Batch is the New Bespoke

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

Michael P. Williams &
Beth Picknally Camden

Penn Libraries
University of Pennsylvania

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

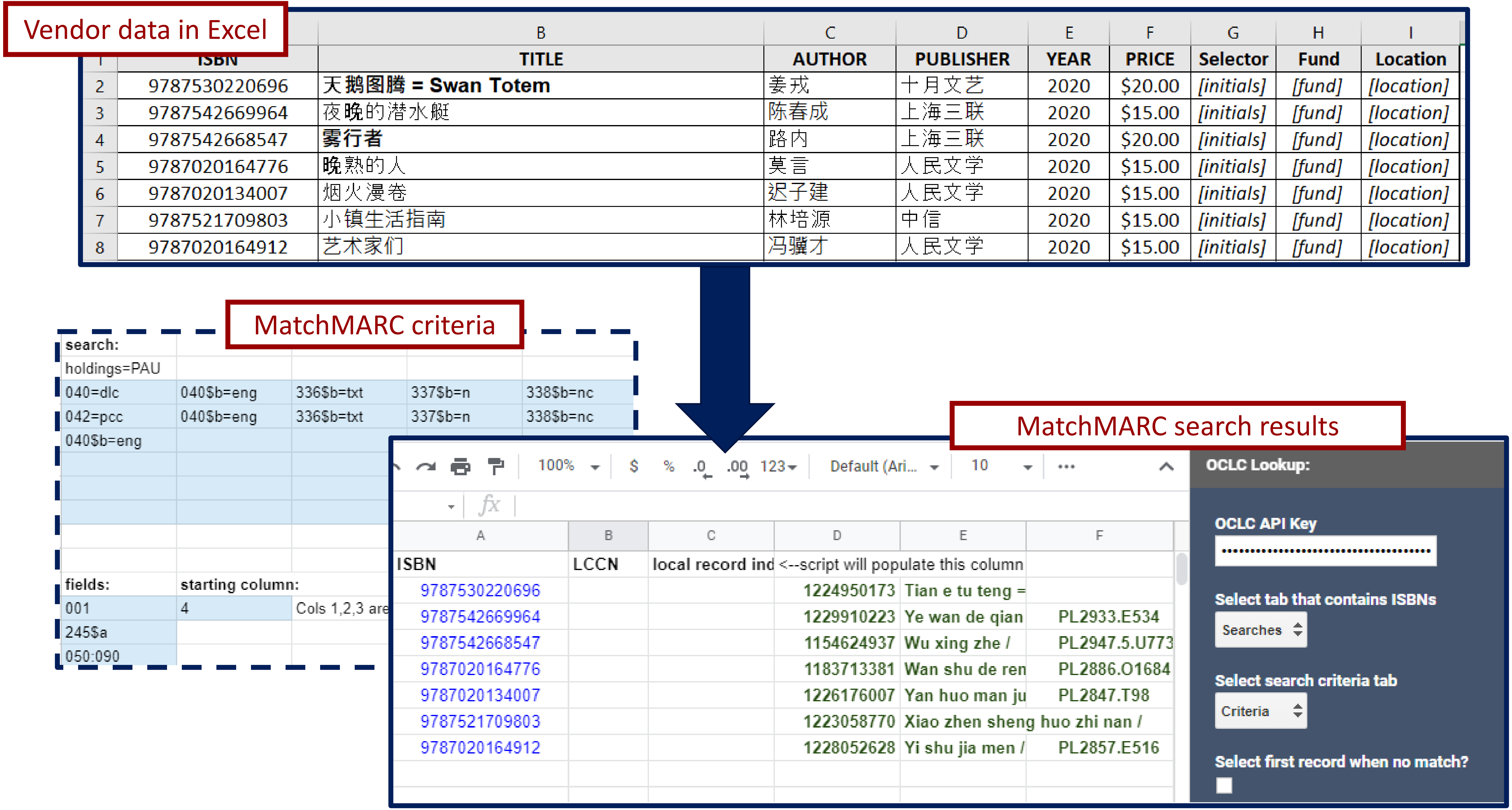
Some Guiding Principles

- 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



Vendor data in Excel

		B	C	D	E	F	G	H	I
1	ISBN	TITLE	AUTHOR	PUBLISHER	YEAR	PRICE	Selector	Fund	Location
2	9787530220696	**天鹅图腾 = Swan Totem**	姜戎	十月文艺	2020	$20.00	*[initials]*	*[fund]*	*[location]*
3	9787542669964	夜晚的潜水艇	陈春成	上海三联	2020	$15.00	*[initials]*	*[fund]*	*[location]*
4	9787542668547	**雾行者**	路内	上海三联	2020	$20.00	*[initials]*	*[fund]*	*[location]*
5	9787020164776	晚熟的人	莫言	人民文学	2020	$15.00	*[initials]*	*[fund]*	*[location]*
6	9787020134007	烟火漫卷	迟子建	人民文学	2020	$15.00	*[initials]*	*[fund]*	*[location]*
7	9787521709803	小镇生活指南	林培源	中信	2020	$15.00	*[initials]*	*[fund]*	*[location]*
8	9787020164912	艺术家们	冯骥才	人民文学	2020	$15.00	*[initials]*	*[fund]*	*[location]*

MatchMARC criteria

search:				
holdings=PAU				
040=dlc	040$b=eng	336$b=txt	337$b=n	338$b=nc
042=pcc	040$b=eng	336$b=txt	337$b=n	338$b=nc
040$b=eng				
fields:	**starting column:**			
001	4	Cols 1,2,3 are		
245$a				
050:090				

MatchMARC search results

A	B	C	D	E	F
ISBN	LCCN	local record ind	<--script will populate this column		
9787530220696			1224950173	Tian e tu teng =	
9787542669964			1229910223	Ye wan de qian	PL2933.E534
9787542668547			1154624937	Wu xing zhe /	PL2947.5.U773
9787020164776			1183713381	Wan shu de ren	PL2886.O1684
9787020134007			1226176007	Yan huo man ju	PL2847.T98
9787521709803			1223058770	Xiao zhen sheng huo zhi nan /	
9787020164912			1228052628	Yi shu jia men /	PL2857.E516

OCLC Lookup:

OCLC API Key

••••••••••••••••••••••••••••••••••••

Select tab that contains ISBNs

Searches

Select search criteria tab

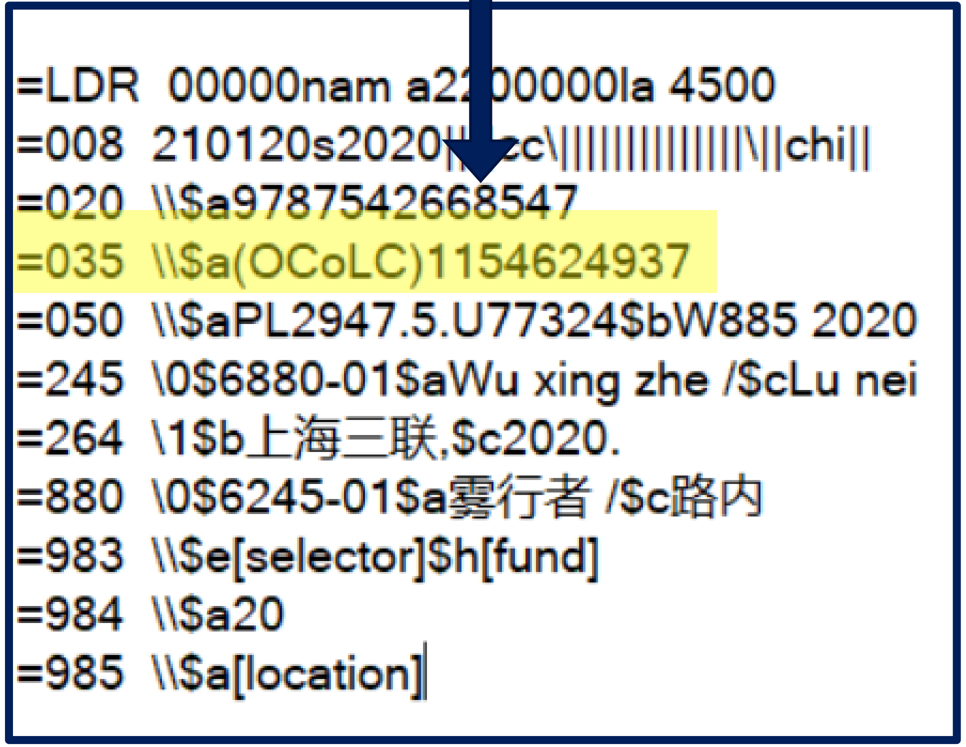
Criteria

Select first record when no match?

□

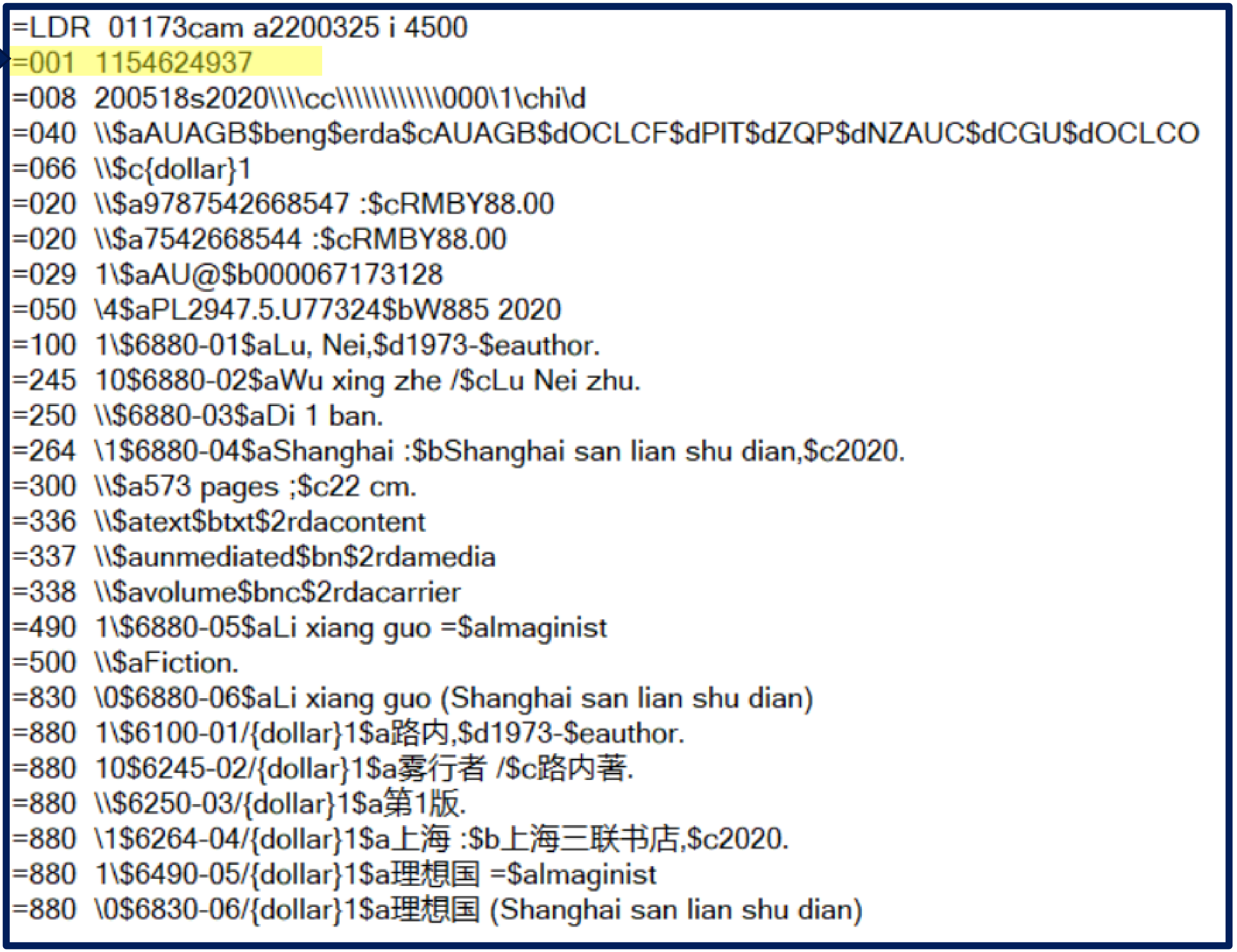
Comparison of Local and OCLC MARC Data

Slim MARC using vendor + MatchMARC data



```
=LDR  00000nam a2200000la 4500
=008  210120s2020||||cc\||||||||||||||\||chi||
=020  \\$a9787542668547
=035  \\$a(OCoLC)1154624937
=050  \\$aPL2947.5.U77324$bW885 2020
=245  \0$6880-01$aWu xing zhe /$cLu nei
=264  \1$b上海三联,$c2020.
=880  \0$6245-01$a雾行者 /$c路内
=983  \\$e[selector]$h[fund]
=984  \\$a20
=985  \\$a[location]
```

MatchMARC search results



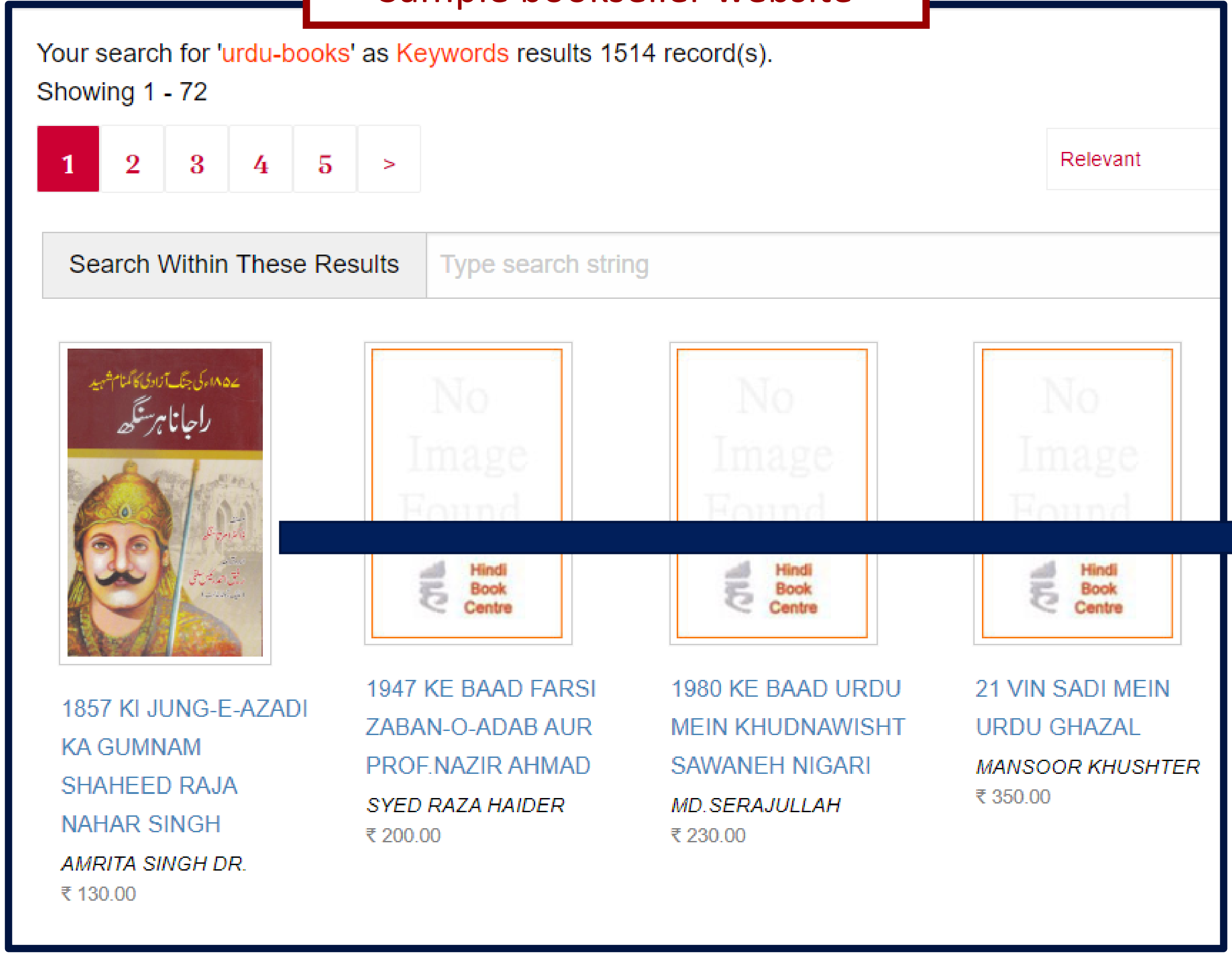
```
=LDR  01173cam a2200325 i 4500
=001  1154624937
=008  200518s2020\\\\cc\\\\\\\\\\\\000\1\chi\d
=040  \\$aAUAGB$beng$erda$cAUAGB$dOCLCF$dPIT$dZQP$dNZAUC$dCGU$dOCLCO
=066  \\$c{dollar}1
=020  \\$a9787542668547 :$cRMBY88.00
=020  \\$a7542668544 :$cRMBY88.00
=029  1\$aAU@$b000067173128
=050  \4$aPL2947.5.U77324$bW885 2020
=100  1\$6880-01$aLu, Nei,$d1973-$eauthor.
=245  10$6880-02$aWu xing zhe /$cLu Nei zhu.
=250  \\$6880-03$aDi 1 ban.
=264  \1$6880-04$aShanghai :$bShanghai san lian shu dian,$c2020.
=300  \\$a573 pages ;$c22 cm.
=336  \\$atext$btxt$2rdacontent
=337  \\$aunmediated$bn$2rdamedia
=338  \\$avolume$bnc$2rdacarrier
=490  1\$6880-05$aLi xiang guo =$aImaginist
=500  \\$aFiction.
=830  \0$6880-06$aLi xiang guo (Shanghai san lian shu dian)
=880  1\$6100-01/{dollar}1$a路内,$d1973-$eauthor.
=880  10$6245-02/{dollar}1$a雾行者 /$c路内著.
=880  \\$6250-03/{dollar}1$a第1版.
=880  \1$6264-04/{dollar}1$a上海 :$b上海三联书店,$c2020.
=880  1\$6490-05/{dollar}1$a理想国 =$aImaginist
=880  \0$6830-06/{dollar}1$a理想国 (Shanghai san lian shu dian)
```

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



Your search for 'urdu-books' as Keywords results 1514 record(s).
Showing 1 - 72

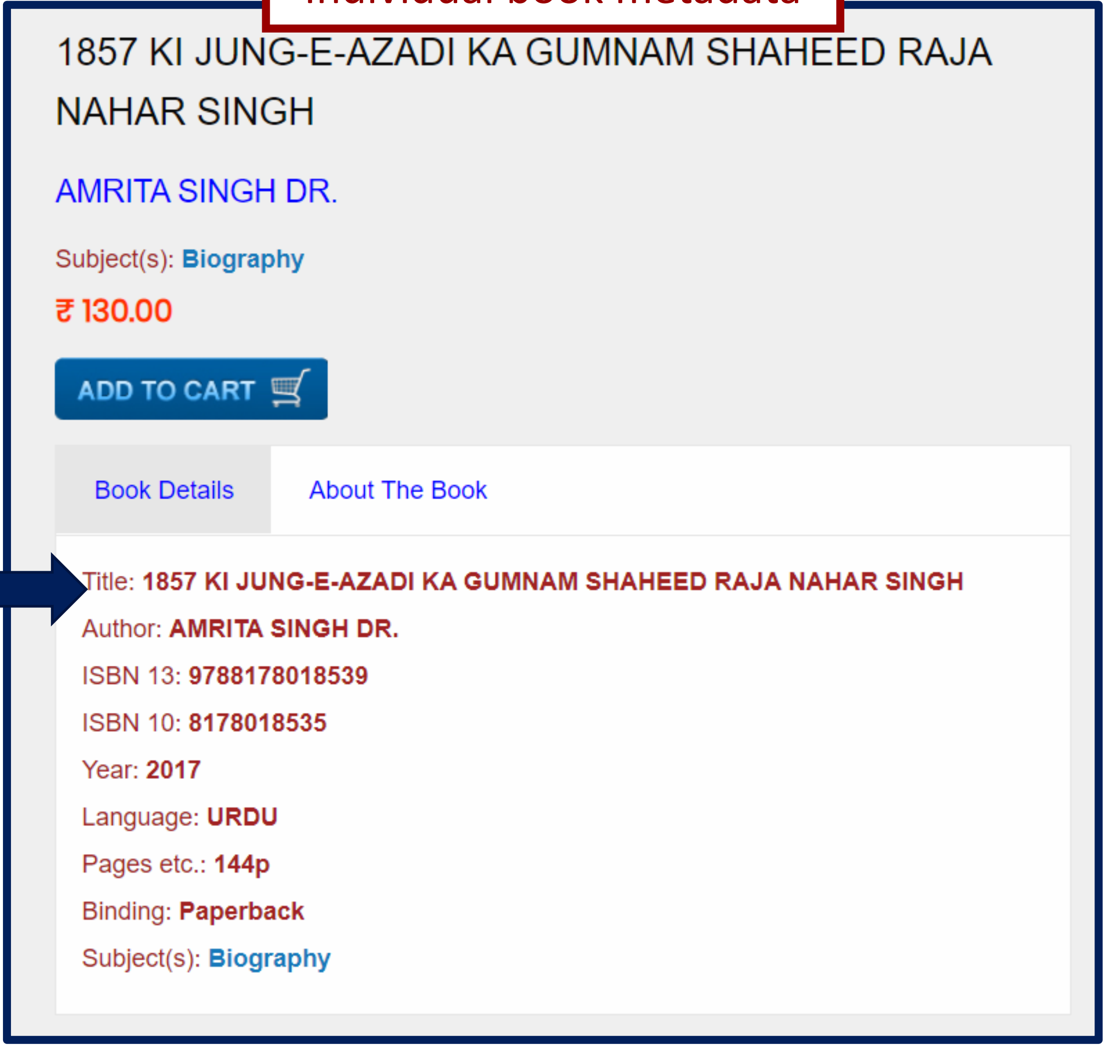
1857 KI JUNG-E-AZADI KA GUMNAM SHAHEED RAJA NAHAR SINGH
AMRITA SINGH DR.
₹ 130.00

1947 KE BAAD FARSI ZABAN-O-ADAB AUR PROF.NAZIR AHMAD
SYED RAZA HAIDER
₹ 200.00

1980 KE BAAD URDU MEIN KHUDNAWISHT SAWANEH NIGARI
MD.SERAJULLAH
₹ 230.00

21 VIN SADI MEIN URDU GHAZAL
MANSOOR KHUSHTER
₹ 350.00

Individual book metadata



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

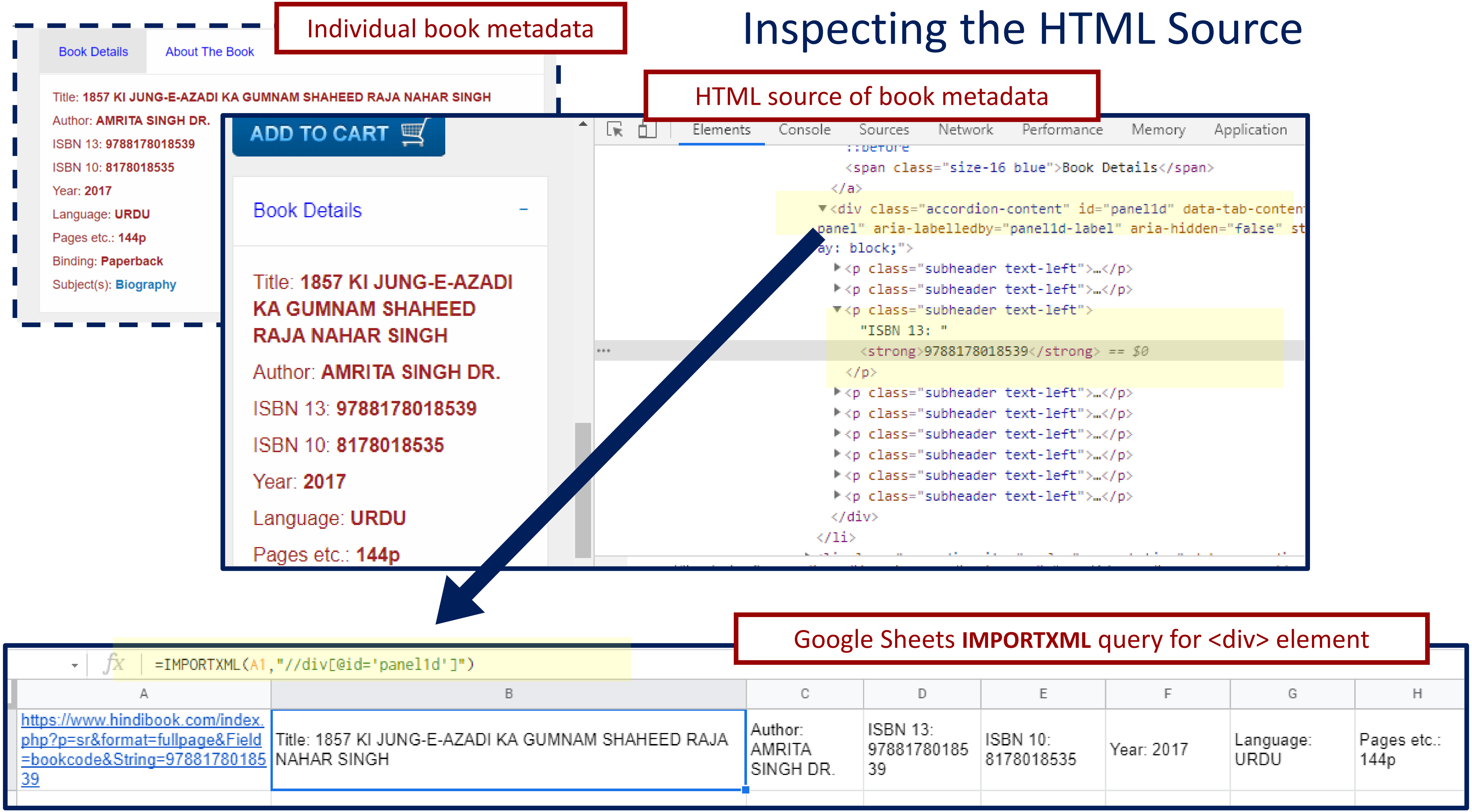
AMRITA SINGH DR.

Subject(s): **Biography**

₹ 130.00

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**

Inspecting the HTML Source



Individual book metadata

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

HTML source of book metadata

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

```
          ::before
          <span class="size-16 blue">Book Details</span>
        </a>
      ▼<div class="accordion-content" id="panel1d" data-tab-conten
      panel" aria-labelledby="panel1d-label" aria-hidden="false" st
      ay: block;">
        ▶<p class="subheader text-left">…</p>
        ▶<p class="subheader text-left">…</p>
        ▼<p class="subheader text-left">
            "ISBN 13: "
            <strong>9788178018539</strong> == $0
          </p>
        ▶<p class="subheader text-left">…</p>
        ▶<p class="subheader text-left">…</p>
        ▶<p class="subheader text-left">…</p>
        ▶<p class="subheader text-left">…</p>
        ▶<p class="subheader text-left">…</p>
        ▶<p class="subheader text-left">…</p>
        </div>
      </li>
```

Google Sheets **IMPORTXML** query for <div> element

```
=IMPORTXML(A1,"//div[@id='panel1d']")
```

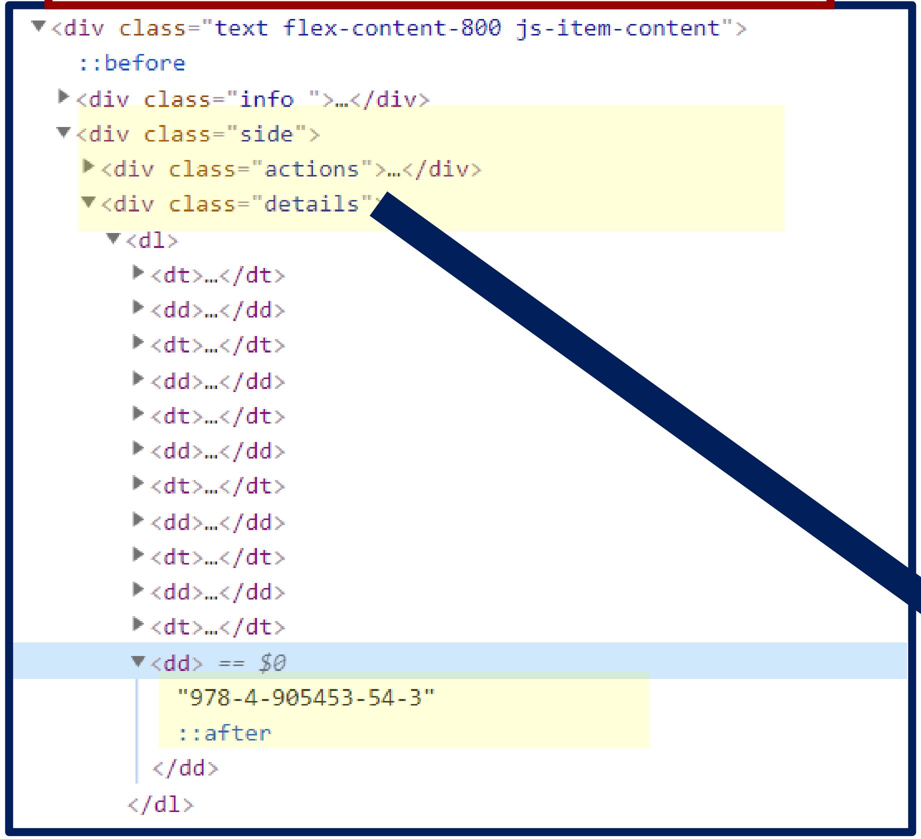
A	B	C	D	E	F	G	H
https://www.hindibook.com/index.php?p=sr&format=fullpage&Field=bookcode&String=9788178018539	Title: 1857 KI JUNG-E-AZADI KA GUMNAM SHAHEED RAJA NAHAR SINGH	Author: AMRITA SINGH DR.	ISBN 13: 9788178018539	ISBN 10: 8178018535	Year: 2017	Language: URDU	Pages etc.: 144p

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

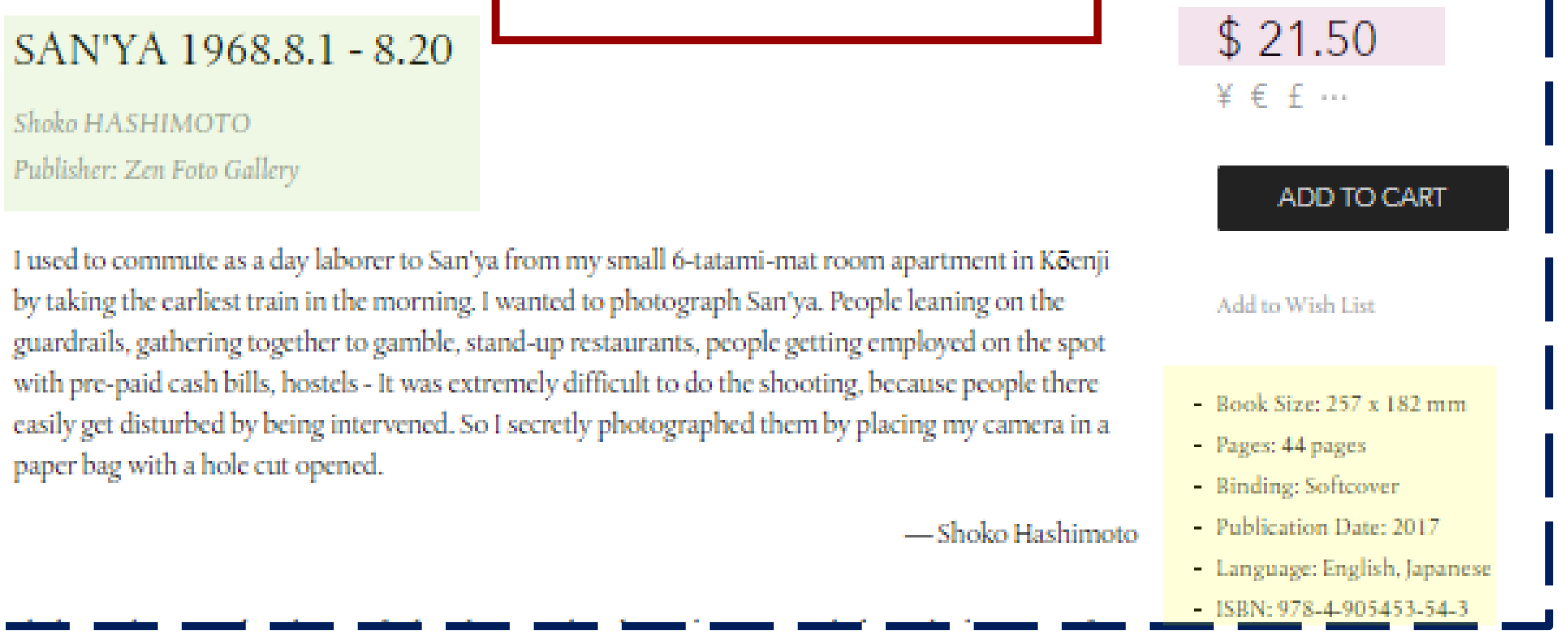
Scraping Semi-Formatted Data

HTML source of ISBN metadata



```
▼<div class="text flex-content-800 js-item-content">
   ::before
  ▶<div class="info ">…</div>
  ▼<div class="side">
    ▶<div class="actions">…</div>
    ▼<div class="details">
      ▼<dl>
        ▶<dt>…</dt>
        ▶<dd>…</dd>
        ▶<dt>…</dt>
        ▶<dd>…</dd>
        ▶<dt>…</dt>
        ▶<dd>…</dd>
        ▶<dt>…</dt>
        ▶<dd>…</dd>
        ▶<dt>…</dt>
        ▶<dd>…</dd>
        ▶<dt>…</dt>
        ▼<dd> == $0
           "978-4-905453-54-3"
           ::after
         </dd>
       </dl>
```

Individual book metadata



SAN'YA 1968.8.1 - 8.20

Shoko HASHIMOTO

Publisher: Zen Foto Gallery

I used to commute as a day laborer to San'ya from my small 6-tatami-mat room apartment in Kōenji by taking the earliest train in the morning. I wanted to photograph San'ya. People leaning on the guardrails, gathering together to gamble, stand-up restaurants, people getting employed on the spot with pre-paid cash bills, hostels - It was extremely difficult to do the shooting, because people there easily get disturbed by being intervened. So I secretly photographed them by placing my camera in a paper bag with a hole cut opened.

— Shoko Hashimoto

$ 21.50

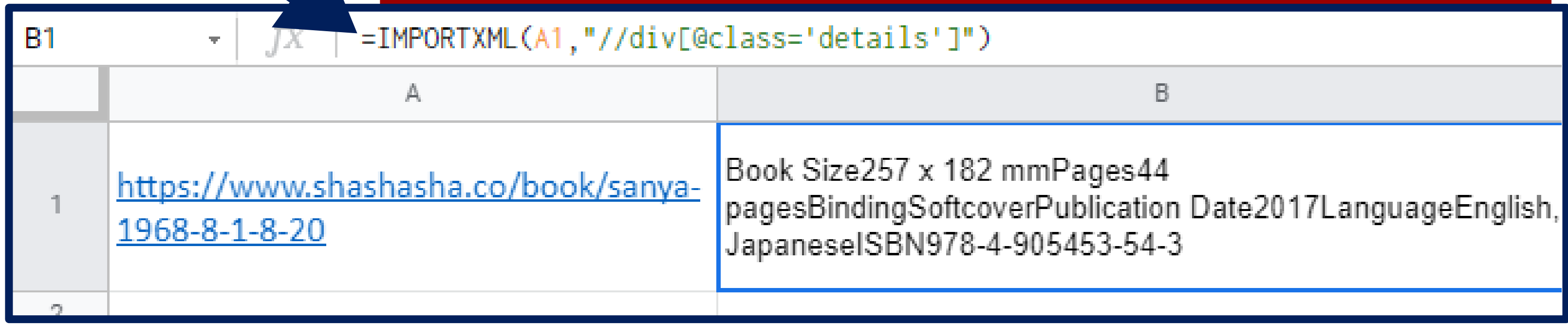
¥ € £ ···

ADD TO CART

Add to Wish List

- Book Size: 257 x 182 mm
- Pages: 44 pages
- Binding: Softcover
- Publication Date: 2017
- Language: English, Japanese
- ISBN: 978-4-905453-54-3

Google Sheets **IMPORTXML** query for <div> element

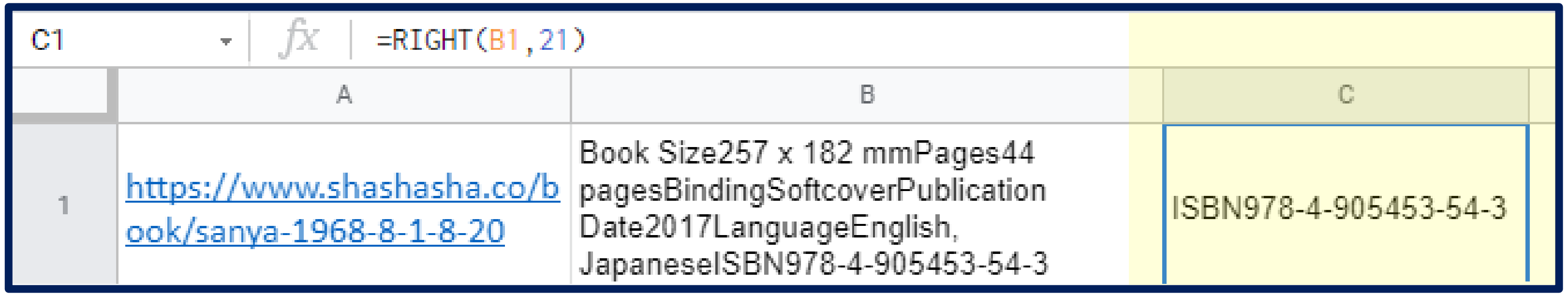


B1 fx =IMPORTXML(A1,"//div[@class='details']")

	A	B
1	https://www.shashasha.co/book/sanya-1968-8-1-8-20	Book Size257 x 182 mmPages44 pagesBindingSoftcoverPublication Date2017LanguageEnglish, JapaneseISBN978-4-905453-54-3
2		

Cleaning Up and Adding to Scraped Data

Google Sheets **RIGHT** formula to get ISBN

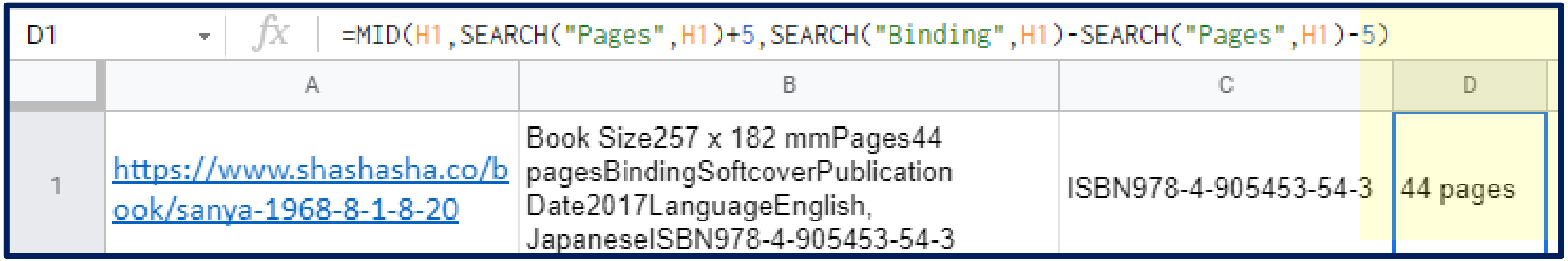


C1

```
=RIGHT(B1,21)
```

	A	B	C
1	https://www.shashasha.co/book/sanya-1968-8-1-8-20	Book Size257 x 182 mmPages44 pagesBindingSoftcoverPublication Date2017LanguageEnglish, JapaneseISBN978-4-905453-54-3	ISBN978-4-905453-54-3

Google Sheets **MID** formula to isolate pagination

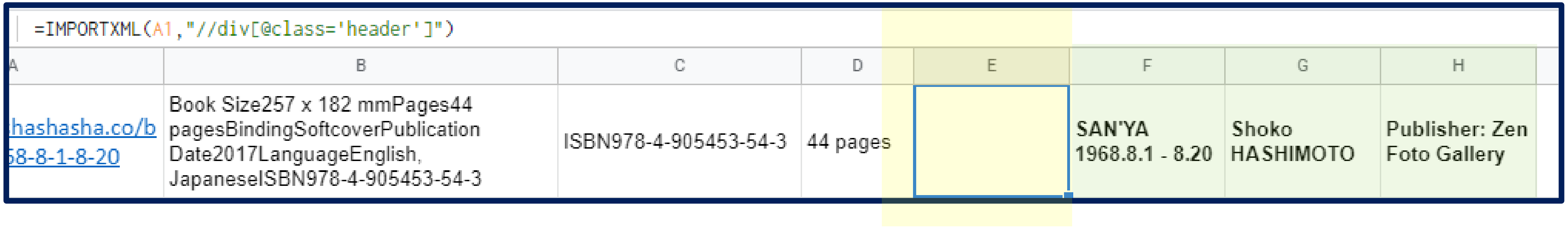


D1

```
=MID(H1,SEARCH("Pages",H1)+5,SEARCH("Binding",H1)-SEARCH("Pages",H1)-5)
```

	A	B	C	D
1	https://www.shashasha.co/book/sanya-1968-8-1-8-20	Book Size257 x 182 mmPages44 pagesBindingSoftcoverPublication Date2017LanguageEnglish, JapaneseISBN978-4-905453-54-3	ISBN978-4-905453-54-3	44 pages

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

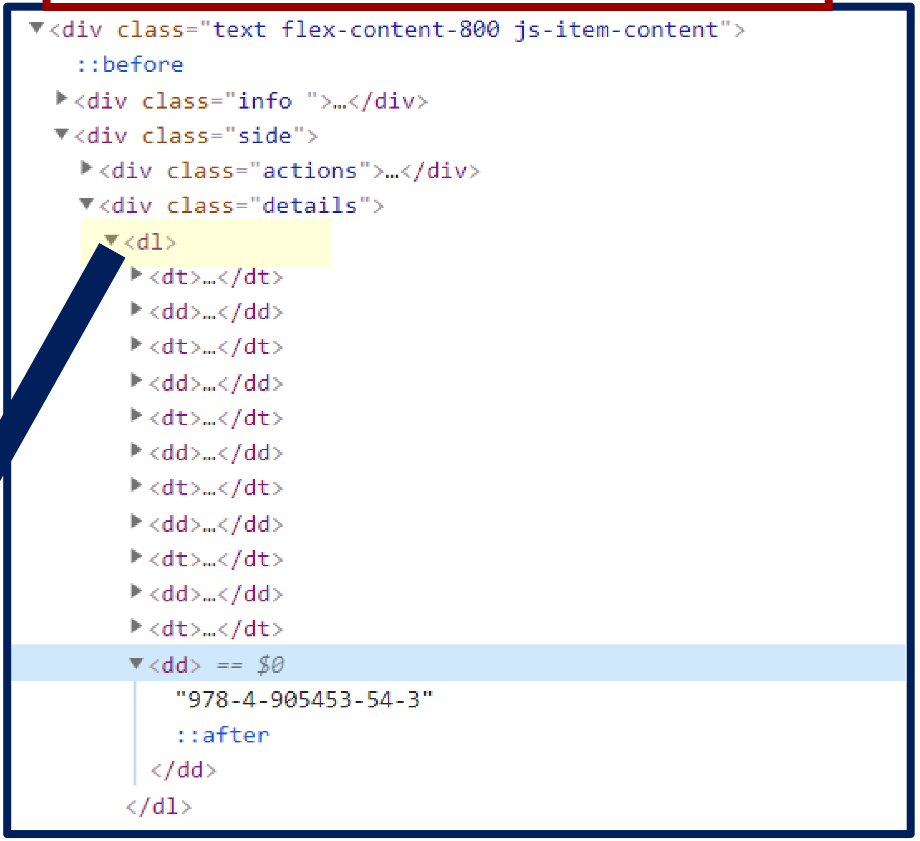


```
=IMPORTXML(A1,"//div[@class='header']")
```

A	B	C	D	E	F	G	H
hashasha.co/b 58-8-1-8-20	Book Size257 x 182 mmPages44 pagesBindingSoftcoverPublication Date2017LanguageEnglish, JapaneseISBN978-4-905453-54-3	ISBN978-4-905453-54-3	44 pages		SAN'YA 1968.8.1 - 8.20	Shoko HASHIMOTO	Publisher: Zen Foto Gallery

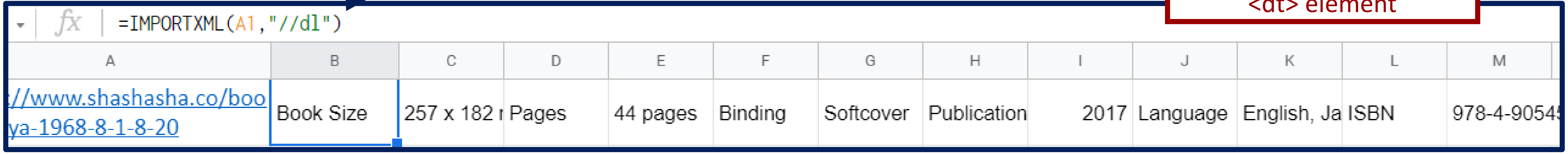
Further Experimentation (and Success!)

HTML source of ISBN metadata



```
▼<div class="text flex-content-800 js-item-content">
   ::before
  ▶<div class="info ">…</div>
  ▼<div class="side">
    ▶<div class="actions">…</div>
    ▼<div class="details">
      ▼<dl>
        ▶<dt>…</dt>
        ▶<dd>…</dd>
        ▶<dt>…</dt>
        ▶<dd>…</dd>
        ▶<dt>…</dt>
        ▶<dd>…</dd>
        ▶<dt>…</dt>
        ▶<dd>…</dd>
        ▶<dt>…</dt>
        ▶<dd>…</dd>
        ▶<dt>…</dt>
        ▼<dd> == $0
            "978-4-905453-54-3"
            ::after
          </dd>
        </dl>
```

Google Sheets **IMPORTXML** query for <dt> element



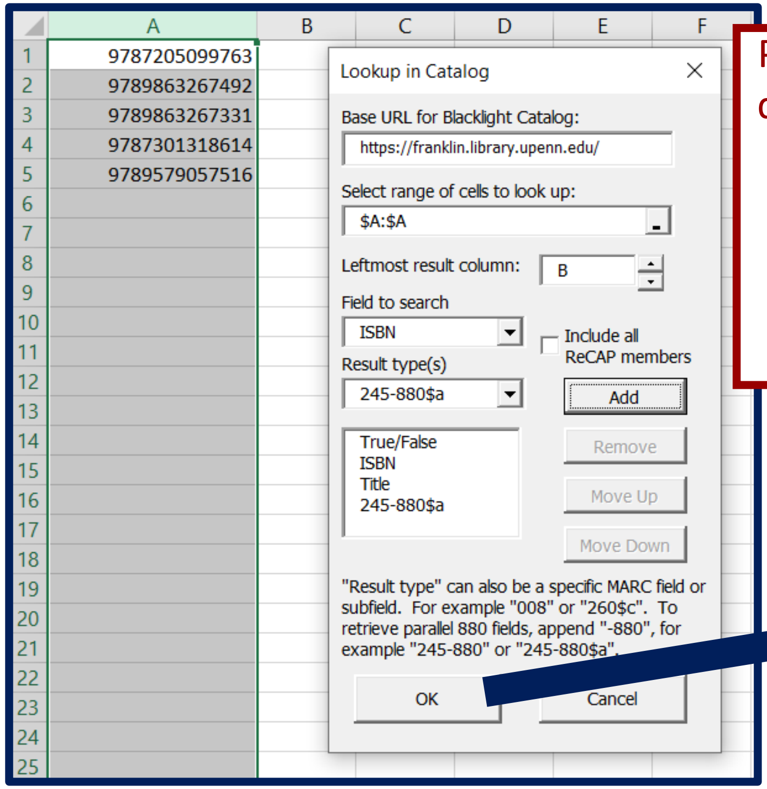
```
fx  =IMPORTXML(A1,"//dl")
```

A	B	C	D	E	F	G	H	I	J	K	L	M
://www.shashasha.co/boo ya-1968-8-1-8-20	Book Size	257 x 182 r	Pages	44 pages	Binding	Softcover	Publication	2017	Language	English, Ja	ISBN	978-4-9054

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



	A
1	9787205099763
2	9789863267492
3	9789863267331
4	9787301318614
5	9789579057516

Lookup in Catalog

Base URL for Blacklight Catalog: https://franklin.library.upenn.edu/

Select range of cells to look up: $A:$A

Leftmost result column: B

Field to search: ISBN

□ Include all ReCAP members

Result type(s): 245-880$a

Add

True/False
ISBN
Title
245-880$a

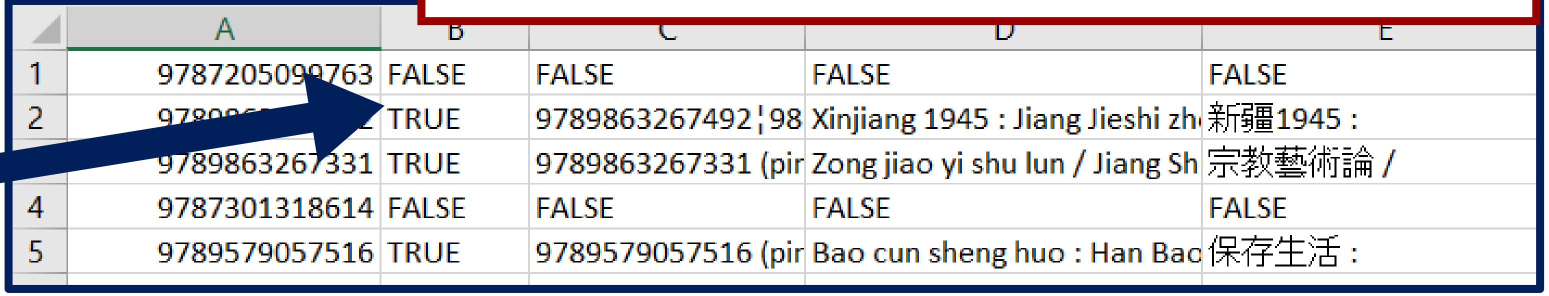
Remove | Move Up | Move Down

"Result type" can also be a specific MARC field or subfield. For example "008" or "260$c". To retrieve parallel 880 fields, append "-880", for example "245-880" or "245-880$a".

OK | Cancel

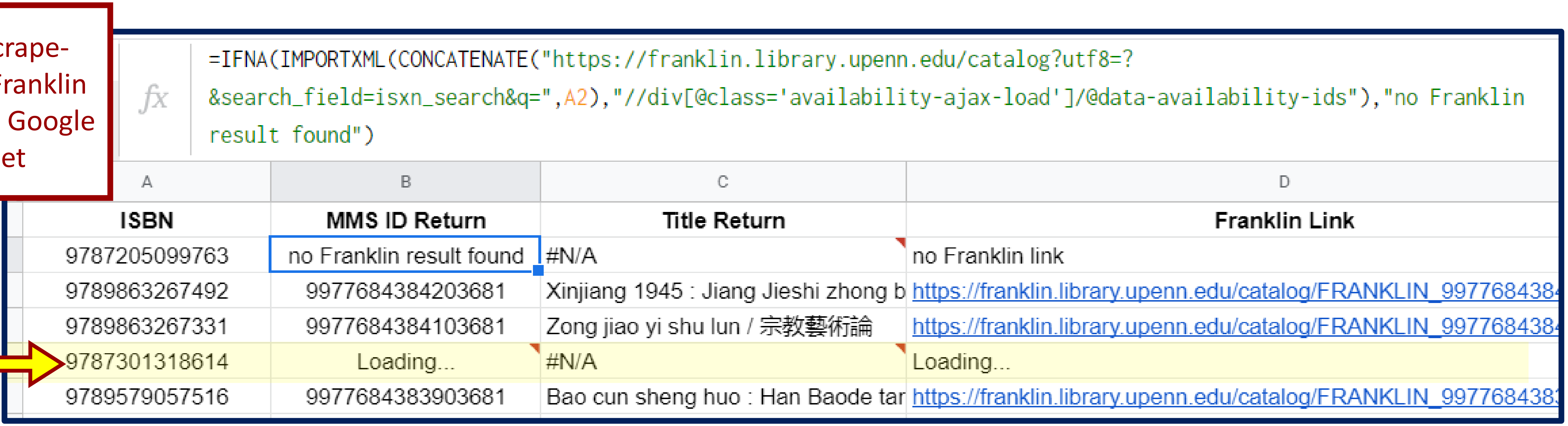
Princeton-developed "Look Up in Local Catalog" interface for Excel

Princeton-developed "Look Up in Local Catalog" Excel results



	A	B	C	D	E
1	9787205099763	FALSE	FALSE	FALSE	FALSE
2	9789863267492	TRUE	9789863267492¦98	Xinjiang 1945 : Jiang Jieshi zh	新疆1945 :
3	9789863267331	TRUE	9789863267331 (pir	Zong jiao yi shu lun / Jiang Sh	宗教藝術論 /
4	9787301318614	FALSE	FALSE	FALSE	FALSE
5	9789579057516	TRUE	9789579057516 (pir	Bao cun sheng huo : Han Bac	保存生活 :

Web-scrape-based "Franklin Checker" Google Sheet



```
=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")
```

A	B	C	D
ISBN	MMS ID Return	Title Return	Franklin Link
9787205099763	no Franklin result found	#N/A	no Franklin link
9789863267492	9977684384203681	Xinjiang 1945 : Jiang Jieshi zhong b	https://franklin.library.upenn.edu/catalog/FRANKLIN_997768438
9789863267331	9977684384103681	Zong jiao yi shu lun / 宗教藝術論	https://franklin.library.upenn.edu/catalog/FRANKLIN_997768438
9787301318614	Loading...	#N/A	Loading...
9789579057516	9977684383903681	Bao cun sheng huo : Han Baode tar	https://franklin.library.upenn.edu/catalog/FRANKLIN_997768438

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
    - January 2021 Core webinar: http://www.ala.org/core/using-matchmarc
- **MarcEdit**
    - Home page: https://marcedit.reeset.net/
    - Translating Delimited Files (Illinois Library LibGuide): https://guides.library.illinois.edu/c.php?g=463460&p=3168299
- **Google Sheets**
    - IMPORTHTML sample usage: https://support.google.com/docs/answer/3093339?hl=en
    - IMPORTXML sample usage: https://support.google.com/docs/answer/3093342?hl=en
    - XPATH syntax (W3Schools): https://www.w3schools.com/xml/xpath_syntax.asp
- **Microsoft Excel formulas**
    - Microsoft text functions reference: https://support.microsoft.com/en-us/office/text-functions-reference-cccd86ad-547d-4ea9-a065-7bb697c2a56e
    - Extract text from strings (Ablebits): https://www.ablebits.com/office-addins-blog/2017/11/15/excel-substring-functions-extract-text/

Contact Us!

Michael P. Williams, *Head, Global Studies Technical Services*

mpw2@upenn.edu

Beth Picknally Camden, *Patricia and Bernard Goldstein Director of Information Processing*

bethpc@upenn.edu

Penn Libraries
UNIVERSITY *of* PENNSYLVANIA