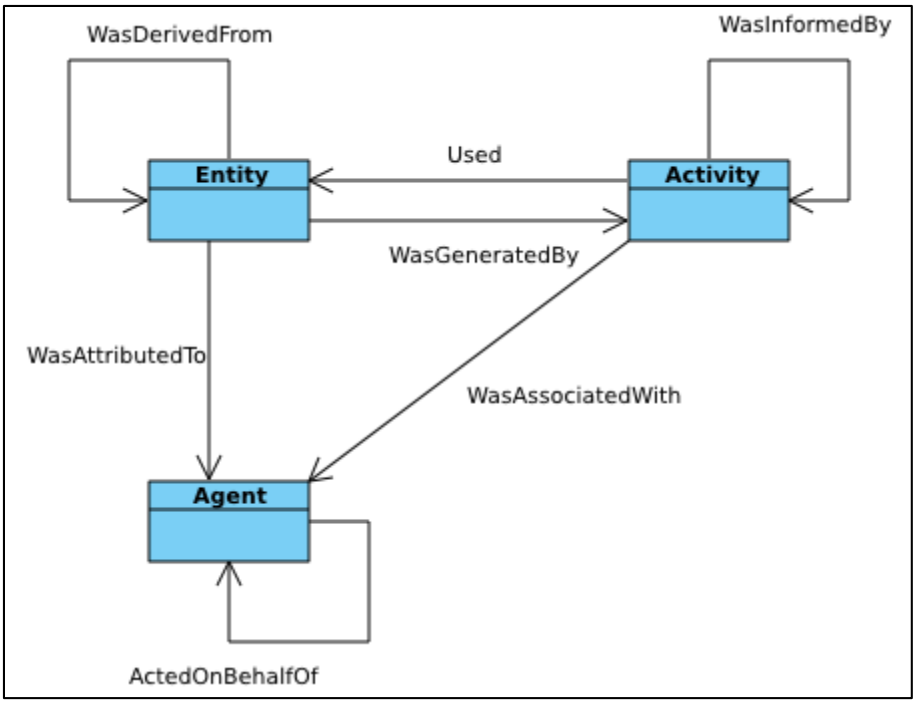

Figure 11: PROV Core Structures (Informative)[18]

SAM is the outgrowth of research efforts examining the essential nature of science datasets. Somewhat more concerned with capturing and providing sufficient amounts of interoperability metadata to adequately facilitate reuse of scientific datasets, in their 2013 RDAP[19] poster, Wickett et al. describe how SAM can be extended and applied to humanities computing data. SAM treats events within the lifecycle of various data resources with a bit more specificity than PROV-O enabling the system architecture to supply additional information to scholars that may inform their confidence in the authoritativeness of particular file objects, empowering them with more tools for precisely selecting the contents of their worksets. This additional functionality comes at the cost of increased verbosity, making implementation of SAM more challenging and requiring a larger amount of storage to adequately capture the provenance metadata that it records.

Of the three of these, PROV-O provides the best functionality for the least impact on any ultimate implementation. It does require an additional type assertion be made for every Workset, i.e.:

:workset1 rdf:type prov:Entity .

[18] Excerpted from PROV-DM: the PROV data model (2013); Accessible at: http://www.w3.org/TR/2013/REC-prov-dm-20130430/

[19] Research Data Access & Preservation Summit (this is an ongoing conference that focuses on issues of data access and preservation).

Any implementation of PROV-O will be able to work in coordination with any versioning structures inherent to the architectural platforms that are employed to implement a Workset Builder based on the Workset model. In the case that the architecture has no affordances for versioning then implementation of PROV-DM, as an extension to the Workset data model, can fulfill requirements for the immutability of citable entities.

3.3. ROLES OF ITEM-LEVEL METADATA AND DESCRIPTION

The technical requirements necessary for worksets also inform those needed to adequately describe the properties that describe its member *items*. Under ideal circumstances the infrastructure resulting from the implementation of the above Workset data model would facilitate the inclusion of any kind of resource. However, much of the technology and many of the techniques needed for such an implementation have yet to be fully developed and, as the HTRC already possesses a fair amount of existing infrastructure, any initial implementation of this data model needs to at least support the kinds of bibliographic resources that the HTDL's corpus contains.

The use cases and technical requirements listed above suggest the need for additional technical requirements at the level of the *items* being gathered into the Worksets. The requirements for these bibliographic resources are as follows:

- A bibliographic resource is a persistent, globally unique entity that can be directly cited;
- A bibliographic resources possesses provenance properties that support both:
    - Change awareness within the HTRC context so that a description of its nature at the time of analysis persists over time and,
    - Awareness of events within the bibliographic resource's lifespan that facilitate its disambiguation from other, similar bibliographic resources, i.e., in support of deduplication, finding the first printings of first editions, etc.; and
- A bibliographic resource must possess a set of metadata rich enough to support its discovery through means of various types of filtration, e.g., if the text transmitted by it is in English then it possess the property of being in English.

3.3.1. BIBLIOGRAPHIC RESOURCES AS GLOBALLY UNIQUE, PERSISTENT ENTITIES (TR6)

Much like worksets, bibliographic resources within the HTRC context must also have globally unique, persistent identifiers. Such identifiers facilitate stability of each workset's underlying data. Like worksets, versioning controls will be necessary to ensure the fidelity of bibliographic resource identifiers. Identifiers will also need to be minted for finer grained entities such as pages or other arbitrary chunks of content, so that scholars who desire more specific kinds of bibliographic resources than whole volumes can be adequately supported. Because some of the granules are very arbitrary in size, versioning at the level of the ingested files is going to be necessary in order to avoid cascading changes in underlying data.

It is an important factor that the use of proxies as workset *items* be discouraged. The reason is twofold:

1. The use of proxies works in opposition of linked data principles where the goal is to link directly to data resources.

2. The use of proxies adds an ambiguous layer of indirection which will be very difficult for analytics modules to accommodate without a great deal of additional engineering.

This requirement supports UC1, UC2, UC4, TR2, and TR3.

3.3.2. BIBLIOGRAPH RESOURCE PROVENANCE (TR7)

As Nurmikko-Fuller et al. (2015) showcased, researchers need a full suite of metadata that deeply describes the provenance of particular bibliographic resources. Such metadata ensures that scholars are able to select the most appropriate resources for their worksets. It also supports a more general change awareness within the workset. Important events in the bibliographic resource's life cycle, such as an OCR text file being superseded, can be captured, recorded, and propagated to the workset entity. This ensures that the data used by the scholar remains stable and citable. Without such measures the overall robustness of the Workset data model will be greatly degraded, making it difficult to cite worksets as unique data products in their own right and impinge upon the ability of scholars to remark on one another's work as reproducible science.

This requirement supports UC1, UC2, TR2, and TR3.

3.3.3. DESCRIPTIVE METADATA FOR BIBLIOGRAPHIC RESOURCES (TR8)

Bibliographic resources must also possess sufficient metadata to allow users to group them by various properties. The most basic level of bibliographic resources – *volumes* in the HTRC context – already possess descriptive metadata in the form of MARC records. Unfortunately many older manuscripts frequently have multiple records that describe their features. The resulting architecture must be able to mine and reconcile the assertions contained within these existing descriptions.

Additionally, the HTRC has already begun building a large store of descriptive metadata at the level of individual pages – the Extracted Features Dataset (Oganisciak et al., 2015). As it becomes available, descriptive metadata needs to be attached to bibliographic resources at every level of granularity. This both expands the options available to researchers and feeds project outputs back into the ecosystem of the whole, allowing the HTRC to realize the benefits of research taking place within its milieu.

This requirement supports UC2, UC3, UC4, TR3, TR4, and TR5.

3.4. THE BIBLIOGRAPHIC RESOURCE DATA MODEL

As noted throughout this document, scholars require access to metadata whose quality and scope is sufficient to facilitate the various workset gathering activities outlined in the technical requirements above. The metadata necessary to meet these needs comes in a variety of kinds.

As we noted above, despite our desire that worksets be able to gather together any kind of resource, a great many barriers need to be overcome. The following model for Bibliographic Resources is intended to provide a foundation upon which Worksets can be extended upon until such a time as the

HTRC's technological infrastructures have evolved beyond their current states. The following essential features of a nascent model of Bibliographic Resources are detailed as follows.

A *Bibliographic Resource* that is gathered into a *Workset* is an information resource and SHOULD have a URI. All Bibliograpic Resources gathered into Worksets MUST be instances of the class dcterms:BibliographicResource. In order to better distinguish among the various abstract entities and granularities that can be described a series of sub-classes of dcterms:BibliographicResource are in the process of being developed. The first two such sub-classes are the classes, htrc:Volume and htrc:Page. Where appropriate, Bibliographic Resources gathered into Worksets SHOULD INSTEAD be instances of the htrc:Volume or htrc:Page classes rather than the dcterms:BibliographicResource class.

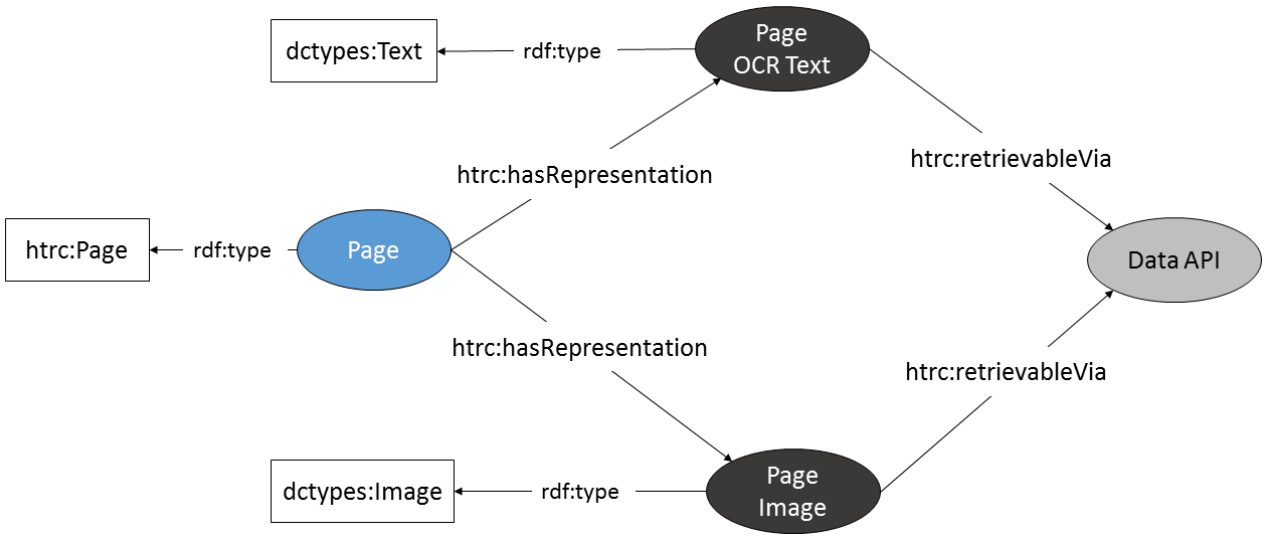

Figure 12: Basic Bibliographic Resource Data Model

Additionally, all Bibliographic Resources possess properties that facilitate their consumption by the HTRC's analytics modules. All Bibliograpic Resources gathered into Worksets MUST (for now) be related to at least 1 representation in the form a computer file that that can be ingested into the analytics module's workflows. This relationship is represented through the htrc:hasRepresentation predicate. Likewise, all *representations* of Bibliographic Resources gathered into Worksets MUST be related to a programmatic method (e.g., an API) from which an appropriate named representation can be retrieved and consumed by the analytics modules. This relationship is represented through the htrc:retrievableVia predicate. The essential data model for Bibliographic Resources is illustrated in Figure 12 (above).

In addition to the basic properties that facilitate the functionality of analytics modules, Bibliograpghic Resources require sufficient metadata, both to complete the account of the Workset's metadata (as illustrated in Figure 8 above and to aid scholars in selecting the most appropriate resources for their Worksets. There are many kinds of metadata that capture various aspects of the *items* gathered into Worksets. The essential data model for Bibliographic Resources will require a number of extensions in order to maximize its potential.

Vocabulary

Entity / Property	Type	Definition	Cardinality
dcterms:BibliographicResource[20]	Class	A book, article, or other documentary resource.	N/A
htrc:Volume	Class	A sub-class of dcterms:BibliographicResource, specifically an abstraction equivalent to a book or bound-format serial and comprising a group of page-level entities.	N/A
htrc:Page	Class	A sub-class of dcterms:BibliographicResource, specifically an abstraction equivalent to a single page-sized chunk of content, which may represent a page from a book, a letter, or content that fits on one side of a single leaf (e.g., of paper, papyrus, vellum, etc.). Sometimes called the logical page or leaf.	N/A
htrc:hasRepresentation	Relationship	The relationship between an abstract entity that constrains some block of content, e.g., a volume or a page, and a file that contains an inscription of that content (that may or may not be decipherable by a human).	1+
htrc:retrievableVia	Relationship	The relationship between a file that represents a Bibliographic Resource and a method for retrieving that file (e.g., an API).	1+

The simplest metadata is asserted through existing metadata records and describes those bibliographic resources that are analogous to books, what the HTRC calls *Volumes*. In addition to this kind of metadata, the use cases and technical requirements clearly illustrate that there is a need for metadata that describes finer grained bibliographic resources such as individual pages, paragraphs, sentences or arbitrary blocks of text. There is also a need to adequately capture provenance relationships between bibliographic entities, as well as more abstract relationships between textual content and the physical artifacts into which they are inscribed.

[20] http://dublincore.org/documents/domain-range/#sect-2

The following four sub-sections takes up each of these issues in turn and discusses one or more possible courses of action before suggesting one to be pursued as an extension to the simplistic workset model laid out above.

3.4.1. ORDINARY BIBLIOGRAPHIC METADATA

Ordinary bibliographic metadata is well represented through whole documents (i.e., records) conforming to the MARC standard. Unfortunately, records are not easily leveraged by the systems that data stores employ. Typically, records must be broken down and, in relational databases, their constituent information divided among multiple tables. RDF-based data stores have a small advantage in that they can better preserve the semantics of the individual assertions contained within a metadata record at the expense of much of the record's document structure.

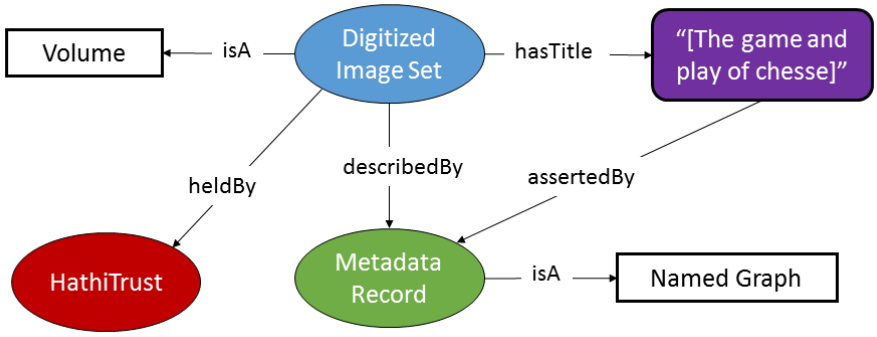

Figure 12: Metadata records and their associations with Volumes

Some of the structure can be preserved by employing named graphs as illustrated in Figure 12 (above). This technique has larger implications for serialization of information out of the data store[21] rather than for search and retrieval within it. Of bigger consideration is which of the many competing MARC-to-RDF standards to employ when producing the graphs.

With the advent of the linked data movement[22] several initiatives emerged that were either directly investigating how to best move existing MARC records from the xml document format to the RDF graph format or were developing vocabularies that could potentially be used for that purpose. Among these are: MODSRDF,[23] BIBFRAME,[24] Schema.org,[25] BIBO,[26] and FRBRoo. A recent paper by Nurmikko-Fuller et al. (2015) conducts a preliminary analysis comparing MODSRDF, BIBFRAME, Schema.org, and FRBRoo. The paper notes that there is a great deal of overlap between MODSRDF, BIBFRAME, and to a lesser extent Schema.org when compared to FRBRoo.

[21] Cf http://www.w3.org/TR/trig/ and http://json-ld.org/spec/ED/json-ld-syntax/20120522/#named-graphs for more details on how to serialize named graphs.
[22] http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/
[23] http://www.loc.gov/standards/mods/modsrdf/v1/
[24] http://bibframe.org/
[25] http://schema.org/
[26] http://bibliontology.com/

As noted above, FRBRoo is an extension of the CIDOC-CRM model and is much more focused on the capture, preservation, and representation of events within a bibliographic resource's lifecycle. Vis-à-vis description of bibliographic resources, it is only marginally helpful. It is much too focused on events to adequately preserve the kinds of descriptive information that appears in MARC. However, as there are also provenance concerns that must be addressed by the system's overarching data model, it will be referenced again in section 3.4.3 below.

BIBO, or more properly the Bibliographic Ontology, is a simplistic standard that leverages existing Dublin Core (DC) descriptive vocabulary while adding in the additional properties necessary to contextualize the what (e.g., the resource is a conference paper) and the where (e.g., presented at conference, reproduced in proceedings, etc.). BIBO is optimized to capture, preserve, and represent the kinds of metadata that are most exploitable for citation construction. MARC metadata is once again not a good fit, as the lossyness of moving from the MARC format into the DC vocabulary is well known (St. Pierre & LaPlant, 1998; NDMSO, 2008). However, once again there is a need elsewhere in the model for metadata that looks like this and use of BIBO in more granular contexts will be taken up below in the next section.

This leaves MODSRDF, BIBFRAME, and Schema.org, all of which have been engineered with either MARC in mind or the kinds of inventory systems that MARC is optimized for in mind. Of these three, MODSRDF at first, looks to be the optimal match. Designed from at the onset to move MARC metadata into XML, MODS[27] has been the go-to metadata schema at the Library of Congress for well over a decade. The problem with MODSRDF is that, with properties like "elementList" and "elementValue", it preserves too much of MODS XML document structure, packing it in alongside the metadata that actually describes the bibliographic resource.

BIBFRAME, another Library of Congress initiative, is to some extent an ongoing exercise in the reinvention of MARC. Designed from the ground up as a linked data vocabulary, BIBFRAME seems to be the next best option. Unfortunately, BIBFRAME's development appears to be diverging from other linked data and RDF-based vocabulary projects. One of its primary problems is verbosity. Within the BIBFRAME universe there is an individualized predicate for each and every standard identifier system in the bibliographic universe. For instance there are separate predicates for such standards as ISBN, ISSN, DOI, etc. Since there are always going to be new identifier schemes being invented, it seems doubtful that development of BIBFRAME will ever end – the numbers of predicates could very well balloon out forever. There are other BIBFRAME predicates that conform to this model of enumerating all possible permutations. Stability is going to be a very elusive state for the BIBFRAME ontology to achieve.

This leaves Schema.org as the clear choice for implementation within the workset's data architecture and that is the recommendation of this white paper. To be perfectly clear though, this is an imperfect solution to a complex issue. Schema.org is certainly not without its own set of issues. Transformation from MARC into Schema.org is still lossy. Schema.org is also more optimized for systems that are

[27] http://www.loc.gov/standards/mods/

designed to allow end users to select something for delivery. However, both OCLC[28] and the University Library at the University of Illinois at Urbana-Champaign (UIUC)[29] are actively converting some of their MARC records into this format. Additionally, neither MODSRDF nor BIBFRAME are stable specifications so adopting Scema.org seems likely to provide results that have the least risk for entropy.

The benefits of adopting Schema.org as the vocabulary for capturing, preserving, and representing Volume-level bibliographic metadata are twofold:

1. The prototyping development team can leverage existing Schema.org graphs from UIUC's library and OCLC, and
2. Since Schema.org is a more stable vocabulary, it is much less likely that the underlying data structures for Volume-level metadata are going to change.

This simplifies implementing metadata services for this level of data in the WCSA prototype, leaving additional time to further develop the functionalities described in the subsequent sections.

3.4.2. METADATA FOR BIBLIOGRAPHIC GRANULES

The issue of representing finer-grained bibliographic resources beyond whole volumes has come up several times through the course of this paper. This is directly related to the desire by digital humanists to be able to define and work with their own units of analysis rather than be limited to those that are artifacts of the digitization process or system architecture design. There are two kinds of metadata that support the representation and exploitation of bibliographic resources that are different from volumes. The first is identity metadata, without a means of specifically referring to these finer granules, it is impossible to gather them into worksets or make use of them in any meaningful way. The second is descriptive metadata, which allows scholars to make informed decisions regarding what to include as units for analysis.

3.4.2.1. IDENTIFIERS FOR BIBLIOGRAPHIC GRANULES

Identity metadata takes two forms, simple identifiers (e.g., IRIs, URIs, etc.) and contextual metadata. Simple identifiers can be used for those cases where the granules are already known due to specific facts of the digitization process and accompanying system architectures. Contextual metadata is required when the bibliographic resource is an arbitrary granule of a larger resource. Some examples of this latter use case range from the relatively self-contained, e.g., a poem on a page, to the highly random, e.g., a 500-word block of text that spans over parts of three different pages.

It has already been suggested that the HTRC take steps to implement the former (Jett et al., 2014). Since the entire HTRC infrastructure is built around artifacts from the HathiTrust Digital Library's corpus there is already an ample foundation to build upon. To the typical end user the digital artifacts appear to be whole volumes. From the point of view of the HTRC's existing system architecture, each volume is an abstraction that comprises a paired set of file objects. Each of the individual file objects in these sets contains the textual content of a single page from a volume.

[28] http://www.oclc.org/home.en.html
[29] http://www.library.illinois.edu/

The existing HTRC architectures already support exploiting page-level granules as distinct bibliographic resources. All the HTRC lacks is vocabulary sufficient for identifying pages as distinct entities. The file objects that represent pages are currently identified through a naming convention that combines the identifier for the volume they belong to with an integer representing their relative position within that volume. Unfortunately, this system has already proven to be imperfect, as the actual relative position of a page within a volume does not always correspond to the integer part of its identifier. To make page-level content addressable, persistent and unique identifiers must be minted for each page. Since the content of each page is represented by a pair of file objects in different formats there also needs to be an abstract entity that captures the page's content. Such an architecture appears in Figure 13.

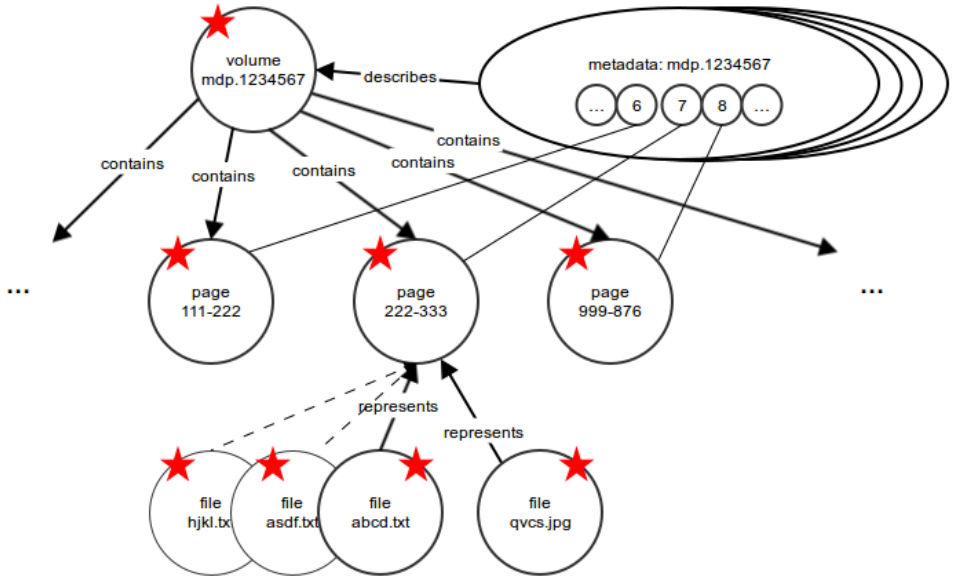

Figure 13: Page-level identifier architecture

Pages are the simple case for the HTRC because volumes come pre-chunked in page-sized granules as a result of the digitization process. To identify finer granules or arbitrary granules requires that pages be further sub-divided. Technically, this could be accomplished by chunking the text content of the existing file objects and storing it in smaller, more numerous file objects. Such an approach would allow the identifier solution suggested for pages to simply be extended to accommodate the smaller granules.

This is not a solution that will scale well. The approach is not efficient and will quickly fill up the system's storage space with large amounts of files that contain duplicate text. A better method would be to store the metadata that contextualizes the granule as a means for chunking the pages in a manner which can be exploited by various HTRC tools to produce arbitrary granules at the point in the system's internal workflows that they are needed. The metadata that gives sufficient contextual information for successful retrieval of the particular bibliographic granule is necessary. Fortunately, such means of articulating such metadata has already been invented.

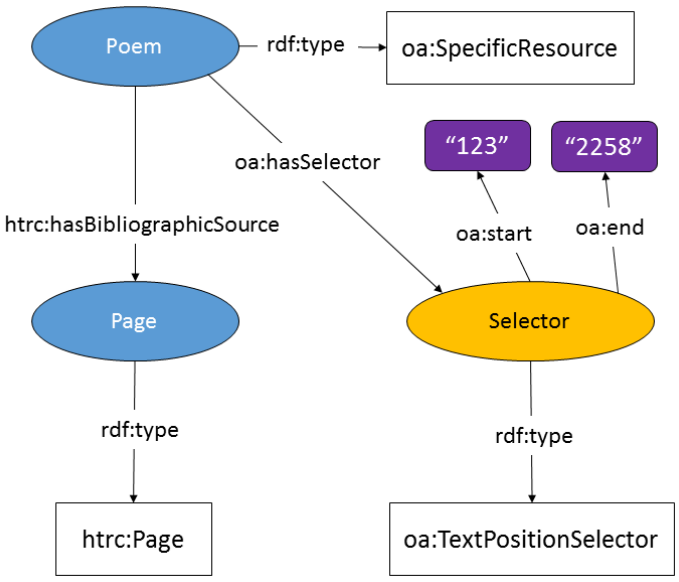

Figure 14: Using Specific Resources to pick out Bibliographic Granules

The Web Annotation Working Group[30] (WAWG) has been developing an interoperability standard for serializing annotations across the web. An outgrowth of the Open Annotation Community Group,[31] which was itself the result of a merger of the Open Annotation Collaboration[32] and the Annotation Ontology,[33] the WAWG has recently published their first public working draft for the Web Annotation specification. This is important because within the documentation of the vocabulary for their annotation model lies a construct – the specific resource[34] – which is perfect for the task needed to support minting identifiers for any arbitrarily sized bibliographic granule.

The specific resource (modeled in Figure 14 above) comes fully equipped with all of the entities and properties needed to describe specific portions of web resources. The proposal is to make a sub-class of the oa:hasSource predicate – called htrc:hasBibliographicSource – and then to use the remaining structures wholesale. In the example in the figure, the bibliographic granule that the workset gathers is a poem. Since it is only a portion of one page of one volume, it is given the type oa:SpecificResource. This is a clear indication to the system architecture that it should expect a source and a selector. In this case the resource that is the object of the htrc:hasBibliographicSource predicate is an htrc:Page, a suggested sub-type for htrc:BibliographicResource. Of equal importance is the object of the oa:hasSelector predicate, which in this case provides the character range that contains the text of the poem on the page.

[30] http://www.w3.org/annotation/
[31] https://www.w3.org/community/openannotation/
[32] http://www.openannotation.org/
[33] https://code.google.com/p/annotation-ontology/wiki/Homepage
[34] http://www.w3.org/TR/2014/WD-annotation-model-20141211/#specifiers-and-specific-resources

3.4.2.2. DESCRIPTIVE METADATA FOR BIBLIOGRAPHIC GRANULES

Once scholars are able to select bibliographic resources whose granularity is a match to their desired units of analysis new metadata, particular to those bibliographic granules, will be produced. Both the HTRC and scholars can realize additional benefits by consuming this metadata and constructing accounts of the individual granules. This will facilitate further use of the corpus by providing enough tools for scholars to search through the collection based on properties possessed by specific bibliographic granules. This will also allow the HTRC to build additional systems that leverage this finer grained data.

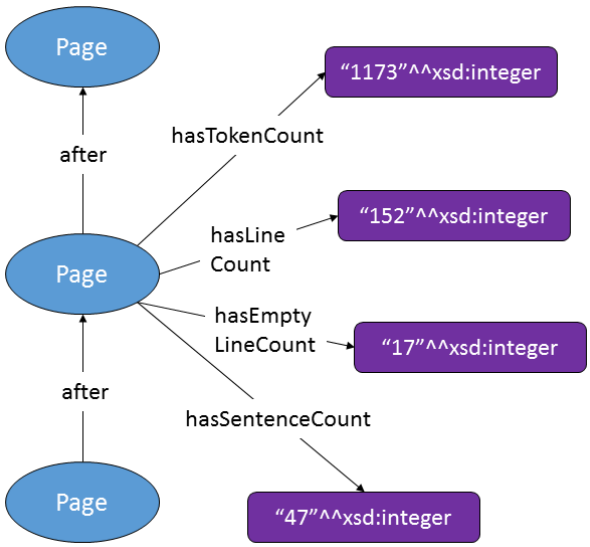

Figure 15: Possible Workset extension to assert Page-level metadata

The HTRC already possesses a vast wealth of such metadata in the form of page-level extracted features. A reliable means for leveraging this data has not been produced yet. There are a number of options that would need to be explored. Consideration for whether or not it makes sense to develop an extension to the workset vocabulary, as illustrated in Figure 15 (above), needs to be made. Alternatives, such as using existing standards like Schema.org need to be attempted. This is an ongoing area of research for linked data initiatives like this one, and a firm proposal, beyond confirming that bibliographic granules need metadata that specifically describes them in order to assure their optimal use by both scholars and the HTRC, is not forthcoming in this report.

Finer-grained resources also require descriptive metadata to facilitate their retrieval and use by the system architecture and other scholars. In the case of arbitrarily-sized granules, such as 500-word text blocks, a means of capturing and linking metadata, similar to those described for pages needs to be explored. In contrast, whole work granules, such as chapters, poems, and stories may be more easily linked to their metadata through existing standards. BIBO stands out as one such standard that might exploited for this purpose, as illustrated in Figure 16.

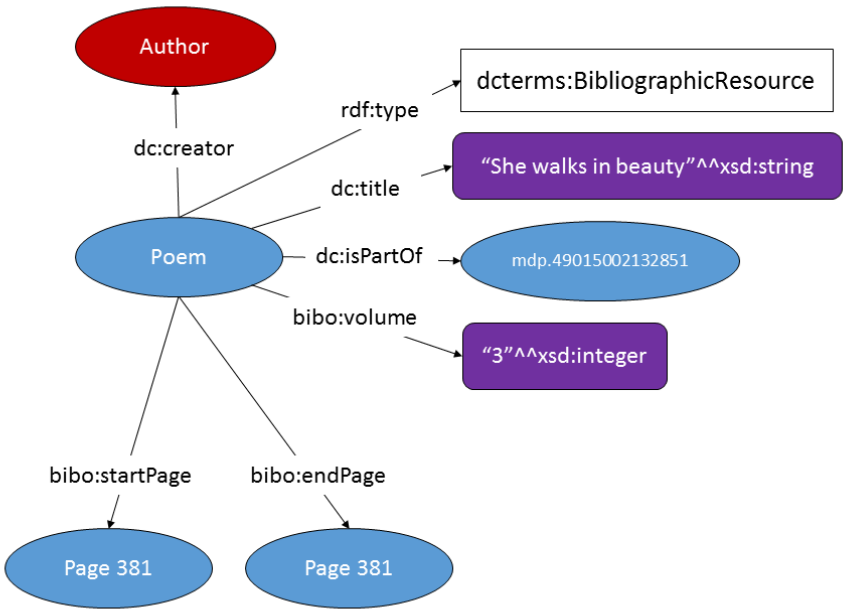

Figure 16: Describing a poem using BIBO

Again, beyond confirming that pages and other granular bibliographic resources require descriptive metadata that is specific to them, this report cannot definitely recommend the exact shape that such descriptive metadata take. Making extensions to the workset architecture laid out in the sections above provides one possible solution while using existing standards like BIBO provides another. Further development is needed to determine the best solution and to answer questions such as:

- How best can extracted features be asserted as metadata describing resources?
- Can extracted features be used to create graphs that conform to existing standards like BIBO or Schema.org?
- Is there a need for cataloging intervention to further facilitate the use of some data products?

It may be the case that ad hoc solutions are called for and different descriptive metadata standards will be needed for the optimal representation of different bibliographic resources. For instance, it has already been argued that the workset vocabulary requires an extension that will facilitate linking extracted features to the pages that they describe. This solution may also be appropriate for arbitrary bibliographic granules such as 500-word text blocks. Using a more formal, off-the-shelf vocabulary, like BIBO, may be more appropriate for granules that contain entire works, such as short stories, poems, etc. Further examination is necessary to make definitive determinations.

3.4.3. PROVENANCE FOR BIBLIOGRAPHIC RESOURCES

As discussed above, the recommendation is to employ PROV-O to ensure the fidelity of worksets as immutable, citable data products. The immutability and citability of the bibliographic resources gathered into them is the cornerstone of any infrastructure that would meet this need. As such, it is

the recommendation of this report that PROV-O be applied to manage the provenance of the various bibliographic resources in the HTRC milieu.

3.4.4. BIBLIOGRAPHIC RESOURCES AS ABSTRACTIONS

The topic of higher-level abstract bibliographic entities was tangentially broached in the section above. The fact is that many of the resources, from worksets to bibliographic granules, are abstractions. That they are abstractions isn't as important in and of itself as what their being abstractions buys the system architecture. In the case of pages, the abstraction gives the architecture a ready means to differentiate between an image file containing a page-sized chunk of text, a text file containing a page-sized chunk of text, and a page-sized chunk of content. This works because the content of the text in each of the two files is the same (neglecting the obvious problems that OCR quality causes). Volumes are likewise abstractions which contain a set of page-sized chunks of content.

FRBR (IFLA, 2009) is a conceptual framework developed by the library and information science professions during the 1990s. Within the FRBR milieu are higher-level abstractions – Expression and Work – that can be leveraged to find all of the different versions of certain narratives by particular authors. Many library catalog systems are being refined to make better use FRBR's entities in support of expanded query response services that can reconcile different descriptions of the same content.

One way for the architectural model described in this report could integrate the FRBR framework within its existing and recommended structures is to reconcile volume-level metadata descriptions with work-level descriptions that are being developed within OCLC.[35] This would serve two purposes:

1. It would provide for an entity that allows scholars to remark directly on an author's narrative content and,
2. It would provide for an entity around which multiple volumes containing the same narrative content can be grouped.

Unfortunately, there are some catches to this approach. One potential problem is that the FRBR framework actually obfuscates some of the kinds of textual and content features that are of interest to scholars for the sake of maintaining its Work-Expression-Manifestation-Item entity quartet. The other potential problem is one that I'm calling the *manifestation problem*.

3.4.4.1. MAPPING HTRC ENTITIES TO FRBR

With regards to the first problem, there are several stumbling blocks to overcome. One of the primary issues is that contemporary descriptive metadata packs in work-level, expression-level, manifestation-level, and (some) item-level metadata into a single undifferentiated set of assertions.[36] It is completely ambiguous to the computer and somewhat ambiguous to the end user, just which one of the four entities each metadata assertion describes.

---

[35] http://www.oclc.org/developer/develop/linked-data/worldcat-entities/worldcat-work-entity.en.html

[36] Aspects of this issue have been brought up before in discussions of the Dublin Core 1:1 principle. See Urban, R. J. (2014) for a thorough discussion of the 1:1 issue.

The overall existing and suggested architectural models are also missing entities at the expression and (probably) item-levels. One would be tempted to argue that the volumes are the items but, in the HTRC context, volumes are just ordered sets of pages and sets of any kind are abstract entities. Under the FRBR framework, Items must be concrete things constructed of patterned matter and energy; they cannot be abstractions, and so a volume in the HTRC context is not analogous to a FRBR Item. The files containing the page-size chunks of text are much more analogous to FRBR Items.

3.4.4.2. THE MANIFESTATION PROBLEM

The manifestation problem is potentially even more troubling. Nothing is damaged by not having every level of FRBR entity explicitly represented. Taking the trouble to add them in on the other hand may result in deleterious effects. To some extent it depends upon the desired amount of fidelity that needs to be accommodated. At full fidelity, capturing all of the relevant entities proves to be a very daunting task. This is because a "new" FRBR Manifestation becomes evident each time the text of a known Manifestation is copied into a new medium. This is especially pertinent to and especially onerous for digital libraries.

The implication is that each time a file object is moved, copied, or used by an agent, a previously unknown Manifestation is discovered.[37] When a file is accessed, a never before seen Manifestation makes itself known within the computer's processing system and when it is rendered to an agent through some output process, yet another Manifestation is discovered.[38] If a digital library were to capture and record just the provenance metadata regarding each of these events, its database's contents would quickly be overwhelmed.

For practical reasons, no digital library would capture a record of every FRBR Manifestation that occurs. All the more reason not to try to extend the underlying architectural model to accommodate every last FRBR entity that plays a role within a data store. For the model proposed by this report, an extension accommodating work level entities will be harmless in the overall scheme of things. A full round of testing through implementation in a prototyping environment will be needed to see what kinds of additional functionality scholars will be able to realize through the additional support of FRBR entities.

3.4.4.3. ALTERNATIVES TO FRBR

An alternative approach that might provide similar functionality, while increasing opportunities for characterizing the nature of the abstract content types being studied, is the Basic Representation Model (BRM), illustrated in Figure 17 below (Wickett et al., 2012b). Taking aspects of the Preservation Model, version 1.0 ([Dubin], 2010) and SAM, BRM makes three simple delineations:

1. There is propositional content (i.e., the content represented by some text, music, images, etc.),

[37] Here I'm taking the Platonist's position that abstract entities are neither created nor destroyed, just discovered or forgotten by particular individuals.

[38] Those experienced in computer system design might observe that new Manifestations are discovered (and new Items created) each and every time a file jumps to a new bus and is encoded or decoded by one microchip or another. Each time a person accesses a file, dozens, if not hundreds, of previously unknown Manifestations would be discovered (and subsequently forgotten when the Items serving as evidence for them are destroyed).

2. There are physical artifacts made from matter and/or energy (i.e., text is inscribed in some medium, whether that be photons being projected from a monitor or ink scratched into some paper), and
3. There are multiple intervening symbol sets which encode the content in manners that make it suitable for transmission (e.g., an author decides to express their story using English text, in the form of a novel, etc.).

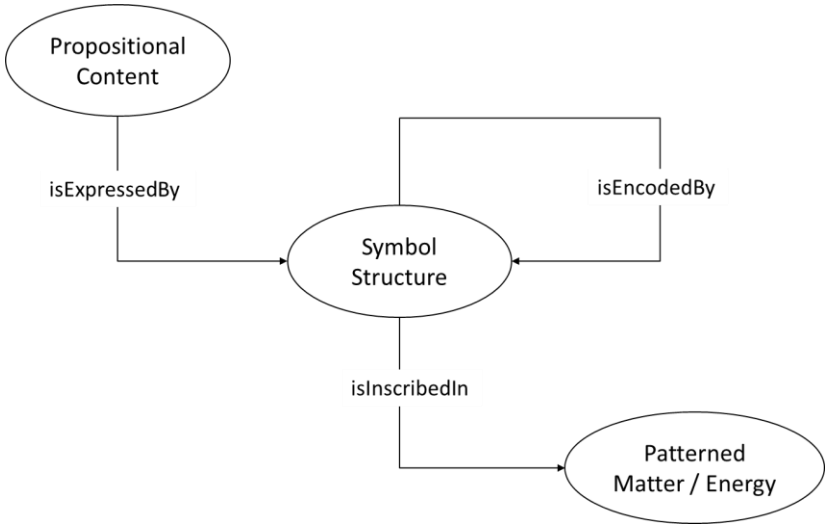

Figure 17: The Basic Representation Model

The primary benefit of the BRM is that we can abstract away the intervening symbol sets at arbitrary levels of granularity. It allows us to reference a much higher abstraction than FRBR's Work in the form of Propositional Content while preserving the FRBR notion of Item in the form Patterned Matter and Energy. The various Symbol Sets can be compacted into a single "Publication" entity or expanded into as many entities as individual scholars think is helpful. Such expansions can be stored in the form of named graphs in a separate data store dedicated to their maintenance and use in support of analytics algorithms designed to exploit them and citations that reference them by kind. Such graphs ultimately serve as evidence remarking with as great or little specificity as to *what* was analyzed as the scholar is willing to assert.

3.4.4.4. OTHER ABSTRACT ENTITIES OF INTEREST

As noted above, under ideal conditions, the formalization of Worksets presented here accommodates the gathering together of any kind of resource, not just those we understand as bibliographic resources. In the long run, to best facilitate scholars' abilities to focus on entities of interest in accordance to their research goals, the implementation resulting from the data model presented here will need to continue to expand and grow beyond even what has already been suggested.

Infrastructure that enables scholars to gather together, among other things, arbitrary named entities and concepts as their units of analysis, needs to be explored. One imagines that such abstract entities would be highly appropriate to a number of revealing analytics processes, such as network analytics. Careful thought needs to be given by the engineers and architects enlisted in building the HTRC's

next generation of technical infrastructures to avoid assumptions that will artificially constrain and dampen efforts to extend the model to accommodate such features.

4. CONCLUSION

This report has set out to describe a set of descriptive and technical requirements derived from documented use cases. It used them to develop a basic conceptual model that describes and makes machine-actionable, scholar-built, digital worksets. Each workset aggregates a selection of bibliographic resources which can be programmatically chosen or hand curated by individual scholars according to their specific research needs.

Table 5: List of Recommendations for Realizing and Extending the Workset Data Model

Recommendation
Implement the basic Workset and Bibliographic Resource models described in Section 3 through new Workset Builder infrastructure.
Develop workflows to leverage existing HTRC MARC metadata for Volumes to better empower scholars to select resources for their Worksets.
Implement identity metadata for bibliographic granules (Page-level relatively easy to implement, finer and more arbitrary granules will require additional development cycles).
Develop and implement descriptive metadata for bibliographic granules (Page-level relatively challenging. (How best to leverage extracted features remains something of an open question.) Other granule levels will require additional development cycles.
Develop and implement provenance metadata at all levels using PROV-O and PROV-DM. (Unless a provenance method that relies solely on infrastructure is instead indentified.)
Develop and implement means of differentiating abstract levels of content from one another. (Relatively moderate at the Page-level. Complicated by indirection and notions like “proxies” which lead to misuse of metadata records acting in the role of avatars representing other entities.)

To fully realize the resulting workset and bibliographic resource data models and better meet the needs that scholars have articulated, a series of recommendations for action have been detailed (Table 3 above). Several of the suggested innovations extending the basic workset and bibliographic resource models are in the process of being actively developed within the context of a prototype triple store that has been established for the experimentally-based development of this model. Others will need additional refinement before they are ready for such provisional deployment. The ultimate goal of this architectural model is to build an articulated data model that affords both scholars and the HTRC a broad range of functionality, from volume deduplication and disambiguation to providing sophisticated metadata that affords opportunities for analysis of finely grained bibliographic resources.

REFERENCES

Companion to Digital Humanities. (2004). Schriebman, S., Siemens, R. & Unsworth, J., eds. Oxford: Blackwell.

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

Downie, J. S. & Aiden, E. L. (2014). Exploring the billions and billions of words in the HathiTrust corpus with Bookworm: HathiTrust + Bookworm. NEH implementation grant. September 2014 – August 2016.

[Dubin, D.] (2010) Preservation model, version 1.0. In Unsworth, J. & Sandore, B. [(eds.)] ECHO DEPository – Phase 2: 2008-2010: final report of project activities. Report for the National Digital Information Infrastructure & Preservation Program. Champaign, IL: University of Illinois at Urbana-Champaign.

Henry, C. & Smith, K. (2010). Ghostlier demarcations: large-scale text digitization projects and their utility for contemporary humanities scholarship. In The idea of order : transforming research collections for 21st century scholarship (pp. 106–115). Council on Library and Information Resources.

IFLA Study Group on FRBR. (2009). Functional requirements for bibliographic records: Final report [revised]. München: K.G. Saur Verlag.

Gooding, P., Terras, M. & Warwick, C. (2013). The myth of the new: Mass digitization, distant reading, and the future of the book. *Literary and Linguistic Computing 28*(4), 629 - 639.

Fenlon, K., Senseney, M., Green, H., Bhattacharyya, S., Willis, C. & Downie, J. S. (2014). Scholar-built collections: A study of user requirements for research in large-scale digital libraries. Paper presented at The 77th ASIS&T Annual Meeting. (Seattle, WA, 31 October – 5 November, 2014).

Hill, L., Janee, G., Dolin, R., Frew, J. & Larsgaard, M. (1999). Collection metadata solutions for digital library applications. *Journal of the American Society for Information Science 50*(13), pp 1169-1181.

Jett, J., Ruan, G., Unnikishnan, L., Fallaw, C., Maden, C. & Cole, T. (2014). Proposal for persistent & unique identifiers. Technical report to the HathiTrust Research Center Executive Committee. Champaign, IL: University of Illinois at Urbana-Champaign.

Lynch, C. (2002). Digital collections, digital libraries, and the digitization of cultural heritage information. *First Monday*, *7*(5).

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. & Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science 331*(6014), pp 176–182.

(NDSMO) Network Development & MARC Standards Office, Library of Congress. (2008). Dublin Core to MARC crosswalk. Technical documentation. Washington, D.C.: Library of Congress. Accessed on 24 February 2015 from: http://www.loc.gov/marc/dccross.html

Nurmikko-Fuller, T., Page, K., Willcox, P., Jett, J., Maden, C., Cole, T., Fallaw, C., Senseney, M. & Downie, J. S. (2015). Building complex research collections in digital libraries: A survey of ontology implications. Short paper submitted to the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, 2015 (Knoxville, TN, 21-25 July 2015).

Organisciak, P., Auvil, L., Bhattacharyya, S. & Downie, J. S. (2015). The HTRC extracted features dataset. Paper accepted for the Joint Canadian Society for the Digital Humanities and Association for Computing in the Humanities Conference. Ottawa, Canada. Forthcoming.

Palmer, C. L. (2004). Thematic research collections. In Schreibman, S., Siemens, R., and Unsworth, J. (Eds.) A Companion to Digital Humanities. Blackwell Publishing, Oxford.

Palmer, C. L. & Knutson, E. (2004). Metadata practices and implications for federated collections. Proceedings of the 67th ASIS&T Annual Meeting (Providence, RI, 12-17 November 2004).

Palmer, C. L., Knutson, E., Twidale, M. & Zavalina, O. (2006). Collection definition in federated digital resource development. Proceedings of the 69th ASIS&T Annual Meeting (Austin, TX, 3-8 November 2006).

Palmer, C. L., Zavalina, O. & Fenlon, K. (2010). Beyond size and search: Building contextual mass in aggregations for scholarly use. Proceedings of the 73rd ASIS&T Annual Meeting (Pittsburgh, PA, 22-27 October 2010).

Palmer, C. L. & Jett, J. (2013). Next generation digital federations: Adding value through collection evaluation, metadata relations, and strategic scaling. Final report submitted to IMLS (LG-06-07-0020). Champaign, IL: University of Illinois at Urbana-Champaign.

Palmer, C. L., Isaac, A., Wickett, K. M., Fenlon, K. & Senseney, M. (2015). Digital collection contexts: iConference 2014 workshop report. CIRSS technical report 20150301. Champaign, IL: Center for Informatics Research in Science and Scholarship. Accessed on 5 May 2015 from: http://hdl.handle.net/2142/73359

Pearsall, E. (2012). Twentieth-century music theory and practice. London: Routledge.

Renear, A. H., Wickett, K. M., Urban, R. J. & Dubin, D. (2008a). The return of the trivial: Formalizing collection/item metadata relationships. Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries 2008 (Pittsburgh, PA, 16-20 June 2008).

Renear, A. H., Wickett, K. M., Urban, R. J., Dubin, D. & Shreeves, S. (2008b). Collection/Item metadata relationships. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2008 (Berlin, Germany, 22-26 September 2008).

St. Pierre, M. & LaPlant, W. P. (1998). Issues in crosswalking content metadata standards. Technical report prepared for NISO. Baltimore, MD: National Information Standards Organization.

Underwood, T. (2012). Topic modeling made just simple enough. The Stone and the Shell [blog], 7 April 2012. Accessed on 2 March 2015 from: http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Urban, R. J. (2014). The 1:1 Principle in the Age of Linked Data. Paper presented at the International Conference on Dublin Core and Metadata Applications DC-2014, Austin Texas, U.S.A. Accessed on 7 May 2015 from: http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/263

Varvel, V. E. J. & Thomer, A. (2011). Google digital humanities awards recipient interviews report (CIRSS Report No. HTRC1101). Technical report prepared for the HathiTrust Digital Library. Champaign, IL: Center for Informatics Research in Science and Scholarship.

Wickett, K. M., Renear, A. H. & Urban, R. J. (2010). Rule categories for collection/item metadata relationships. Proceedings of the 73rd ASIS&T Annual Meeting (Pittsburgh, PA, 22-27 October 2010).

Wickett, K. M., Renear, A. H., & Furner, J. (2011). Are collections sets? Proceedings of the 74th ASIS&T Annual Meeting (New Orleans, LA, 9-13 October 2011).

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

Wickett, K. M., Thomer, A., Sacchi, S., Baker, K. S. & Dubin, D. (2012a). What dataset descriptions actually describe: Using the systematic assertion model to connect theory and practice. Poster presented at the 2012 ASIS&T Research Data Access & Preservation Summit. Baltimore, MD.

Wickett, K. M., Sacchi, S., Dubin, D. & Renear A. H. (2012b). Identifying content and levels of representation in scientific data. Proceedings of the 75h ASIS&T Annual Meeting (Baltimore, MD, 26-30 October 2012).

Wickett, K. M., Isaac, A., Fenlon, K., Doerr, M., Meghini, C., Palmer, C. L. & Jett, J. (2013a). Modeling cultural collections for digital aggregation and exchange environments. CIRSS Technical Report. Champaign, IL: University of Illinois at Urbana-Champaign.

Wickett, K. M., Dubin, D., Senseney, M. & Almas, B. (2013b). Extending the systematic assertion model for humanities research. Poster presented at the 2013 ASIS&T Research Data Access & Preservation Summit. New Orleans, LA.

Wickett, K. M., Isaac, A., Doerr, M., Fenlon, K., Meghini, C. & Palmer, C. L. (2014). Representing cultural collections in the digital aggregation and exchange environments. D-Lib Magazine 20(5/6). Accessed 24 February 2015 from: http://www.dlib.org/dlib/may14/wickett/05wickett.html

Zavalina, O. L., Palmer, C. L., Jackson, A. S. & Han, M. J. (2008). Assessing descriptive substance in free-text collection-level metadata. Proceedings of the 8th International Conference on Dublin Core and Metadata Applications (Berlin, Germany, 22-25 September 2008).

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

APPENDIX A: HTRC WORKSET XSD

```
<?xml version="1.0" encoding="UTF-8"?>
<schema xmlns="http://www.w3.org/2001/XMLSchema"
        targetNamespace="http://registry.htrc.i3.illinois.edu/entities/workset"
        xmlns:tns="http://registry.htrc.i3.illinois.edu/entities/workset"
        elementFormDefault="qualified">

  <include schemaLocation="comment.xsd" />
  <include schemaLocation="tag.xsd" />
  <include schemaLocation="volume.xsd" />

        <complexType name="WorksetMeta">
            <sequence>
                <element name="version" type="long" minOccurs="0" />
                <element name="name" type="string" />
                <element name="description" type="string" />
                <element name="author" type="string" minOccurs="0" />
                <element name="rating" minOccurs="0">
      <simpleType>
        <restriction base="nonNegativeInteger">
          <maxInclusive value="5" />
        </restriction>
      </simpleType>
                </element>
                <element name="avgRating" type="float" minOccurs="0" />
                <element name="lastModified" type="dateTime" minOccurs="0" />
                <element name="lastModifiedBy" type="string" minOccurs="0" />
                <element ref="tns:tags" minOccurs="0" />
                <element ref="tns:comments" minOccurs="0" />
                <element name="volumeCount" type="int" minOccurs="0" />
                <element name="public" type="boolean" minOccurs="0" />
            </sequence>
        </complexType>

        <complexType name="WorksetContent">
            <sequence>
                <element ref="tns:volumes" />
            </sequence>
        </complexType>

        <complexType name="Workset">
            <sequence>
                <element name="metadata" type="tns:WorksetMeta" />
                <element name="content" type="tns:WorksetContent" minOccurs="0" />
            </sequence>
        </complexType>

        <complexType name="Worksets">
    <sequence>
      <element ref="tns:workset" minOccurs="0" maxOccurs="unbounded" />
    </sequence>
        </complexType>

        <element name="workset" type="tns:Workset">
            <unique name="TagUnique">
                <selector xpath="tns:metadata/tns:tags/tns:tag" />
                <field xpath="." />
            </unique>
        </element>

        <element name="worksets" type="tns:Worksets" />

</schema>
```

APPENDIX B: HTRC COMMENT XSD

```
<?xml version="1.0" encoding="UTF-8"?>
<schema xmlns="http://www.w3.org/2001/XMLSchema"
    targetNamespace="http://registry.htrc.i3.illinois.edu/entities/workset"
    xmlns:tns="http://registry.htrc.i3.illinois.edu/entities/workset"
    elementFormDefault="qualified">

    <complexType name="Comment">
        <sequence>
            <element name="author" type="string" />
            <element name="text" type="string" />
            <element name="created" type="dateTime" minOccurs="0" />
            <element name="lastModified" type="dateTime" minOccurs="0" />
        </sequence>
    </complexType>

    <complexType name="Comments">
        <sequence>
            <element name="comment" type="tns:Comment" maxOccurs="unbounded" />
        </sequence>
    </complexType>

    <element name="comments" type="tns:Comments" />

</schema>
```

APPENDIX C: HTRC TAG XSD

```
<?xml version="1.0" encoding="UTF-8"?>
<schema xmlns="http://www.w3.org/2001/XMLSchema"
    targetNamespace="http://registry.htrc.i3.illinois.edu/entities/workset"
    xmlns:tns="http://registry.htrc.i3.illinois.edu/entities/workset"
    elementFormDefault="qualified">

    <complexType name="Tags">
        <sequence>
            <element name="tag" type="string" maxOccurs="unbounded" />
        </sequence>
    </complexType>

    <element name="tags" type="tns:Tags" />

</schema>
```

APPENDIX D: HTRC VOLUME XSD

```
<?xml version="1.0" encoding="UTF-8"?>
<schema xmlns="http://www.w3.org/2001/XMLSchema"
   targetNamespace="http://registry.htrc.i3.illinois.edu/entities/workset"
   xmlns:tns="http://registry.htrc.i3.illinois.edu/entities/workset"
   elementFormDefault="qualified">

   <complexType name="Property">
                        <attribute name="name" type="string" />
                        <attribute name="value" type="string" />
   </complexType>

   <complexType name="Properties">
      <sequence>
         <element name="property" type="tns:Property" maxOccurs="unbounded" />
      </sequence>
   </complexType>

   <complexType name="Volume">
      <sequence>
         <element name="id" type="string" />
         <element name="properties" type="tns:Properties" minOccurs="0" />
      </sequence>
   </complexType>

   <complexType name="Volumes">
      <sequence>
         <element name="volume" type="tns:Volume" maxOccurs="unbounded" />
      </sequence>
   </complexType>

   <element name="volumes" type="tns:Volumes" />

</schema>
```