Enhancing Serials Metadata with Artificial Intelligence: A Case Study Using ChatGPT and Alma Cloud Apps

Marlene van Ballegooie & Juliya Borie

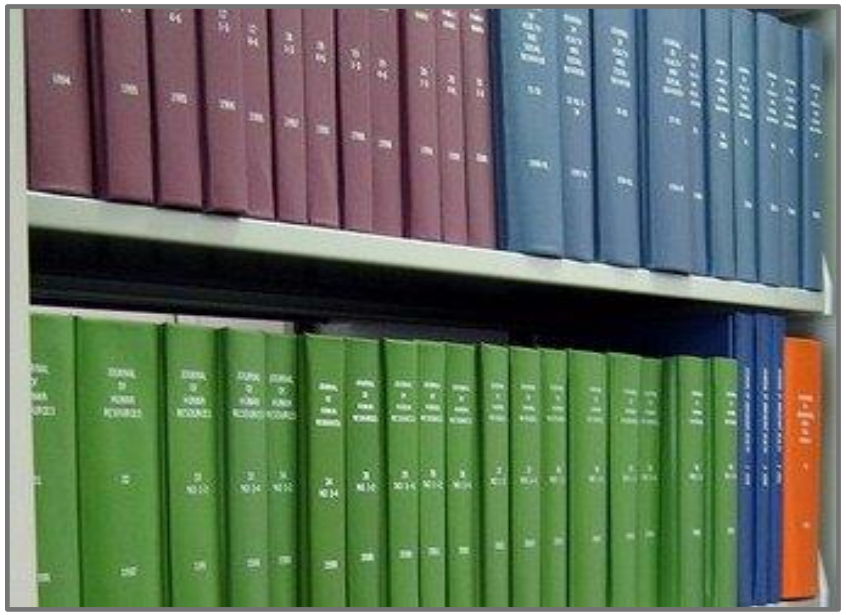
IGeLU 2024 - Copenhagen

Background

- University of Toronto Libraries is one of the largest academic library systems in North America
- In 2005, the University of Toronto established Downsview, an environmentally controlled high-density storage and preservation facility.
- Downsview plays a crucial role in alleviating the space constraints faced by on-campus libraries

The Downsview Serials Collection

- Over the years, libraries from across UTL and our Keep@Downsview shared print partnership have contributed serials to the Downsview offsite storage facility
- Each serial volume must be itemized to be included in the Downsview facility
- Metadata work required for serial volumes is often done in time-limited projects and with the assistance of student employees



Variations in Serial Item Descriptions

- MANY variations in the ways the Description field has been captured over time.
- Discrepancies in the serial metadata has caused some difficulty and confusion among researchers

1946 9/12 (sept-dec)	FALL 1969
163 3869-3874	81 NOS.1-3
BD.184-185	V.5, no. 1, July 1899
no.001-072	v.69 PT. 1 (1986:JAN./JUNE)
v.3:APR.SEPT./OCT.DEC.(1954)	v.0:no.0(1988:Aug.)

Impact of Metadata Quality on User Experience

- Research conducted in 2019 to assess library user's experience in navigating serial metadata
- Study revealed that item records for serial volumes often caused confusion, particularly when the year was lacking in the description, and when unbound serials were bundled together to simplify processing tasks
- Summary holdings statements were also problematic for users, as they are difficult to follow and often not accurate

Addressing Metadata Quality With AI

- With the explosion of generative AI, new possibilities emerge to use artificial intelligence in our work
- Explored ChatGPT for its data analysis and programming capabilities.
- Idea was to use ChatGPT to create simple utilities to enhance and standardize serial item metadata
- Developed a series of five related projects to improve the representation of Downsview serials in Alma/Primo

Project 1: Generating Serial Item Metadata

- Created a series of prompts to analyze the serial item description, extract data elements for the Enumeration and Chronology fields, and rebuild the item description according to a standardized format.

Description	Enum A	Enum B	Chron I	Chron J	New_Description
v.47 no.3/4 Fall/Winter 2017	47	3-4	2017	23-24	v.47:no.3-4(2017:fall-winter)

- Need to be very specific in what you are looking for and you need to iterate repeatedly to capture all data elements. No single prompt can account for every data element.

Sample Prompt: Extract Volume Information

This spreadsheet contains data related to serial publications.

Step 1 - Can you look at the Description field and find a list of all serial items that start with "v." or "V." or "v" or "V" or "vol." or "Vol." and then have a number or a range of numbers that are separated by a dash "-" or a slash "/".

Step 2 - After finding these items, I need you to take the number, or the range of numbers, and copy the number to the Enum A field.

Step 3 - For each of these items, add the designation "v." to the field named "Enum A Unit".

Step 4 - In this analysis it is very important to save all fields as TEXT. If there is no value, do not report "nan", instead just leave the field blank.

Step 5. Save the output to a new spreadsheet named "Update_Serial_Items_Task_1.xlsx

- For each main task, we broke it down into a series of steps for ChatGPT to follow
- Once perfected, ChatGPT created a Python script so that we could replicate the task in another environment
- Use ChatGPT to turn conversations into software utilities

Data Extraction Tasks

- Extract volume information
- Extract issue numbers, parts and pages (no., pt., p.)
- Extract years and ranges of years
- Extract months and seasons and convert to numeric
- Extract supplements, indexes, and "lacking" information
- Adjust enumeration fields for items lacking "volumes" - i.e. move continuously numbered issues to Enum A

Jupyter Notebooks

- Jupyter notebooks work with code blocks, called cells, which can be executed independently to run code
- The programming software is used to process the files sequentially – saving the result set at each step of the way
- Each task builds on the next

jupyter Extract_Serial_Item_Information Last Checkpoint: 2024-07-19 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

In [ ]:
```python
#Task 1 - Extract Volume information

import pandas as pd
import re

def process_serial_items(file_path, output_path):
    # Load the data
    data = pd.read_excel(file_path)

    # Define the regex pattern for volume and number extraction
    volume_pattern = re.compile(r'\b[vV][ol.]*\s*(\d+([/-]\d+)?)')

    # Extract the volume number or range and update Enum A
    data['Enum A'] = data['Description'].apply(lambda x: volume_pattern.search(str(x)).group(1) if volume_pattern.se

    # Convert all data to string, ensuring empty fields are blank instead of 'nan'
    data = data.astype(str).replace('nan', '')

    # Save the updated dataframe to a new Excel file
    data.to_excel(output_path, index=False)

    return f"Data processed and saved to {output_path}"

# Example usage
if __name__ == "__main__":
    file_path = 'files/Downsview_No_Enum_Chron.xlsx'  # Specify the path to your input file
    output_path = 'files/Update_Serial_Items_Task_1.xlsx'  # Specify the path for the output file
    print(process_serial_items(file_path, output_path))
```

In [ ]:
```python
#Task 2 - Extract dates from descriptions

import pandas as pd
import re
```

Results from scripts produced by ChatGPT

Before

MMS Id	Barcode	Description	Enum A	Enum B	Chron I	Chron J
991105980409106196	39212416060032	v.70 PT.2 (1987:JULY/DEC.)				
991105980409106196	39212416060024	v.70 PT.2 (1987:JAN./JUNE)				
991106140709506196	39250822110140	V.12 no.3 (1989 Fall)				
991106334826006196	39292503090100	v.45:no.3(2017 Jun)				
991106334826006196	39292503090118	v.45:no.2(2017 Apr)				
991106334826006196	39292503090126	v.45:no.1(2017 Feb)				
991106450458006196	39111621100049	V.5, no. 6, May 1900				
991106450458006196	39111621100130	V.5, no. 5, March 1900				
991106464006106196	39260521120068	V.11:no.1 (1972:MAR.)				
991106476076606196	39270806110015	v.5 Nos.4 (Dec. 1986)				
991106640102006196	39300412030032	v.71 no.13-26 (JUNE-DEC. 1958)				
991106640102006196	39300412030016	v.71 no.1-12 (JAN.-JUNE 1958)				
991106640102006196	39300412030024	v.69 no.13-26 (JUNE-DEC. 1956)				

After

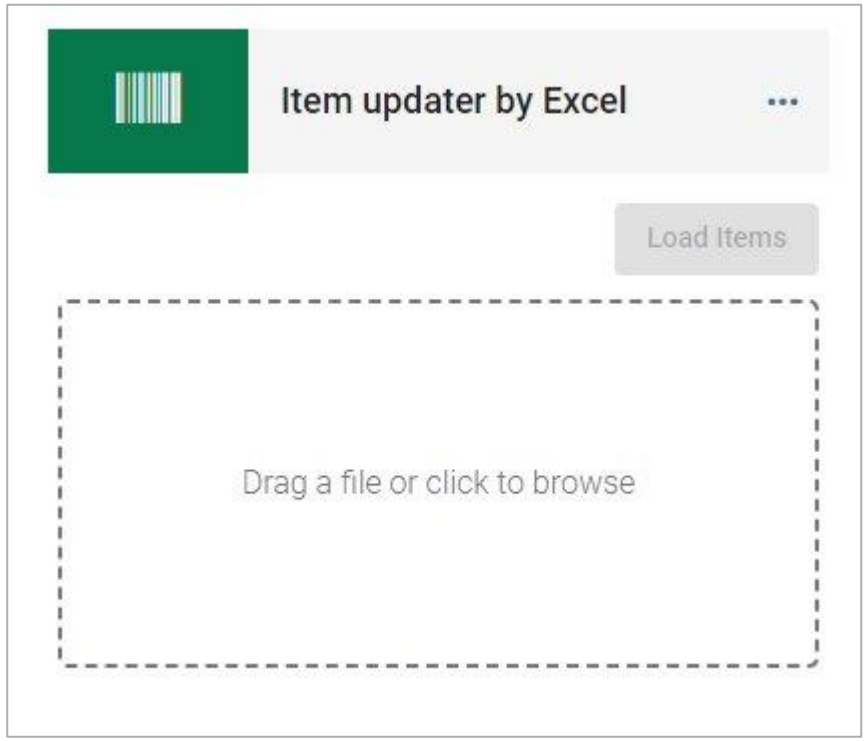
MMS Id	Barcode	Description	Enum A	Enum B	Chron I	Chron J
991105980409106196	39212416060032	v.70:pt.2(1987:Jul.-Dec.)	70	2	1987	07-12
991105980409106196	39212416060024	v.70:pt.2(1987:Jan.-Jun.)	70	2	1987	01-06
991106140709506196	39250822110140	v.12:no.3(1989:fall)	12	3	1989	23
991106334826006196	39292503090100	v.45:no.3(2017:Jun.)	45	3	2017	06
991106334826006196	39292503090118	v.45:no.2(2017:Apr.)	45	2	2017	04
991106334826006196	39292503090126	v.45:no.1(2017:Feb.)	45	1	2017	02
991106450458006196	39111621100049	v.5:no.6(1900:May)	5	6	1900	05
991106450458006196	39111621100130	v.5:no.5(1900:Mar.)	5	5	1900	03
991106464006106196	39260521120068	v.11:no.1(1972:Mar.)	11	1	1972	03
991106476076606196	39270806110015	v.5:no.4(1986:Dec.)	5	4	1986	12
991106640102006196	39300412030032	v.71:no.13-26(1958:Jun.-Dec.)	71	13-26	1958	06-12
991106640102006196	39300412030016	v.71:no.1-12(1958:Jan.-Jun.)	71	1-12	1958	01-06
991106640102006196	39300412030024	v.69:no.13-26(1956:Jun.-Dec.)	69	13-26	1956	06-12

Data Checking

- Revised item data is split into Excel spreadsheets based on specific patterns
- This assists with the necessary data checking process – rather than making your eyes concentrate on ALL patterns, we group items into similar patterns for better focus on the details
- Since Downsview serials are not browsable on the shelves, extra care must be taken to ensure the data is correct
- Data checking is the most significant bottleneck in the whole process

Reload into Alma with Item Updater by Excel

- Once items are verified, data is reformatted with column headers required by the Item Updater by Excel tool
- A script is used to split the file into 2500 row chunks for easy loading into Alma
- Very slow loading, but effective
- To date: 250,000 items updated!

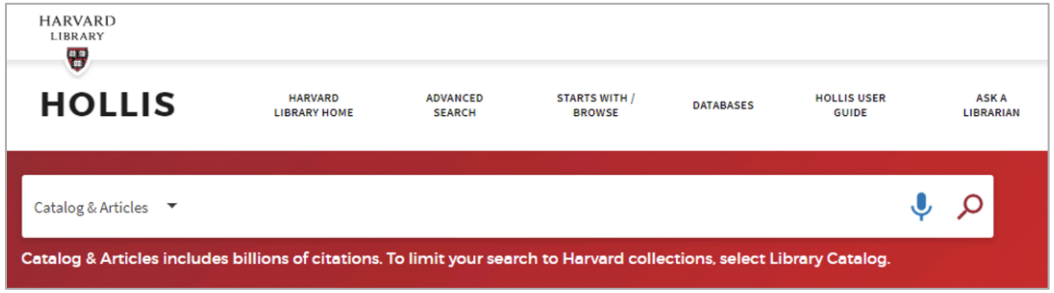


Project 2: Enhancing Serial Items with External Data

- The idea was to enhance serial item records with issue dates obtained from JSTOR
- JSTOR provides fairly standardized metadata related to serial coverage
- ChatGPT can assist in extracting issue level data from text files and merging dates into local metadata

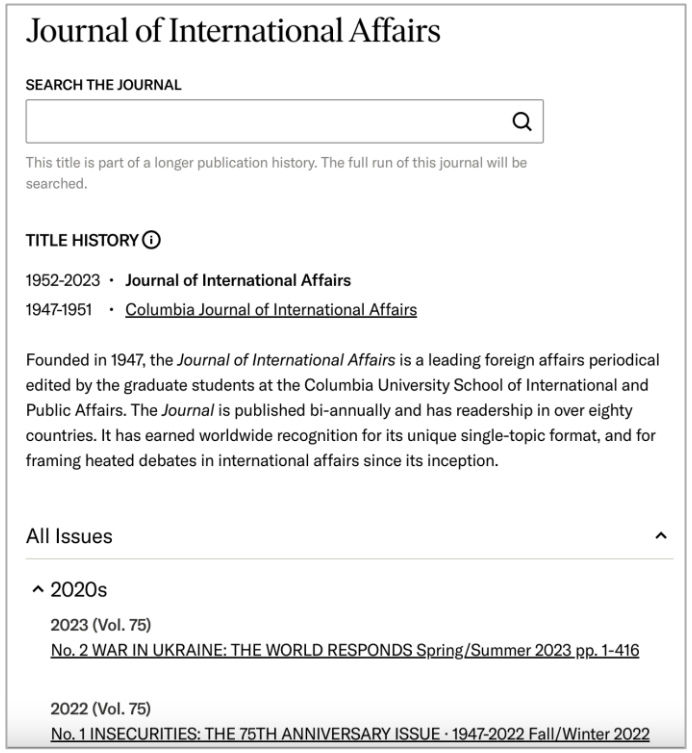
JSTOR

国立国会図書館
National Diet Library, Japan



Obtain Data from JSTOR

Copy and Paste



Journal of International Affairs

SEARCH THE JOURNAL

This title is part of a longer publication history. The full run of this journal will be searched.

TITLE HISTORY

1952-2023 • **Journal of International Affairs**

1947-1951 • Columbia Journal of International Affairs

Founded in 1947, the *Journal of International Affairs* is a leading foreign affairs periodical edited by the graduate students at the Columbia University School of International and Public Affairs. The *Journal* is published bi-annually and has readership in over eighty countries. It has earned worldwide recognition for its unique single-topic format, and for framing heated debates in international affairs since its inception.

All Issues

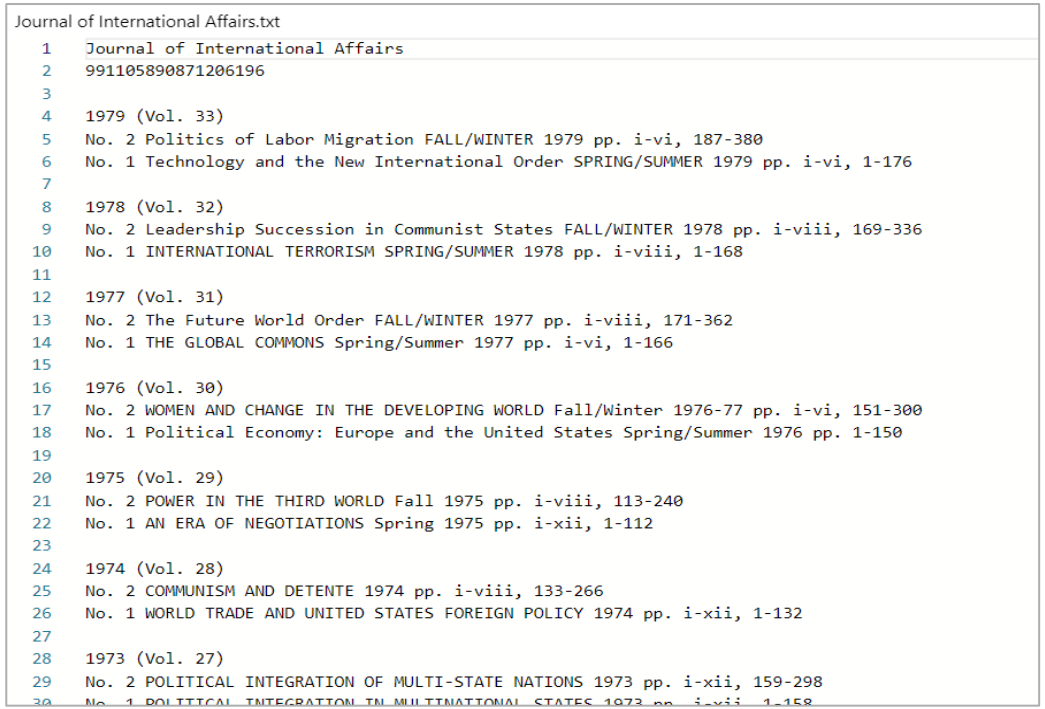
2020s

2023 (Vol. 75)

No. 2 WAR IN UKRAINE: THE WORLD RESPONDS Spring/Summer 2023 pp. 1-416

2022 (Vol. 75)

No. 1 INSECURITIES: THE 75TH ANNIVERSARY ISSUE · 1947-2022 Fall/Winter 2022

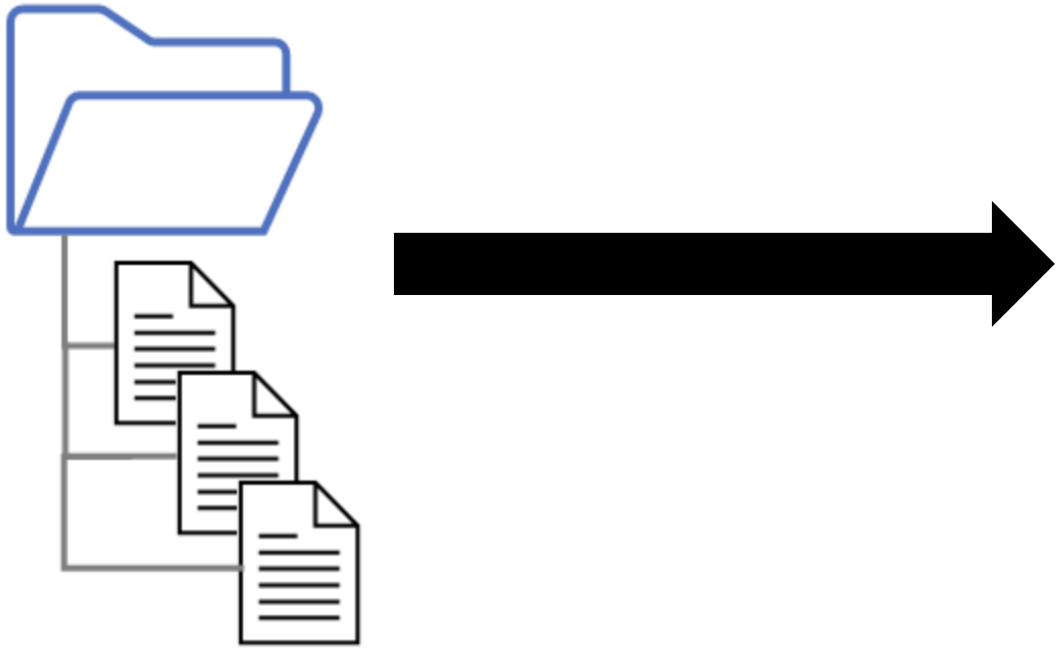


Journal of International Affairs.txt

```
1   Journal of International Affairs
2   991105890871206196
3
4   1979 (Vol. 33)
5   No. 2 Politics of Labor Migration FALL/WINTER 1979 pp. i-vi, 187-380
6   No. 1 Technology and the New International Order SPRING/SUMMER 1979 pp. i-vi, 1-176
7
8   1978 (Vol. 32)
9   No. 2 Leadership Succession in Communist States FALL/WINTER 1978 pp. i-viii, 169-336
10  No. 1 INTERNATIONAL TERRORISM SPRING/SUMMER 1978 pp. i-viii, 1-168
11
12  1977 (Vol. 31)
13  No. 2 The Future World Order FALL/WINTER 1977 pp. i-viii, 171-362
14  No. 1 THE GLOBAL COMMONS Spring/Summer 1977 pp. i-vi, 1-166
15
16  1976 (Vol. 30)
17  No. 2 WOMEN AND CHANGE IN THE DEVELOPING WORLD Fall/Winter 1976-77 pp. i-vi, 151-300
18  No. 1 Political Economy: Europe and the United States Spring/Summer 1976 pp. 1-150
19
20  1975 (Vol. 29)
21  No. 2 POWER IN THE THIRD WORLD Fall 1975 pp. i-viii, 113-240
22  No. 1 AN ERA OF NEGOTIATIONS Spring 1975 pp. i-xii, 1-112
23
24  1974 (Vol. 28)
25  No. 2 COMMUNISM AND DETENTE 1974 pp. i-viii, 133-266
26  No. 1 WORLD TRADE AND UNITED STATES FOREIGN POLICY 1974 pp. i-xii, 1-132
27
28  1973 (Vol. 27)
29  No. 2 POLITICAL INTEGRATION OF MULTI-STATE NATIONS 1973 pp. i-xii, 159-298
30  No. 1 POLITICAL INTEGRATION IN MULTINATIONAL STATES 1973 pp. i-xii, 1-158
```

Extract Issue Level Data

- ChatGPT helped to create a Python script that reads multiple text files containing issue level metadata and extract the MMS Id, title, issue and date into a single Excel file.



MMS_ID	TITLE	ENUM_A	CHRON_I
991106372666206196	Social Problems	38	1991
991106372666206196	Social Problems	37	1990
991106372666206196	Social Problems	36	1989
991106372666206196	Social Problems	35	1988
991106372666206196	Social Problems	34	1987
991106372666206196	Social Problems	33	1986
991106372666206196	Social Problems	33	1985
991106372666206196	Social Problems	32	1985
991106372666206196	Social Problems	32	1984
991106372666206196	Social Problems	31	1984
991106372666206196	Social Problems	31	1983
991106372666206196	Social Problems	30	1983
991106372666206196	Social Problems	30	1982
991106372666206196	Social Problems	29	1982
991106372666206196	Social Problems	29	1981
991106372666206196	Social Problems	28	1981
991106372666206196	Social Problems	28	1980
991106372666206196	Social Problems	27	1980
991106372666206196	Social Problems	27	1979
991106372666206196	Social Problems	26	1979
991106372666206196	Social Problems	26	1978
991106372666206196	Social Problems	25	1978
991106372666206196	Social Problems	25	1977
991106372666206196	Social Problems	24	1977
991106372666206196	Social Problems	24	1976
991106372666206196	Social Problems	23	1976
991106372666206196	Social Problems	23	1975
991106372666206196	Social Problems	22	1975
991106372666206196	Social Problems	22	1974

Merge Dates into Item Metadata

- ChatGPT helped to create a second Python script that matches the JSTOR data and the local data on the MMS Id and volume number. Where there is a match, dates are copied into the Chron I field.
- Script accounts for bundled issues by returning a range of dates.
- With dates added to Chron I, this enables the Description field to updated with dates

Description	description	enumeration_a	enumeration_b	chronology_i
v.5	v.5(1961)	5		1961
v.3-4	v.3-4(1959-1960)	3-4		1959-1960
v.1-2	v.1-2(1957-1958)	1-2		1957-1958
v.12	v.12(1981)	12		1981
v.11	v.11(1980)	11		1980
v.10	v.10(1979)	10		1979
v.9	v.9(1978)	9		1978
v.8	v.8(1977)	8		1977
v.7	v.7(1976)	7		1976
v.6	v.6(1975)	6		1975
v.16	v.16(1995)	16		1995
v.17	v.17(1996)	17		1996
v.18	v.18(1997)	18		1997
v.19	v.19(1998)	19		1998
v.20	v.20(1999)	20		1999
v.22	v.22(2001)	22		2001

Project 3: Retrospective Itemization of Serials

- ChatGPT is also being used to facilitate the process of retrospective itemization
- Idea is to collect serial item information in an Excel spreadsheet and use ChatGPT to convert the data into MARCXML for loading into Alma

Barcode	Call Number	Description	Volume	Year
31761118754985	PG8501 .L52	v.88(2023)	88	2023
31761118754969	PG8501 .L52	v.87(2022)	87	2022
31761118754977	PG8501 .L52	v.86(2022)	86	2022
31761118643501	PG8501 .L52	v.85(2021)	85	2021
31761118643394	PG8501 .L52	v.84(2021)	84	2021
31761118643287	PG8501 .L52	v.83(2020)	83	2020
31761112191754	PG8501 .L52	v.82(2020)	82	2020
31761112170626	PG8501 .L52	v.81(2019)	81	2019

Collecting Serial Item Metadata

- Created a template to support itemization projects
- Predictive functionality in Excel + generating Description using AI

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	MMS Id	Title	Library Code	Location Code	Call Number	Material Typ	Barcode	Description	Enum A	Enum B	Chron I	Chron J	arrival_date	
2	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517677	v.41(1999)	41		1999		2024-07-18Z	
3	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517669	v.42(2000)	42		2000		2024-07-18Z	
4	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517651	v.43(2000)	43		2000		2024-07-18Z	
5	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517586	v.44(2001)	44		2001		2024-07-18Z	
6	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517644	v.45(2001)	45		2001		2024-07-18Z	
7	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517636	v.48(2003)	48		2003		2024-07-18Z	
8	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517628	v.49(2003)	49		2003		2024-07-18Z	
9	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517610	v.50(2004)	50		2004		2024-07-18Z	
10	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517602	v.51(2004)	51		2004		2024-07-18Z	
11	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517594	v.52(2005)	52		2005		2024-07-18Z	
12	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517578	v.53(2005)	53		2005		2024-07-18Z	
13	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517560	v.54(2006)	54		2006		2024-07-18Z	
14	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517552	v.55(2006)	55		2006		2024-07-18Z	
15	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517545	v.56(2007)	56		2007		2024-07-18Z	
16	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517537	v.57(2007)	57		2007		2024-07-18Z	
17	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517529	v.58(2008)	58		2008		2024-07-18Z	
18	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517511	v.59(2008)	59		2008		2024-07-18Z	
19	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517503	v.60(2009)	60		2009		2024-07-18Z	
20	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517495	v.61(2009)	61		2009		2024-07-18Z	
21	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517487	v.62-63(2010)	62-63		2010		2024-07-18Z	
22	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517479	v.64-65(2011)	64-65		2011		2024-07-18Z	
23	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517420	v.66(2012)	66		2012		2024-07-18Z	
24	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517412	v.67(2012)	67		2012		2024-07-18Z	
25	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517404	v.68(2013)	68		2013		2024-07-18Z	
26	991106472301106196	Acta linguistica Lithuanica.	ROBARTS	STACKS	PG8501 .L52	ISSUE	31761120517396	v.69(2013)	69		2013		2024-07-18Z	

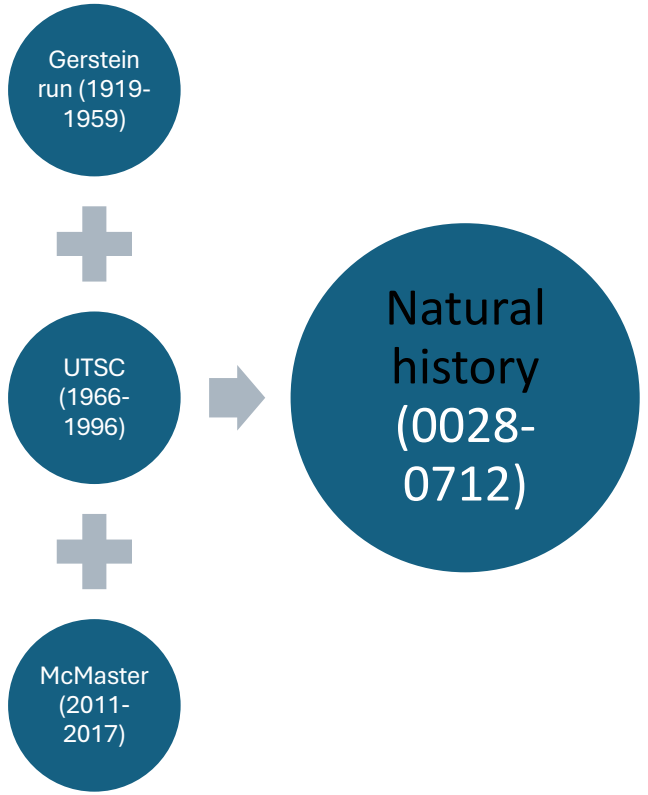
Converting Data to MARCXML

- Used ChatGPT to create a Python script that iterates through the spreadsheet and writes the data into MARCXML format
- Item records can then be attached to existing bibliographic records in Alma
- Potential for time saving and less mistakes

```
<?xml version="1.0" encoding="UTF-8" ?>
    <marc:collection xmlns:marc="http://www.loc.gov/MARC21/slim" xmlns:xsi="http://www.w
    schema/MARC21slim.xsd">
    <marc:record>
    <marc:leader>00190nam a2200061Ia 45e0</marc:leader>
    <marc:controlfield tag="001">991106472301106196</marc:controlfield>
    <marc:controlfield tag="008">230821s9999  xx      000 0 und d</marc:controlfield>
    <marc:datafield tag="949" ind1=" " ind2=" ">
    <marc:subfield code="a">PG8501 .L52</marc:subfield>
    <marc:subfield code="i">31761120517677</marc:subfield>
    <marc:subfield code="l">STACKS</marc:subfield>
    <marc:subfield code="m">ROBARTS</marc:subfield>
    <marc:subfield code="f">ISSUE</marc:subfield>
    <marc:subfield code="n">v.41(1999)</marc:subfield>
    <marc:subfield code="o">41</marc:subfield>
    <marc:subfield code="p"></marc:subfield>
    <marc:subfield code="q">1999</marc:subfield>
    <marc:subfield code="r"></marc:subfield>
    <marc:subfield code="s">20240718</marc:subfield>
    </marc:datafield>
    </marc:record>
<marc:record>
    <marc:leader>00190nam a2200061Ia 45e0</marc:leader>
    <marc:controlfield tag="001">991106472301106196</marc:controlfield>
    <marc:controlfield tag="008">230821s9999  xx      000 0 und d</marc:controlfield>
    <marc:datafield tag="949" ind1=" " ind2=" ">
    <marc:subfield code="a">PG8501 .L52</marc:subfield>
    <marc:subfield code="i">31761120517669</marc:subfield>
    <marc:subfield code="l">STACKS</marc:subfield>
    <marc:subfield code="m">ROBARTS</marc:subfield>
    <marc:subfield code="f">ISSUE</marc:subfield>
    <marc:subfield code="n">v.42(2000)</marc:subfield>
    <marc:subfield code="o">42</marc:subfield>
    <marc:subfield code="p"></marc:subfield>
    <marc:subfield code="q">2000</marc:subfield>
    <marc:subfield code="r"></marc:subfield>
    <marc:subfield code="s">20240718</marc:subfield>
    </marc:datafield>
    </marc:record>
```

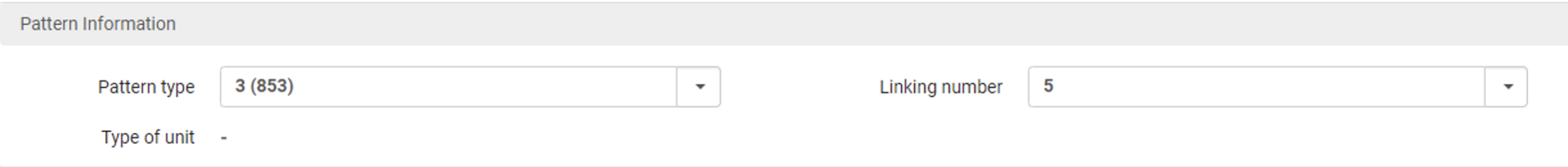
Project 4: Regenerating Summary Holdings

- Historically, libraries transferred "items" to Downsview, little attention was paid to ensuring summary holdings remained up to date
- "Holdings drift" more apparent in Alma
- As different libraries contribute to Downsview, legacy holdings no longer reflect the journal run



Creating Summary Holdings

- Prediction patterns in holding records are necessary for generating summary holdings
- Can summary holdings be created retrospectively?
- Yes! Add a valid prediction pattern to the holdings record (can be done using a normalization rule).
- Add pattern type and linking number to the items – update using the Item Updater by Excel tool. Use Alma "Change Holdings Information" job to force a save action.



Project 5: Recording Gaps in Serial Holdings

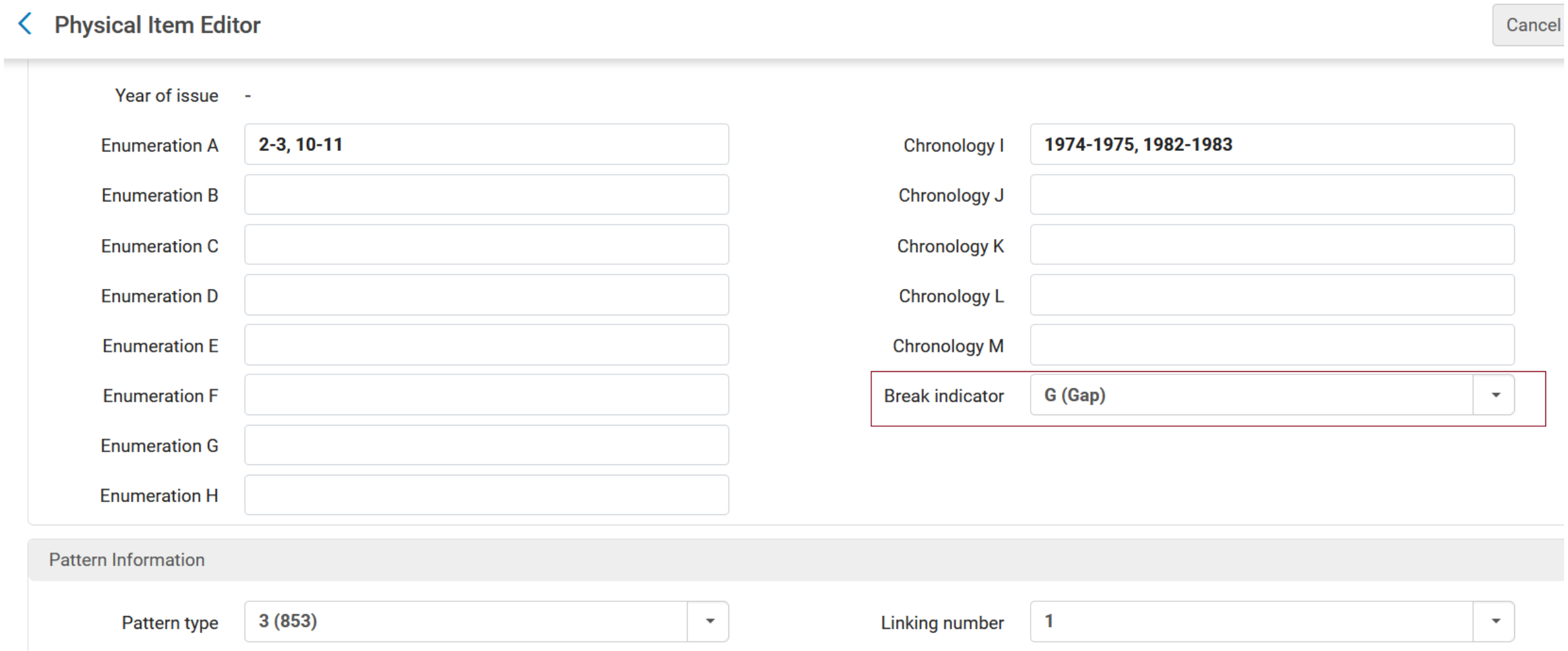
- NISO/ANSI standard specifies to record and display data in a positive sense; that is, emphasize that which is held rather than that which is not.
- Specificity in indicating gaps – **Break indicator**

UTL at Downsview
Available , Item may be requested ; HB1 .B764

Holdings:

1989-1992
1993:no.1
1993:no.2
1994-1997
2011

Break Indicator in ENUM/CHRON



Challenges and Opportunities

- Challenges
  - Item data verification is very time consuming and mundane work. Still looking for accurate AI solutions!
  - Itemization granularity varies – we can only work with what is provided
  - Determining the top level of enumeration and chronology without visual cues from the serial title page.
- Opportunities
  - Better standardization of metadata = better user experience
  - Multilingual captions to better represent the items
  - Working with ChatGPT exercises our computational thinking abilities – a key skill in working with AI in the future

Next Steps

- Investigate more automated methods to review results of transformations
- Full-scale automation of holdings summary regeneration, paying particular attention to gaps in holdings
- Use artificial intelligence to assess the quality of Downsview serial bibliographic records

Related Work

van Ballegooie, Marlene, and Juliya Borie. “Inside-Out and Outside-In: A Holistic Approach to Metadata Assessment for an Off-Site Storage Collection.” *The Serials Librarian*, vol. 78, no. 1–4, 2020, pp. 86–92, https://doi.org/10.1080/0361526X.2020.1697600.

Marlene van Ballegooie
m.vanballegooie@utoronto.ca

Juliya Borie
juliya.borie@utoronto.ca