Batch is the New Bespoke

Integrating Web Scraping, Batch Metadata Tools, and Streamlined Cataloging Workflows into Technical Services

Michael P. Williams & Beth Picknally Camden
Introduction

• Creating Global Studies Technical Services (GSTS)
  • Study & pilot
  • Creating a unit of language experts
  • Standardize practices, especially acquisitions & fiscal

• “One stop” model
  • Need for technology
  • Fill staffing gaps
  • Advantages of a responsive LMS
  • External tools
Some Guiding Principles

• Determine the scope of, and embed, the highest-quality metadata for with as little manual work as possible (e.g. reduce copy/paste)
• Piggyback on existing software applications (e.g. Excel/Google Sheets) or plugins with as little manual customization as possible
• Stretch skillsets of staff to become competent and agile users of those software
• Create workflows that are open to change and transparent
• Move books off tech services shelves and get them into patron spaces as quickly as possible
MatchMARC + MarcEdit Workflow

• Use tabular data (with ISBNs in Excel) from vendor or minimal staff input to build useful catalog records

• MatchMARC relies on the presence of ISBNs (or LCCNs), then queries OCLC and returns data requested using an API Key (http://www.ala.org/core/using-matchmarc).

• Retrieved data can be embedded in original Excel file, then transformed to MARC records with MarcEdit’s Delimited Text Translator. These form the basis of acquisitions records.

• With retrieved OCLC numbers, MatchMARC can email records as MarcXML. These can be overlaid onto acquisitions records with the OCLC number as a match point and “batch-catalog” the books in advance.
Using Vendor/Local Data in MatchMARC

<table>
<thead>
<tr>
<th>Vendor data in Excel</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Excel Spreadsheet" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MatchMARC criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image2.png" alt="MatchMARC Criteria" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MatchMARC search results</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3.png" alt="MatchMARC Search Results" /></td>
</tr>
</tbody>
</table>
Comparison of Local and OCLC MARC Data

Slim MARC using vendor + MatchMARC data

MatchMARC search results
Web Scraping for ISBNs (and More!)

• We don’t always have ISBNs (or LCCNs) to plug into MatchMARC, but the web is rich in data sources.

• Google Sheets formulas fetch structured data from known URLs:
  • **IMPORTHTML** targets list (<ul>, <ol>) or table (<table>) data
  • **IMPORTXML** targets specific HTML tags (e.g. <div> or <span> or <h2>), *provided the HTML was not generated by JavaScript*
  • **IMPORTDATA** query a .csv or .tsv online. *(In many cases, these could just be copied and pasted into Excel)*.
  • Google limits “import” transactions and sometimes process is slow/hangs
Identifying Sources of Book Metadata

Sample bookseller website:

Individual book metadata:

Title: 1857 KI JUNG-E-AZADI KA GUMNAM SHAHEED RAJA NAHAR SINGH

AMRITA SINGH DR.

Subject(s): Biography

₹ 130.00

ADD TO CART

Book Details

Title: 1857 KI JUNG-E-AZADI KA GUMNAM SHAHEED RAJA NAHAR SINGH

Author: AMRITA SINGH DR.

ISBN 13: 9788178018539

ISBN 10: 8178018535

Year: 2017

Language: URDU

Pages etc.: 144p

Binding: Paperback

Subject(s): Biography
Individual book metadata

HTML source of book metadata

Google Sheets IMPORTXML query for <div> element
Formatting Scraped Data

• Web data is structured—but not always in the way we want
• Google Sheets may interpret web data in unexpected ways (hard line breaks, split across/down cells, etc.)
• Sometimes data requires more complex cleanup with Excel/Google Sheet formulas
• Individual customizations may be required per project/task
• Process is iterative and requires trial and error
Scraping Semi-Formatted Data

Individual book metadata

HTML source of ISBN metadata

Google Sheets IMPORTXML query for <div> element
## Cleaning Up and Adding to Scraped Data

### Google Sheets RIGHT formula to get ISBN

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td><strong>B</strong></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td>1</td>
<td><a href="https://www.shashasha.co/book/sanya-1968-8-1-8-20">https://www.shashasha.co/book/sanya-1968-8-1-8-20</a></td>
<td><strong>Book Size</strong>257 x 182 mm<strong>Pages</strong>44 <strong>Binding</strong>Softcover <strong>Publication</strong>Date2017 <strong>Language</strong>English, Japanese <strong>ISBN</strong>978-4-905453-54-3</td>
</tr>
</tbody>
</table>

### Google Sheets MID formula to isolate pagination

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td><strong>B</strong></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td>1</td>
<td><a href="https://www.shashasha.co/book/sanya-1968-8-1-8-20">https://www.shashasha.co/book/sanya-1968-8-1-8-20</a></td>
<td><strong>Book Size</strong>257 x 182 mm<strong>Pages</strong>44 <strong>Binding</strong>Softcover <strong>Publication</strong>Date2017 <strong>Language</strong>English, Japanese <strong>ISBN</strong>978-4-905453-54-3</td>
</tr>
</tbody>
</table>

### Additional Google Sheets IMPORTXML formula to get title, author, publisher

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td><strong>B</strong></td>
<td><strong>C</strong></td>
<td><strong>D</strong></td>
<td><strong>E</strong></td>
</tr>
<tr>
<td>shashasha.co/book/sanya-1968-8-1-8-20</td>
<td><strong>Book Size</strong>257 x 182 mm<strong>Pages</strong>44 <strong>Binding</strong>Softcover <strong>Publication</strong>Date2017 <strong>Language</strong>English, Japanese <strong>ISBN</strong>978-4-905453-54-3</td>
<td>ISBN978-4-905453-54-3</td>
<td>44 pages</td>
<td>SAN'YA 1968.8.1 - 8.20</td>
</tr>
</tbody>
</table>
Further Experimentation (and Success!)

HTML source of ISBN metadata

Google Sheets
IMPORTXML query for <dt> element
Batch Duplicate Check

• Use tabular data (with ISBNs in Excel) from vendor or minimal staff input to query catalog in bulk
  • Locally designed Google Sheet method relies on **IMPORTXML** (checks catalog website), but can be subject to slow/stalled performance
  • “Look Up in Local Catalog” Excel Add-In developed at Princeton’s East Asian Library (relies on Blacklight catalog structure)
• Penn Libraries’ LMS has a “build a set” from ISBN function, but results don’t easily integrate into tabular data
### Tabular Duplicate Check Methods

#### Princeton-developed “Look Up in Local Catalog” interface for Excel

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9787205099763</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
<tr>
<td>2</td>
<td>9789863267331</td>
<td>TRUE</td>
<td>9789863267331</td>
<td>98 Xinjiang 1945 : Jiang Jieshi zhong b</td>
</tr>
<tr>
<td>3</td>
<td>9789863267331</td>
<td>TRUE</td>
<td>9789863267331</td>
<td>(pir Zong jiao yi shu lun / Jiang Sh</td>
</tr>
<tr>
<td>4</td>
<td>9787301318614</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
<tr>
<td>5</td>
<td>9789579057516</td>
<td>TRUE</td>
<td>9789579057516</td>
<td>(pir Bao cun sheng huo : Han Bao</td>
</tr>
</tbody>
</table>

#### Princeton-developed “Look Up in Local Catalog” Excel results

1. **Formula**:
   
   ```excel
   =IFNA(IMPORTXML(CONCATENATE("https://franklin.library.upenn.edu/catalog?utf8=?
   &search_field=isxn_search&q="A2","/div[@class='availability-ajax-load']/@data-availability-ids"),"no Franklin result found")
   ```

2. **Table**:

<table>
<thead>
<tr>
<th>ISBN</th>
<th>MMS ID Return</th>
<th>Title Return</th>
<th>Franklin Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>9787205099763</td>
<td>no Franklin result found</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
   | 9789863267492 | 997768438203681 | Xinjiang 1945 : Jiang Jieshi zhong b | [Link](https://franklin.library.upenn.edu/catalog/FRANKLIN_978768438203681)
   | 9789863267331 | 997768438103681 | Zong jiao yi shu lun / 宗教藝術論 | [Link](https://franklin.library.upenn.edu/catalog/FRANKLIN_978768438103681)
   | 9787301318614 | Loading... | #/A | |
   | 9789579057516 | 9977684383903681 | Bao cun sheng huo : Han Baode | [Link](https://franklin.library.upenn.edu/catalog/FRANKLIN_9787684383903681)

#### Web-scrape-based “Franklin Checker” Google Sheet

Comment: slow search
Putting It All Together (With Local Ingenuity!)

• Instead of focusing on individual, repeatable data transactions (copy, paste, repeat), intellectual labor can move to custom, batch-enabled processes (analyze, design, fetch, enhance).

• All data is different!
  • Not every book has an ISBN
  • Not every ISBN will hit an OCLC record
  • Target records may be incorrect (duplicated ISBNs)
  • Target records may be of little use (electronic records vs print, non-English-language records, vendor records, etc.)
  • Not every page can be scraped or clean easily

• Understanding the realities and limits of existing and negotiating those with constituents (selectors, patrons, and more), e.g. less-than-full records
General Outcomes

- Order records for batch orders (beyond just titles!)
- Current budget info for bibliographers is embedded in records
- Staff skills expansion (Excel, MarcEdit, etc.)
Outcomes: Abbreviated Cataloging

- Catalog for remote storage
  - No call numbers
  - Minimal metadata (using MARC Encoding Level 3)
  - Accurate 008 (country, language, date)
- Focus areas:
  - Languages without expertise
  - Older backlogs in understaffed areas
- Patrons can request without intervention
Additional Resources

- **MatchMARC**
  - Google Sheets Add-on: [https://workspace.google.com/marketplace/app/matchmarc/903511321480](https://workspace.google.com/marketplace/app/matchmarc/903511321480)
  - January 2021 Core webinar: [http://www.ala.org/core/using-matchmarc](http://www.ala.org/core/using-matchmarc)

- **MarcEdit**
  - Home page: [https://marcedit.reeset.net/](https://marcedit.reeset.net/)

- **Google Sheets**
  - IMPORTHTML sample usage: [https://support.google.com/docs/answer/3093339?hl=en](https://support.google.com/docs/answer/3093339?hl=en)
  - IMPORTXML sample usage: [https://support.google.com/docs/answer/3093342?hl=en](https://support.google.com/docs/answer/3093342?hl=en)
  - XPATH syntax (W3Schools): [https://www.w3schools.com/xml/xpath_syntax.asp](https://www.w3schools.com/xml/xpath_syntax.asp)

- **Microsoft Excel formulas**