Journal of Open Humanities Data

Jett, J et al 2016 The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. *Journal of Open Humanities Data,* 2: e1, DOI: http://dx.doi.org/10.5334/johd.3

ISSUES IN HUMANITIES RESEARCH DATA

The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections

Jacob Jett[1], Timothy W. Cole[1], Chistopher Maden[1] and J. Stephen Downie[1]

[1] Center for Informatics Research in Science & Scholarship, Graduate School of Library & Information Sciences, University of Illinois at Urbana-Champaign, US
jjett2@illinois.edu

Corresponding author: Jacob Jett

The HathiTrust Digital Library (HTDL) is a digital library containing about 14 million volumes which comprise billions of pages of content. The HathiTrust Research Center (HTRC) is a collaborative research initiative jointly led by Indiana University and the University of Illinois at Urbana-Champaign. This paper describes the development of a collections data model by the Workset Creation for Scholarly Analysis project, a HTRC research initiative funded by the Andrew W. Mellon Foundation. The resulting HTRC Workset data model is designed to aid humanities scholars by helping them to describe selected portions of the HTDL corpus that serve as the objects of their research. The resulting worksets are persistent, citable, and can be assessed by other scholars for reuse in additional research processes.

Keywords: research collections; formalisms; data models; digital libraries; large-scale corpuses

1 Introduction & Context

The HathiTrust Digital Library (HTDL) is a digital library containing 13.95 million volumes, comprising several billion pages of digitized text. The HathiTrust Research Center (HTRC) is a collaborative research initiative jointly based at Indiana University and the University of Illinois at Urbana-Champaign that provides support to researchers and humanities scholars who wish to exploit the HTDL's vast treasure trove of data. The Workset Creation for Scholarly Analysis (WCSA) project was an 18-month research initiative funded by the Andrew W. Mellon Foundation and was engaged in exploring methods for enriching the metadata that describes the HTDL's corpus, augmenting traditional string-based metadata with linked data, and formalizing the notion of *collections*, *research collections,* and *worksets* in the HTRC context.

The ideas of *research collection* and *workset* are central to the expectations for the kind of scholarly workflow (illustrated in **Figure 1** below) that the HTRC expects its digital humanist users to employ when they aggregate and ingest their specified research materials into particular analytics pipelines. As these notions are the cornerstone upon which any resulting ontology can be built, the WCSA project first had to understand them. This article reports on the outcomes of efforts to develop formal definitions of collections, research collections, and worksets in first order logic and a basic ontology capable of capturing various metadata that describe them.

A key outcome of the WCSA project is the HTRC's Non-Consumptive Workset ontology that is designed to work within the context of the HTRC's scholarly workflow. Worksets serve as a major input source for those tools and allow humanities scholars to gather together their research materials in a manner that is reminiscent to their traditional method of developing research collections. As **Figure 1** illustrates no philosophical or institutional limitation is put upon the scholars in such a way as to prescribe from which corpuses they may gather their research materials. They may gather materials from the HTDL's corpus or from other corpuses external to the HTDL. They may mix materials together. They may even gather the analytical results from previous work into new worksets for the purposes of additional analyses. In order to achieve this result the ontology remains agnostic with respect to *what* can be gathered into a non-consumptive workset.

The first part of this article lays out the theoretical firmament upon which the ontology rests. A series of formal definitions begins with collections, in the most general sense, and narrows to worksets, in the most specific sense. We next discuss a working ontology derived from the formal definitions that fully develops a number of properties vital for distinguishing the various kinds of collections from one another. We close with a discussion of additional work to refine and extend the workset ontology. Among this work is the need for additional extensions and refinements that enable the ontology to better target more granular intellectual objects that scholars may be

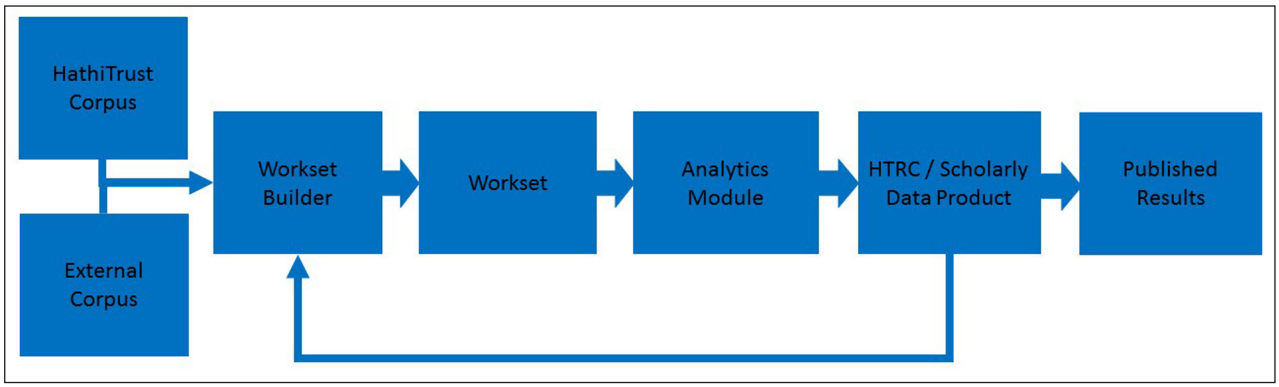

Figure 1: HTRC's Scholarly Workflow.

interested in or that are more appropriate for their desired analyses (e.g., worksets of pages, poems, images, or other bibliographic granules rather than ones composed entirely of whole volumes).

2 Formal Definitions for Collections

The notion of *collection* and the determination of what the label "collection" actually describes continues to be challenging for digital libraries (see [3, 6, 9, 13, 17]; among others). That scholars actively undertake collection development activities as a normal method for furthering their research seems incontrovertible [3, 8, 10, 11, 12]. While the preceding authors all speak in terms of *research collections*, it is also clear from their accounts that a great deal of curatorial effort is produced by scholars as they assemble their *research collections*. Since archival, library, museum collections as wholes are also the products of careful curatorial processes, it made sense to us to distinguish between *curated collections* and *research collections*. We proceed by first defining *collections* in general and then producing narrower definitions for *curated collections* and *research collections* before finally arriving at a definition for *worksets*. The following formal definitions take it for granted that *worksets*, as a central feature of the HTRC's scholarly workflow, are a kind of *research collection*.[1]

Previous work describing the kinds of relationships that obtain between *collections* and the things gathered into them relied on the binary predicate *isGatheredInto*(*x*,*y*) as the key property that supplies a *collection's* identity conditions [14, 15]. It is described at length in Wickett et al. [16] and axiom A1 below is derived directly from that work. This binary predicate comes directly from the Dublin Core Metadata Initiative's definition for collections,[2] which we have adapted as D1.

> D1: If and only 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:

$$\text{A1: } \forall y\,(\exists x\ isGatheredInto(x, y) \leftrightarrow Collection(y))$$

This formalization seems to satisfy the general requirements of the definition set forth in D1; however, in the HTRC's context a more specific definition is needed as the scope of their use case context is that of *worksets*, which are a kind of *research collection*. In order to narrow the formalization's scope, we first postulated what properties would be necessary to distinguish *research collections* in particular from other kinds of *collections*.

While there are many relevant themes that interweave throughout the various accounts, two particular themes emerge from among them [3, 8, 10, 11, 12].

- Research collections are the products of curatorial effort, i.e., they are created by an entity through some means of selection.
- Research collections serve a specific role within a scholarly research workflow, i.e., they gathered together in accord to some motivating purpose.

In order for something to be a *research collection* both of these things need to be true. However, if we consider archival, library, and museum collections, it is the case that only the first theme need be true in their case. This was a clear indication to us that *research collections* are, in fact, a kind of *curated collection*. It was necessary for us to then distinguish *curated collections* in general from *research collections* specifically before we could proceed to develop formal definitions for *research collections* and *worksets*.

That the things gathered into a *curated collection* are purposefully selected according to some criteria that are defined by some agent (typically the collection's curator), seems to be central to the concept of a *curated collection*. We set about formalizing this notion of *curated collections* by first developing the following definition, D2:

> D2: 'If and only 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.'

Since this definition, which we take to be the most accurate one for *curated collections*, invokes sets and undefined functions of the things being gathered, it is rather difficult to produce an axiom using just first-order logic.

To properly represent its true nuances requires the expressiveness of second-order logic which is a heavier weight solution than we hoped for. Rather than wrestle with these finer distinctions here, we instead chose to reformulate D2 in a manner that was better expressed using just first-order logic but admit it is also more of a gloss than an explication of the true nature of *curated collections*. We defer a fuller exploration of the notion of *curated collection* to a future date.

The reformed definition (D2') provides enough of a workaround to let us continue our use of first-order logic.

> D2': 'If and only if something, *x*, meets some criterion, *c*,[3] 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:

> A2': $\forall y\ (\exists x \exists c \exists w\ (isGatheredInto(x,y) \wedge meets Criterion(x,c) \wedge definedby(c,w)) \leftrightarrow Curated Collection(y))$[4]

Having for the time being adequately accounted for the process of selection, we were finally in a position to formally consider *research collections*. The primary distinction that we needed to showcase is that while every *research collection* is a *curated collection*, not every *curated collection* is fit for research (i.e., not every *curated collection* was curated with the express purpose of answering one or more research questions).[5] With this understanding we developed the following definition, D3.

> D3: 'If and only 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.'

Which we represent through axiom A3.

> A3: $\forall y((CuratedCollection(y) \wedge \exists z\ hasResearch Motivation(y,z)) \leftrightarrow ResearchCollection\ (y))$

Having finally arrived at a proposed definition for *research collections*, we next turned to considering how *worksets* (in this case specifically *non-consumptive worksets*) differed from other kinds of *research collections*. The primary distinction was provided by the HTRC and its *non-consumptive research paradigm*.

The *non-consumptive research paradigm* is a notion that emerged from the rejected Google Books settlement in the case of *Authors Guild et al. v. Google Inc.* [9]. The idea around non-consumptive research is that humanities scholars and researchers can employ computational science techniques to large corpuses of text that fall within the auspices of copyright. The settlement[6] vaguely defines "non-consumptive research" as any research that involves computational analysis on books wherein the researchers do not, themselves, have direct access to the text of those books such that they might read them or reproduce large portions of text from them. The entire *workset* notion has been developed by the HTRC as a means to refer to *research collections* that are expected to operate within the confines of a non-consumptive computational environment.

This understanding allowed us to produce a formal definition (D4) for *worksets* specifically.

> D4: 'If and only 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.'

Which in turn can be represented by the following axiom:

> A4: $\forall y\ ((ResearchCollection(y) \wedge intended ForUse(y,a)) \leftrightarrow Workset(y))$,
> 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.

Taken collectively, each of the four definitions provide important requirements for the kinds of properties and attributes for which any resulting ontology must provide. The core ontology set forth in the next section provides the essential framework from which a data model and serializations may be derived.

3 HTRC Workset Ontology

Adequately scoping the definitions for predicate and entity types is the most important factor in the development of a working ontology. The following paragraphs discuss the scoping of the *isGatheredInto* predicate in particular. They are followed by subsections describing the core entity and predicate arrived at through the formalization exercises. This discussion is followed by a discussion of additional entities and predicates derived from descriptive requirements suggested by the formalization exercise above.

The ultimate work that any ontology must accomplish is the provision for sufficient vocabulary access points that both a coherent descriptive account of a *workset* and its contents can be provided to end users on demand and so that various information retrieval operations (such as faceted retrieval and sorting, among others) can be accomplished. Since at its heart, every *workset* is a *collection*, the predicate *isGatheredInto(x,y)* is possibly the most important property described by the ontology.

Having determined this, the next question to answer was whether or not we should align our ontology with other ongoing collections modeling efforts. The notion of the predicate, *isGatheredInto*, originated with the DCMI and its Collections Application Profile (CAP). Rather than produce a new predicate which seemed like a specialization of one of their existing predicates, the DCMI-CAP chose to reuse the much broader scoped *dc:isPartOf* predicate in

their collection model. The predicate *dc:isPartOf* is a very general purpose predicate that describes a merelogical relationship that obtains between two resources.

Because the HTRC context is much more constrained (partially due to the sensitivity of the data being gathered into *worksets*) and because we wanted to highlight the specialized semantics of the *isGatheredInto* relationships, a predicate with a much narrower scope than *dc:isPartOf* was preferable. We next turned to the much more recent work of researchers working in collaboration with developers at Europeana to extend the Europeana Data Model (EDM) with collection entities. The proposed addition of a new property *edm:isGatheredInto* has been made [17] which matches both the theoretical scope described in A1 and the functional needs of the HTRC's nascent infrastructure. Because the work at the EDM is based on the existing work on *collections* and because convergence of data models expected to serialize metadata in Web settings is highly desirable from interoperability and efficiency perspectives we made a conscious decision to align our efforts with theirs (inasmuch as the differing scope of our use cases made it possible) and adopted the *edm:isGatheredInto* predicate. The HTRC Workset Ontology uses this relationship as the focus point from which all of its other properties are designed to support.

The key words "MUST," "MUST NOT," "REQUIRED," "SHALL," "SHALL NOT," "SHOULD," "SHOULD NOT," "RECOMMENDED," "MAY," and "OPTIONAL" in this section are to be interpreted as described in [2].

3.1 Core Ontology

The cornerstone of the linked data approach to metadata is the ontology. Each ontology provides both machines and humans with a relatively unambiguous vocabulary of terms that are employed for the sharing of data and around which various information retrieval system features can be designed. As **Figure 1** illustrates, moving structured information from module to module is an important aspect of the HTRC's evolving architecture making a linked data approach valuable. The core of the HTRC Workset ontology is the *htrc:Workset* entity class. We describe it as a sub-type of the older *dcmi:Collection* type established as part of the DCMI-CAP vocabulary. The core property is *edm:isGatheredInto* (and is derived directly from D1). Our expectation is that implementers will actually use its reciprocal, *edm:gathers*.

There are two reasons for this. Most compellingly, in graph-based models such as the HTRC Workset Ontology, the directionality of predicates is extremely important for indicating when a property is contingent and when it is not. For example, that a *collection gathers [a] thing* is a fundamental property of *collections*. The converse, that a *thing isGatheredInto [a] collection* is merely a contingent property of that *thing*. This is an important distinction to make as it is a clear indication that collection membership is a role from the context of a collection's items but the presence of those items is an existential identity condition for that collection from the collection's context. The other, less compelling reason is that actual implementers are going to prefer abbreviated predicates as it makes optimization work easier.

Under the purview of the definitions developed in Section 2, in the HTRC context, every *workset* that exists is related to at least one *item* (although in practice each *workset* gathers together hundreds or thousands of *items*). A summary the fundamental class and relationship appears in **Table 1** below.

3.2 Metadata Features Derived From Formal Definitions

If the scope of the ontology was *collections* in general then we could move directly to metadata features that can easily be generated by computer architectures or that are based on user expectations (e.g., the number of *items* in the collection, the title of the collection, and a brief description of the collection, among other things). The scope of the HTRC Workset Ontology is much narrower than *collections* in general and, definitions D2' through D4 suggest a number of additional characteristics that differentiate *worksets* from other kinds of *collections*.

These features are recorded as metadata through the predicates listed in **Table 2** below. Definition D2' adds the concepts of curators (through the presence of agent *w*) and curatorial criteria (through the presence of criterion *c*). These are captured in the ontology through the use of the predicates *dcterms:creator* and *htrc:hasCriterion*. This is the metadata that allows *curated collections* to be distinguished from other kinds of *collections*. Definition D3 adds the requirement of a *research motivation* which the ontology captures through the predicate *htrc:hasResearchMotivation* which allows *research collections* to be distinguished from other kinds of *curated collections*. Finally definition D4 adds the requirement that every *workset* is a *research collection* that is intended to be analyzed within a non-consumptive computational environment and is recorded through the predicate *htrc:intendedForUse*.

While these are key distinctions for the formal definitions of a *workset*, an actual implementation of the resulting ontology will necessitate an amount of data

Entity / Property	Type	Definition
htrc:Workset	Class	A sub-type of dcmi:Collection with an additional Expectation constraint. An instance of the htrc:Workset class MUST be associated with a Workset as defined in aggregate by D1–D4.
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.

Table 1: HTRC Workset Core 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[7]	1+

Table 2: HTRC Workset Core Metadata Vocabulary.

entry on the part of the digital humanities scholar. For this reason the cardinality of several of the properties are zero or more (i.e., they are optional) and for those that are one or more, we expect that it is the case that the metadata can be derived automatically. In the creator's case through their secure credentials at the time they create their *workset*. In the case of the *usage intention*, since all *worksets* share at the expectation that they will be employed within a non-consumptive computational environment, it can be hardcoded. We have left the option open to the end user to add additional *usage intentions* so that they may, at their prerogative, go so far as to record precisely what computational analytics processes they intend the *workset's items* be analyzed with. Since not every *workset* will comprise *items* suitable for analysis by every analytics algorithm our intention here is that scholars indicate to one another precisely which algorithms their *workset's* contents are suitable for.

4 Additional Features and Future Work

During the process of formalizing the HTRC's notion of *worksets* a large number of use cases were compiled and analyzed. These use cases, along with features inherent to computational environments, led to the inclusion of a number of additional metadata features in the HTRC Workset Ontology that are not discussed in this article. A table containing a complete accounting of the HTRC Workset Ontology appears in **Appendix A**.

Our future work includes continuing to analyze methodologies for asserting what kinds of things have been gathered into the *worksets*. As alluded to above, we have a large number of use cases stating that the granularity of things that can be gathered into *worksets* needs to be as fine as possible [5], to the point that a *workset* may comprise individual poems, paragraphs, or even smaller tokens. Similarly, there is a great deal of desire on the part of scholars to gather together resources from different corpuses, reuse data products, and even exploit intermediary results from cleaning algorithms. While these inputs are illustrated in **Figure 1**, more work needs to be done developing an *item*-level ontology that makes these kinds of intermixtures easier. We made an exploratory start on this work in our technical report [7] and expect to continue developing the ontology in that direction.

Another area that we expect to explore in the future is how best to communicate the scope of the *kinds of things* gathered into *worksets*. Past approaches such as the DCMI-CAP tried to capture this kind of information through the predicate *dcterms:itemType*. Unfortunately content type is particular to each *item* in the *workset* and so having a first-class property of the *workset* itself attempt to capture this information seems like a solution with dubious semantics. One different approach, which we suggested in our technical report, is to define a series of *collection-by-content-type* classes; however, since the semantics of asserting multiple RDF *types* for an entity are not well established, we have yet to conclude that this is actually a viable approach to capturing this kind of metadata.

Finally, we realize axiom A2' is at best a gloss. We anticipate delving more deeply into the true nature of *curated collections* in order to arrive at a more precise definition using more expressive, higher orders of logic. By better defining the relationships between sets of curatorial criteria and *items* possessing some factor that matches one or more of those criteria we hope to extend the ontology in such a way that it will be able to realize the benefits of previous work describing the relationships that obtain between *collections* and *items* [14, 15].

Competing Interests

The authors declare that they have no competing interests.

Appendix A
HTRC Workset Vocabulary (Entities)

Entity	Type	Definition
htrc:Workset	Class	A sub-type of dcmi:Collection with an additional Expectation constraint. An instance of the htrc:Workset class MUST be associated with a Workset.
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.

These collections by content type (e.g., htrc:TextCollection, htrc:ImageCollection, etc.) are mutually exclusive with one another and are expected to be co-types with htrc:Workset.

Appendix B
HTRC Workset Vocabulary (Predicates)

Predicate	Domain	Range	Cardinality
edm:gathers	dcmi:Collection	rdfs:Resource	1+
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	1+
dcterms:extent	htrc:Workset	xsd:integer	1
dcterms:created	htrc:Workset	xsd:date	1
dcterms:publisher	htrc:Workset	dcterms:Agent	0+
dcterms:title	htrc:Workset	xsd:string	1
dcterms:abstract	htrc:Workset	rdfs:Resource or rdfs:Literal	0 or 1
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+

[dcmi]: http://purl.org/dc/dcmitype/
[dcterms]: http://purl.org/dc/terms/
[edm]: http://www.europeana.eu/schemas/edm/
[htrc]: http://wcsa.htrc.illinois.edu/#
[rdfs]: http://www.w3.org/2000/01/rdf-schema#
[xsd]: http://www.w3.org/2001/XMLSchema#

Notes

[1] Although we leave the door open as to whether or not they are a kind of *scholarly research collection* – a distinction that seems to have dubious value as defining what a *scholar* is seems even more difficult than defining what a *collection* is.

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

[3] Note that the notion of *x* being selected according to the wishes of *w* through the means of matching *c* falls out of this gloss. This is addressed in greater detail directly below in (4).

[4] Note that we could probably further simplify this by eliminating the notion of criterion and just noting that x is selected according to w through the following two axioms:

A2"a: $\forall c\ (\exists x \exists w((meetsCriterion(x,c) \wedge definedBy(c,w)) \leftrightarrow selectedAccordingTo(x,w)))$
and
A2"b: $\forall y\ (\exists x \exists w\ (isGatheredInto(x,y) \wedge selectedAccordingTo(x,w)) \leftrightarrow CuratedCollection(y))$

In this interpretation the role of criterion *c*, that curatorial agent *w* has defined, is abstracted away, thereby highlighting curatorial agent *w*'s role in the selection of *x* for the collection. The HTRC Workset Ontology also neglects to mention the actual selecting process itself. This is not because we undervalue it but rather it is because it is both the case that representing *processes* through metadata can be cumbersome and that the fact that *x* has been selected for *y* in accordance to *w*'s wishes is something that can be inferred by *x* meeting criterion *c* and *w* defining that criterion *c*. We plan a more thorough investigation of this intricate group of relationships in the future.

[5] It seems likely that there will be a great deal of overlap between curatorial criteria and research motivation; however, for the purposes of developing a metadata vocabulary, distinguishing between these two concepts allows us to provide places to record the specifics of curatorial policy separately from the generalizations of motivating research questions.

[6] As summarized in [1] and [4].

[7] These are expected to be drawn from an ontology of concepts such as SKOS.

References

1. **Band, J** 2008 A guide for the perplexed: Libraries and the Google library project settlement. *LLRX.com: Legal Research*, 14-December-2008. Available from http://www.llrx.com/features/googleprojectsettlement.htm (Accessed on 16 September 2015).
2. **Bradner, S** 1997 Key words for use in RFCs to Indicate Requirement Levels [RFC2119]. *IETF Best Current Practice*. Available from https://tools.ietf.org/html/rfc2119 (Accessed on 5 October 2015).
3. **Currall, J, Moss, M** and **Stuart, S** 2004 What is a collection? *Archivaria, 58*: 131–146.
4. **Erway, R** 2009 Impact of the Google Book settlement on libraries. *OCLC Research Report*. Available from: http://www.oclc.org/content/dam/research/publications/library/2009/2009-01.pdf (Accessed on 16 September 2015).
5. **Fenlon, K, Senseney, M, Green, H, Bhattacharyya, S, Willis, C** and **Downie, J S** 2014 Scholar-built collections: A study of user requirements for research in large-scale digital libraries. *Proceedings of the 77th ASIS&T Annual Meeting*. Seattle, WA (31 October – 5 November, 2014).
6. **Hill, L, Janee, G, Dolin, R, Frew, J** and **Larsgaard, M** 1999 Collection metadata solutions for digital library applications. *Journal of the American Society for Information Science, 50*(13): 1169–1181. DOI: http://dx.doi.org/10.1002/(SICI)1097-4571(1999)50:13<1169::AID-ASI3>3.0.CO;2-3
7. **Jett, J** 2015 Modeling worksets in the HathiTrust Research Center. CIRSS Technical Report WCSA0715. Champaign, IL: University of Illinois at Urbana-Champaign. Available From http://hdl.handle.net/2142/78149.

8. **Lynch, C** 2002 Digital collections, digital libraries, and the digitization of cultural heritage information. *First Monday, 7*(5). DOI: http://dx.doi.org/10.5210/fm.v7i5.949
9. **Mullin, J** 2013 Google Books ruled legal in massive win for fair use (updated). *Ars Technica: Law & Disorder / Civilization & Discontents*, 14-November-2013. Available from http://arstechnica.com/tech-policy/2013/11/google-books-ruled-legal-in-massive-win-for-fair-use/ (Accessed on 16 September 2015).
10. **Palmer, C L** 2004 Thematic research collections. In Schreibman, S, Siemens, R and Unsworth, J (Eds.), *A Companion to Digital Humanities*. Blackwell Publishing, Oxford. DOI: http://dx.doi.org/10.1002/9780470999875.ch24
11. **Palmer, C L** and **Knutson, E** 2004 Metadata practices and implications for federated collections. *Proceedings of the 67th ASIS&T Annual Meeting*. Providence, RI (12–17 November 2004).
12. **Palmer, C L, Knutson, E, Twidale, M** and **Zavalina, O** 2006 Collection definition in federated digital resource development. *Proceedings of the 69th ASIS&T Annual Meeting*. Austin, TX (3–8 November 2006).
13. **Palmer, C L, Isaac, A, Wickett, K M, Fenlon, K** and **Senseney, M** 2015 *Digital collection contexts: iConference 2014 workshop report*. CIRSS technical report 20150301. Champaign, IL: Center for Informatics Research in Science and Scholarship. Available from http://hdl.handle.net/2142/73359 (Accessed on 5 May 2015).
14. **Renear, A H, Wickett, K M, Urban, R J, Dubin, D** and **Shreeves, S** 2008 Collection/Item metadata relationships. *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2008* (Berlin, Germany, 22–26 September 2008).
15. **Wickett, K M, Renear, A H** and **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).
16. **Wickett, K M, Renear, A H** and **Furner, J** 2011 Are collections sets? *Proceedings of the 74th ASIS&T Annual Meeting*. New Orleans, LA (9–13 October 2011).
17. **Wickett, K M, Isaac, A, Fenlon, K, Doerr, M, Meghini, C, Palmer, C L** and **Jett, J** 2013 *Modeling cultural collections for digital aggregation and exchange environments*. CIRSS Technical Report. Champaign, IL: University of Illinois at Urbana-Champaign.

How to cite this article: Jett, J, Cole, T W, Maden, C and Downie, J S 2016 The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. *Journal of Open Humanities Data* 2: e1, DOI: http://dx.doi.org/10.5334/johd.3

Published: 18 March 2016