HathiTrust Research Center
Workset Creation for Scholarly Analysis
Project Updates

Jacob Jett
Colleen Fallaw
Timothy W. Cole
J. Stephen Downie
Megan Senseney
HathiTrust
Digital Library
Offers 13.2 million volumes digitized from libraries around the world

University of Michigan

HathiTrust
Research Center
Develops tools and infrastructure to make this massive amount of digital text more useful for researchers

University of Illinois | Indiana University
# Non-consumptive Research

<table>
<thead>
<tr>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Scholarly analysis of copyright-protected resources, without violating intellectual property laws.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>• A machine acts on the copyright-protected content as an agent of the scholar, only releasing work products that cannot be used to reproduce the protected content.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Workset Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>• From scholar’s perspective, a workset is a referential collection used as a citable description of the data analyzed.</td>
</tr>
</tbody>
</table>
Shaping Workset Definition

- HTRC computational requirements
- Scholar needs analysis and citability
- WCSA prototyping projects

workset definition
## HTRC Worksets Today

<table>
<thead>
<tr>
<th>Search and Select</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workset Builder</td>
</tr>
<tr>
<td>Solr Proxy API</td>
</tr>
<tr>
<td>Custom (MODS DB/ Hathifiles)</td>
</tr>
</tbody>
</table>

***Set of Volume IDs + metadata in Registry***

<table>
<thead>
<tr>
<th>Portal</th>
</tr>
</thead>
<tbody>
<tr>
<td>HathiTrust Custom PD Dataset</td>
</tr>
</tbody>
</table>

## Scholarly Analysis

<table>
<thead>
<tr>
<th>Agent in Portal</th>
<th>Data API</th>
<th>Distributed Collaboration</th>
</tr>
</thead>
</table>
Current Limits

Selection
- MARC records
- full text search

Granularity
- HathiTrust volumes

Citability
- Invisible updates to OCR text, image files, ordering
Workset Creation For Scholarly Analysis

Refining workset definition and tools from a scholarly needs perspective

- User Needs Studies
- Prototyping Projects
- Formal Modeling
User Needs Studies

• Conducted through a series of focus groups and interviews

• Primary Findings:
  – “Collections serve as research products in their own right”
  – “Collections serve as reusable resources”
  – “Units of analysis described by respondents varied widely in both abstraction and granularity.”

Prototyping Projects

- **Oxford University**
  - ELEPHâT: Early English Print in HathiTrust, a Linked Semantic Workset Prototype

- **Texas A&M University**
  - Workset Creation through Image Analysis of Document Pages

- **University of Maryland (MITH)**
  - Distributed Metadata Correction and Annotation

- **Waikato University**
  - Semantic Analysis of Documents from the HathiTrust Corpus

Find out more: http://worksets.htrc.illinois.edu/worksets/
Emerging Themes

• Scholar-centric approach
• Metadata/Workset/Analysis Cycle
• Mix resources from HathiTrust and other
• Linked Open Data (RDF)
• Granularity/Identifiers
  – Page
  – Arbitrary (e.g., images, poems, chapters, 500-word blocks of text)
Metadata/Workset/Analysis Cycle

- Metadata is data.
- Some data is metadata.
  - Extracted Features
  - Genres
  - Author Gender
  - Alignment with other resources
  - Image location and type
  - Corrections
  - Annotations
  - Context
Oxford University
ELEPHãT: Early English Print in HathiTrust, a Linked Semantic Workset Prototype

Prototype

• Software that exposes the necessary metadata from individual collections for building aggregate worksets drawn from multiple sources.
• Prototype will build integrated worksets from the HathiTrust and the Early English Books Online Text Creation Partnership (EEBO-TCP) collection, which focuses on high quality images and accurate transcriptions of items usually found in libraries’ special collections.

Tasks

• Learn use cases from scholars
• Coordinate on RDF Representations of bibliographic resources
• Develop software that can use SPARQL queries to align resources
Texas A&M University
Image Analysis of Document Pages

Prototype

• Software that uses the visual characteristics of digitized printed pages to identify documents that contain three types of visually distinctive materials of interest to humanities researchers: poetry, music, and illustrations.
• Prototype will include a store of information about identified image page location and type.

Tasks

• Learn use cases from scholars
• Develop software that can identify images
• Develop API for interacting with the recorded image info
• Build a demo app that uses the API
Prototype

- A set of services and interfaces that pull metadata records from the HathiTrust, correct and annotate these records using standardized vocabularies, gather corrections and annotations from other teams or scholars, and export enhanced metadata in formats suitable for publication as linked data.

Tasks

- Iteratively integrate feedback from scholars on prototype versions
- Use HathiTrust Bib API to pull MARC XML records
- Create CSV / CSV for the Web files for use in OpenRefine
- Develop prototype interface for scholar metadata corrections
Prototype

- A suite of tools that analyze documents by the semantics of their content and metadata. Clustering documents by semantic similarity will open up a wealth of opportunities for scholarly research.
- Prototype application will focus on areas of Maori & Pacific Studies, and Historical Anthropology.

Tasks

- Iteratively integrate feedback from scholars on prototype versions.
- Build knowledge base integrated from multiple sources
  - Mining Wikipedia
  - Topic modeling of HTRC resources
- Develop search interface based on knowledge base, with interactive disambiguation
Workset Data Modeling

- Goals
- Uses Cases
- Functional Requirements
- Model
  - Entities
  - Properties
  - Visualizations
- Next Steps
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
  – Allowing some data to be inferred rather than stored in multiple places (e.g., quantity of things in the workset)
  – Removing the silo-like barriers of document-like data (i.e., metadata records)
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
Representational Requirements for Bibliographic Resources

- Bibliographic resources are persistent and globally unique;
- Bibliographic resources possess provenance properties that support:
  - Change awareness and,
  - Disambiguation from other, similar bibliographic resources; and
- Bibliographic resources must possess metadata rich enough to support their discovery.
## Entities & Properties

<table>
<thead>
<tr>
<th>Entity</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>htrc:Workset</td>
<td>A sub-type of dcmi:Collection</td>
</tr>
<tr>
<td>dcterms:BibliographicResource</td>
<td>“A book, article, or other documentary resource.”</td>
</tr>
<tr>
<td>htrc:Volume</td>
<td>A sub-type of htrc:BibliographicResource; analogous to a book</td>
</tr>
<tr>
<td>dcterms:Agent</td>
<td>“A resource that acts or has the power to act.”</td>
</tr>
<tr>
<td>htrc:Record</td>
<td>A resource that describes a bibliographic resource.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Domain</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>htrc:isGatheredInto</td>
<td>dcterms:BibliographicResource</td>
<td>htrc:Workset</td>
</tr>
<tr>
<td>dcterms:title</td>
<td>htrc:Workset</td>
<td>xsd:string</td>
</tr>
<tr>
<td>dcterms:abstract</td>
<td>htrc:Workset</td>
<td>rdfs:Resource or rdfs:Literal</td>
</tr>
<tr>
<td>dcterms:created</td>
<td>htrc:Workset</td>
<td>xsd:date</td>
</tr>
<tr>
<td>dc:creator</td>
<td>htrc:Workset</td>
<td>dcterms:Agent</td>
</tr>
<tr>
<td>htrc:bibliographicRecord</td>
<td>dcterms:BibliographicResource</td>
<td>htrc:Record</td>
</tr>
</tbody>
</table>
## Some Inferred Properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Source of Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection Identifier [dc:identifier]</td>
<td>Synonymous with the Workset entity itself.</td>
</tr>
<tr>
<td>Size [dcterms:extent]</td>
<td>Can be derived from the number of htrc:isGatheredInto predicates in the graph.</td>
</tr>
<tr>
<td>Item Type [cld:itemType]</td>
<td>Can be inferred from the metadata for the resources gathered into the Workset.</td>
</tr>
<tr>
<td>Item Format [cld:itemFormat]</td>
<td>Can be inferred from the metadata for the resources gathered into the Workset.</td>
</tr>
<tr>
<td>Rights [dc:rights]</td>
<td>Can be inferred from the metadata for the resources gathered into the Workset.</td>
</tr>
<tr>
<td>Access Rights [dcterms:accessRights]</td>
<td>Can be inferred from the metadata for the resources gathered into the Workset.</td>
</tr>
</tbody>
</table>
Model visualized as a graph
Associating Metadata Records with Volumes to...

- **Volume** isA Digitized Image Set
- Digitized Image Set hasTitle [The game and play of chess]
- HathiTrust heldBy Digitized Image Set
- Digitized Image Set describedBy Metadata Record
- Metadata Record assertedBy Metadata Record
- Metadata Record isA Named Graph
...support disambiguation and provenance claims
Next steps for modeling work

• Support for finer grained workset members (a.k.a. Bibliographic Granules)
• Support for better provenance metadata, ensuring “immutability” of entities in the model
• Support for more abstract bibliographic entities (e.g., Works)
Off the deep end...

Using Web Annotation to capture metadata that contextualizes granule locality

Poem \( \text{rdf:type} \) oa:SpecificResource

Page \( \text{htrc:hasBibliographicSource} \)

Selector \( \text{oa:TextPositionSelector} \)

“123” \( \text{oa:start} \)

“2258” \( \text{oa:end} \)
Down the rabbit hole...

Describing pages with extracted features

Page
  after
  Page
    hasTokenCount
      "1173"^^xsd:integer
    hasLineCount
      "152"^^xsd:integer
    hasEmptyLineCount
      "17"^^xsd:integer
    hasSentenceCount
      "47"^^xsd:integer

Page
  after

Describing whole-work granules with BIBO

http://vif.org/vif/95230688
  rdf:type htrc:BibliographicResource
    dcc:creator
    dc:title
      "She walks in beauty"^^xsd:string
    dc:isPartOf
      mdp:49015002132851
    bibo:volume
      "3"^^xsd:integer

Poem
  bibo:startPage
    Page 381
  bibo:endPage
    Page 381
A bridge too far...

Using BRM to better articulate Works, Expressions, Manifestations, Items and everything in between.