HathiTrust Research Center
Workset Data Model

Jacob Jett
Chris Maden
Timothy W. Cole
J. Stephen Downie

21 October 2015, E-Research Round Table, GSLIS
Workset Data Modeling Goals

• Formalize notion of worksets as a kind of scholarly research collection
• Analyze accumulated use cases and derive functional requirements
• Experiment with RDF and graph-based technologies
Use Cases

• Worksets are research products
  – Must be persistently citable
  – Must be able to gather resources from outside of the HTRC context

• Worksets must afford scholars the freedom to specify the granularity of their units of analysis
  – Must be able to gather specific pages, chapters, poems, etc.
  – Must be able to chunk text arbitrarily

• Worksets must support the non-consumptive research paradigm

• Worksets afford leveraging of data in graph form
  – Allow some data to be inferred rather than redundantly stored
  – Support linking at more granular levels
Representational Requirements for Worksets

• A workset is a container – analogous to a research collection;
• A workset is a persistent, globally unique entity;
• A workset possesses properties supporting change awareness;
• A workset’s membership requirements must be flexible enough to allow for:
  – Granularity of resources and
  – Source of resources; and
• A workset’s descriptive properties are informed by the properties of its members.
Implications

• Need to be agnostic with regards to *what* can be gathered into a workset – i.e., worksets gather bibliographic resources

• Metadata describing bibliographic resources must be rich enough to
  – Facilitate selection by scholars
  – Support inferencing by the system

• Bibliographic resources must also be immutable
Simple Conceptual Model

Analytic Module
- Name
- Description
- Creator
- Date Created
- Last Date Changed

Workset
- Title
- Description
- Creator
- Extent
- Date Created
- Last Date Changed
- Research Questions
- Intended Algorithms

Resource
- Title
- Description
- Creator
- Extent
- Date Created
- Last Date Changed

isAnalyzedBy
1..+

isGatheredInto
1..+
Worksets are Collections

∀y(∃x isGatheredInto(x, y) ↔ Collection(y))¹

Worksets are special kinds of collections

• They are a kind of curated collection
  – Items in them match collecting criteria
  – Those criteria are established by some agent
• They are a kind of research collection
  – They have a motivating research purpose
• They are intended for use with analytics algorithms
Core Workset Metadata

Creator_1

Selection Criterion_1

dcterms:creator

htrc:hasCriterion

htrc:hasResearch Motivation

htrc:intendedFor UseWith

Research Motivation_1

Analytics Algorithm_1
Some Workset metadata is intrinsic to digital architectures

- `dcterms:created`
- `dcterms:extent`
- "date"
- "quantity"
Some Workset metadata is a byproduct of human interactions.
Some Workset metadata is derived from its items.
Derived metadata is derived according to established rules (Wickett et al., 2010)

- $\forall y \forall z \left( \left( \text{language}(y, z) \land \text{Collection}(y) \right) \rightarrow \exists x \left( \text{isGatheredInto}(x, y) \land \text{language}(x, z) \right) \right)$

- $\forall y \forall z \left( \left( \text{temporalCoverage}(y, z) \land \text{Collection}(y) \right) \rightarrow \exists x \left( \text{isGatheredInto}(x, y) \land \exists w \left( \text{date}(x, w) \land \text{temporalWithin}(w, z) \right) \right) \right)$

- $\forall y \forall z \left( \left( \text{itemType}(y, z) \land \text{Collection}(y) \right) \rightarrow \exists x \left( \text{isGatheredInto}(x, y) \land \exists w \left( \text{type}(x, w) \land \text{generalizes}(w, z) \right) \right) \right)$

- $\forall y \forall z \left( \left( \text{itemFormat}(y, z) \land \text{Collection}(y) \right) \rightarrow \exists x \left( \text{isGatheredInto}(x, y) \land \text{format}(x, z) \right) \right)$
## Complete Vocabulary

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Domain</th>
<th>Range</th>
<th>Cardinality</th>
</tr>
</thead>
<tbody>
<tr>
<td>edm:isGatheredInto</td>
<td>dcmi:Collection</td>
<td>rdfs:Resource</td>
<td>1+</td>
</tr>
<tr>
<td>dcterms:creator</td>
<td>htrc:Workset</td>
<td>dcterms:Agent</td>
<td>1+</td>
</tr>
<tr>
<td>htrc:hasCriterion</td>
<td>htrc:Workset</td>
<td>rdfs:Resource or rdfs:Literal</td>
<td>0+</td>
</tr>
<tr>
<td>htrc:hasResearchMotivation</td>
<td>htrc:Workset</td>
<td>rdfs:Resource or rdfs:Literal</td>
<td>0+</td>
</tr>
<tr>
<td>htrc:intendedForUseWith</td>
<td>htrc:Workset</td>
<td>rdfs:Resource or rdfs:Literal</td>
<td>1+</td>
</tr>
<tr>
<td>dcterms:extent</td>
<td>htrc:Workset</td>
<td>xsd:integer</td>
<td>1</td>
</tr>
<tr>
<td>dcterms:created</td>
<td>htrc:Workset</td>
<td>xsd:date</td>
<td>1</td>
</tr>
<tr>
<td>dcterms:publisher</td>
<td>htrc:Workset</td>
<td>dcterms:Agent</td>
<td>0+</td>
</tr>
<tr>
<td>dcterms:title</td>
<td>htrc:Workset</td>
<td>xsd:string</td>
<td>1</td>
</tr>
<tr>
<td>dcterms:abstract</td>
<td>htrc:Workset</td>
<td>rdfs:Resource or rdfs:Literal</td>
<td>0 or 1</td>
</tr>
<tr>
<td>dcterms:language</td>
<td>htrc:Workset</td>
<td>rdfs:Resource or rdfs:Literal</td>
<td>1+</td>
</tr>
<tr>
<td>dcterms:temporal</td>
<td>htrc:Workset</td>
<td>rdfs:Resource or rdfs:Literal</td>
<td>1</td>
</tr>
<tr>
<td>dcterms:format</td>
<td>htrc:Workset</td>
<td>rdfs:Resource or rdfs:Literal</td>
<td>0+</td>
</tr>
</tbody>
</table>
Future Work

• Item-level Data Model
  – Expanding the granularity of what can be an item
    • Looking at Web Annotation standard to facilitate this
    • Also pursuing the use of persistent identifiers for pages
  – Overcoming implementation side issues with item sources
    • Best practices for identifying and retrieving documents from corpuses outside of the HathiTrust Digital Library
    • Proof-of-concept testing strongly indicates that this can be done

• Resolve open questions of how to represent worksets by content type of their items
  – Can we use the rdf:type predicate to express this?
  – Or would our own predicate be a better solution?