Assessment of API-generated Links to HathiTrust, Internet Archive, and Google Books Affiliation

Andrew Hart
Chris Holden
Margaretta Yarborough

5/8/14
Background

• UNC-Chapel Hill uses application programming interfaces (API) to link patrons to digitized surrogates of works in the public domain.
• These APIs provide links to Google Books, Internet Archive, and HathiTrust, among others.
Background

- For the most part, these API-generated links seem to work fine.
- However, there are some problematic and confusing links to these digitized resources.
What are the problems?

• Problems include:
  ▫ Linking to the wrong edition of the work
  ▫ Linking to the wrong volume of a multi-volume work
  ▫ Digitized surrogate lacking complete text or missing other materials (such as illustrations).
  ▫ Linking to the wrong work entirely

• Problems stem from metadata issues on UNC’s end, metadata issues on the repositories’ side, or problems with the image capture of the item.
The project

- Derived a random sample of OCLC records used in both the UNC and HathiTrust catalogs.
- Physical copy of each item examined, comparing it to the digital copies linked in the catalog.
- Items with discrepancies in the image or the metadata were flagged for further review.
I. BIB RECORD MATCHES ITEM IN HAND?

Y___N___ title
Y___N___ author
Y___N___ publisher
Y___N___ publication date
Y___N___ publication place
Y___N___ pagination
Y___N___ n/a ___ edition
Y___N___ n/a ___ series
Y___N___ part of multi-vol. set?
   If yes, # of vols. in bib record: _______

II. NUMBER OF ITEMS PER LINK IN ENDECA:

___ HathiTrust
___ Internet Archive—links within IA
___ Google
___ TOTAL LINKED ITEMS

III. LINKED ITEMS MATCH ITEM IN HAND?

Sample vol. not dig., alternate vols. viewed:
HT: ______  IA: ______  G: ______

Y = Yes; N = No; n/a = does not apply

<table>
<thead>
<tr>
<th></th>
<th>HT</th>
<th>IA</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>title</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>author</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pub.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pub.date</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pub. pl.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pag.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>edition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>series</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>illus.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>frontis.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all vols. dig.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sample vol. dig.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>source (HT, IA, G)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

856 field: No Yes
Results and Preliminary Observations

- 289 items with digital surrogates selected, to provide a statistically significant sample.
- Of these, 6 items were not able to be located. A total of 283 items were physically examined and compared to online surrogates.
- These 283 items had 508 digital surrogates available.
  - Some items had no links even when there was a digital surrogate available.
  - Other items had multiple links to different websites.
Results and Preliminary Observations

- Links present in our catalog:
  - DocSouth: 1
  - Google Books: 138
  - Government Publications Office: 2
  - Harvard: 1
  - HathiTrust: 236
  - Internet Archive: 116

- Results don’t add up to 508, because a few items had digital surrogates that were not linked in our catalog.
Results and Preliminary Observations

• 145 links (28.54%) reviewed in greater detail because of an identified discrepancy between the physical item and the digital surrogate.

• Many of these discrepancies determined to be minor.
  ▫ Slightly different pagination
  ▫ Missing frontispiece
  ▫ Metadata mismatch but digitized images match item
Results and Preliminary Observations

- 38 digital items (7.48% of examined items; 26.21% of items with discrepancies) were determined to be “not good enough.”
- Examples of items “not good enough” include:
  - Different edition of same work
  - Different volume of same series
  - Warped or unreadable text
  - Absence of crucial supplementary material (e.g., foldout maps, illustrations)
Results and Preliminary Observations

- Of these 38 “not good enough” items, 7 were not linked in our catalog at all.
- Of the remaining 31 links:
  - 13 are missing materials (including 4 with fold-out maps that would be essential to the work)
  - 1 is a bad scan
  - 6 lead to the wrong edition of the work
  - 4 have our book attached to the wrong WorldCat record
  - 2 stem from local cataloging errors
  - 5 stem from bibliographic problems surrounding multi-volume sets and serials.
Results and Preliminary Observations

• Clear majority of links lead to the expected digital copy without a problem.
• Problems are far more likely to come from the scanned image than the metadata.
• Multi-volume works are more likely to have issues
• Links flagged as “not good enough” have an overwhelming chance of Google Books being the source of the scan.
• Missing materials (pages, illustrations, maps) are the biggest problem.
• Copyright date and pagination are the two categories most likely to be a mismatch between our item and the digitized surrogate.
Questions for Further Exploration

• Are there cost-effective and feasible measures that could be taken to...
  ▫ Fill in missing content for digital manifestations?
  ▫ Provide a link to a better digital copy?
  ▫ Identify and correct metadata errors?

• Is there a way to build on cooperative cataloging programs to allow these sorts of corrections?
Special thanks to

• Anne Conway
• Wanda Gunther
• Gina Suarez

for their ideas, problem-solving, and many hours of work on the project!