Rethinking HathiTrust Metadata to Support Workset Creation for Scholarly Analysis
Katrina Fenlon, Timothy Cole, Myung-Ja Han, Craig Willis, Colleen Fallaw
University of Illinois at Urbana-Champaign (USA)

Research question
Existing bibliographic metadata records underlying the HathiTrust Digital Library are inadequate to support workset creation for scholarly analysis. How can we enrich metadata & item representations in the corpus to enable the creation of worksets supportive of scholarly analysis?

A MARC record (select fields omitted for brevity):

The HathiTrust Research Center

Limitations of MARC for workset creation
• Missing properties of interest – author gender, nationality, ...
• ‘Subject’ may be empty or describe form rather than scope.
• No means to link to thesauri, external authorities.
• Some fields require extensive parsing.
• Catalogors rarely use full expressiveness.

Why Worksets?
• Scholars regularly create worksets – selecting & gathering materials from disparate sources to answer specific research questions
• Scholars need sophisticated tools for the management and manipulation of “custom collections” of digital texts

Requirements (gathered from user studies):
• Allow scholars to gather not just primary items in the HathiTrust corpus (books), but also metadata & granular, intra-book content.
• Allow integration of external sources, such as linked datasets, secondary literature, and references.
• Allow identification and description of worksets so that they function as persistent, reusable scholarly resources.

Workset Creation for Scholarly Analysis objectives
• Allow HT users to formally gather selected subsets of the HathiTrust corpus together for computational analysis.
• Enable routine computational analysis across subsets of materials in the HathiTrust corpus
• Engage scholars in tool design
• Enrich metadata in the HathiTrust corpus
• Formalize the notion of worksets

Key questions being investigated by WCSA
• Given sparseness of HathiTrust records, can we enrich the corpus metadata by distilling analytics over full text?
• Can we deploy/modify off-the-shelf tools, e.g., to determine language(s) of the text, temporal / spatial coverage, etc.?
• Can we augment string-based metadata with URIs for entities – e.g., names, subjects, place of publication, etc.?
• Can we formalize the notion of worksets in HT, e.g., defining the necessary elements of a workset? How do we balance rigor with extensibility and flexibility?

Moving from MARC to an RDF-centric architecture

The HathiTrust (HT) is a repository of over 10 million volumes (3 billion pages) of text. The HathiTrust Research Center (HTRC) enables computational access to the HT corpus. http://www.hathitrust.org/htrc/

Use of MARC Fields in HathiTrust bibliographic records.

Percent of records with value for MARC property
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Title
Publisher
Subject – Tropical
Subject – Geographic
Subject – Temporal
Fiction Literary Form
LC / Dewey Class No.

Use of MARC Fields considered in combination

<table>
<thead>
<tr>
<th>Field Type</th>
<th>% of all records</th>
<th>% of records w/ genre/fiction</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC Classification</td>
<td>45%</td>
<td>40%</td>
</tr>
<tr>
<td>Topic &amp; geographic subject(s)</td>
<td>36%</td>
<td>9%</td>
</tr>
<tr>
<td>Topic subject(s) only</td>
<td>36%</td>
<td>14%</td>
</tr>
<tr>
<td>Geographic subject(s) only</td>
<td>1%</td>
<td>&lt; 1%</td>
</tr>
</tbody>
</table>