Processing a Mongolian Books Collection Using MatchMarc and Annif

Enhancing Acquisition and Cataloging Efficiency and Discoverability with AI Technology

Erica Lu 吕轶堃
Head of Global Studies Technical Services
Van Pelt Libraries, University of Pennsylvania

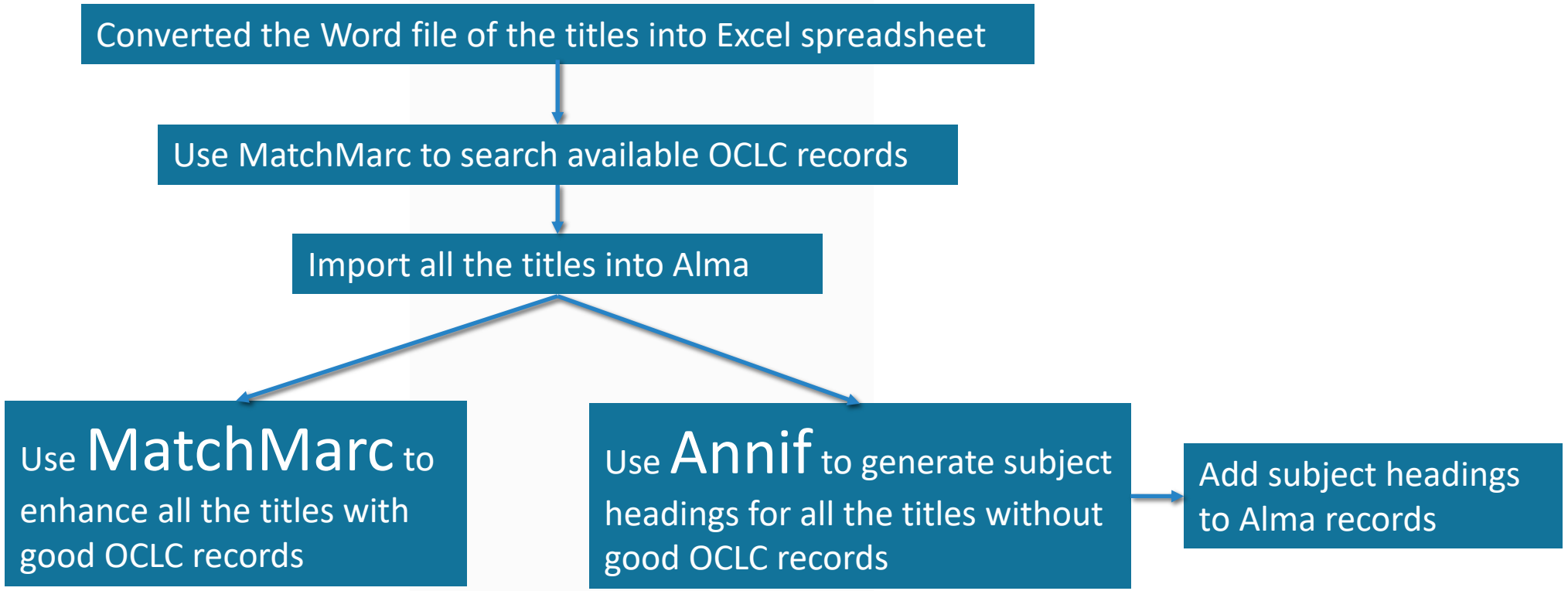
Penn Libraries
University of Pennsylvania

*Certain illustrations featured in this presentation were generated with the assistance of ChatGPT's image creation capabilities.

Introduction

- Collected during fieldwork in Mongolia (early 1990s).
- ~150 books on land degradation in Xilingol League (锡林格勒盟).
- Focus: Mongolian herding practices, ecological grasslands, and nomadic lifestyles. Includes unique books of its kind in North America.
- Languages: Mongolian traditional script, Mongolian Cyrillic, and Chinese.
- Strengthening one of the few North American Mongolian collections.

Project Workflow



- MatchMARC is a Google Apps Script published as a Google Sheets Addon. It is a tool that automates MARC record searches. It uses OCLC's WorldCat Search API. *

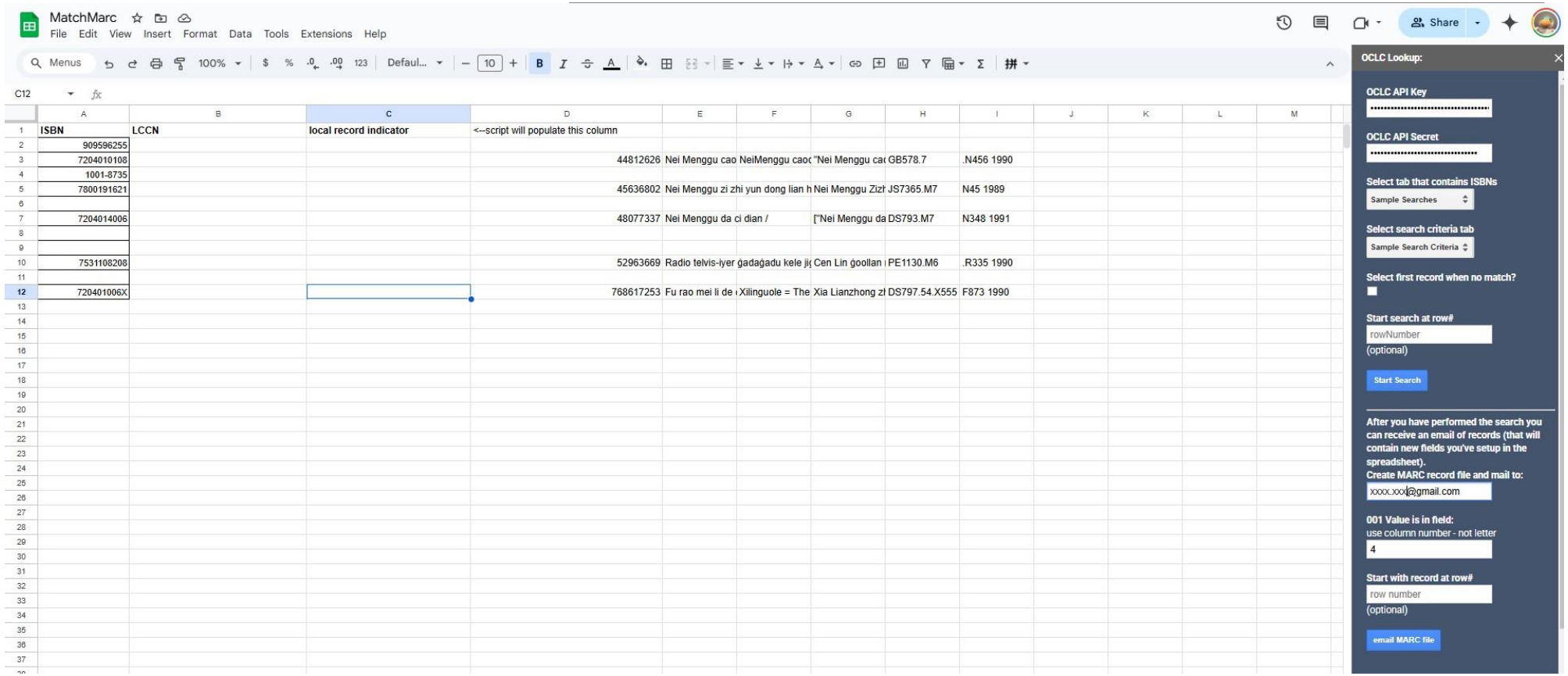
A1 search:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	**search:**																
2	holdings=PAU	<-- set your OCLC Symbol here or clear this cell if you don't want it to look for your own holdings															
3	040=DLC	040$b=eng	336$b=txt	337$b=n	338$b=nc			<-- all must be true for a 'match'									
4	042=pcc	040$b=eng	336$b=txt	337$b=n	338$b=nc			<-- all must be true for a 'match'									
5	040$b=eng	337$b=n	338$b=nc					<-- all must be true for a 'match'									
6	040$b=eng	337$a=unmediated															
7	040$b=eng	338$a=volume															
8																	
9																	
10																	
11	**fields:**	**starting column:**															
12	001	4	Cols 1,2,3 are used for ISBN, LCCN, 'local record found indicator'. 082:092:050 notation will print first field found.														
13	245$a																
14	245$b																
15	245$c																
16	050$a:090$a																
17	050$b:090$b																
18	100$a:110$a																
19																	
20																	
21																	
22																	
23																	
24																	
25																	
26																	
27																	
28																	
29																	
30																	
31																	
32																	
33																	
34																	
35																	
36																	
37																	
38																	

Sample Search Criteria | Sample Searches

*https://workspace.google.com/marketplace/app/matchmarc/903511321480

- MatchMARC is a Google Apps Script published as a Google Sheets Addon. It is a tool that automates MARC record searches. It uses OCLC's WorldCat Search API. *



*https://workspace.google.com/marketplace/app/matchmarc/903511321480

MatchMarc

LDR			01038nam#a22002295i#4500
001			9979515608803681
008			241025s9999||||xx#||||||||||||#||mon||
005			20241025195332.0
020			‡a [1007-1113?]
035			‡a (OCoLC)261615251#
245	0	0	‡a [Öbör Mongɣol-un Baǧsi-yin Yeke Surɣaǧuli-yin erdem sinjilegen-ü sedkül?] : : ‡b [Neyigem-ün sinjilekü uqaǧan-u keblel ?]
246			‡a Journal of Inner Mongolia Teachers' University : Philiosophy & Social Sciences
246			‡a 内蒙古师大学报 : 哲社蒙文版
264		1	‡a Huhehaote : ‡b Nei Menggu shi da xue bao bian ji bu,
300			‡a volumes
336			‡a text ‡b txt ‡2 rdacontent
337			‡a unmediated ‡b n ‡2 rdamedia
338			‡a volume ‡b nc ‡2 rdacarrier
500			‡a Mongolian edition ; volume55 (1992:no.3)
546			‡a In Mongolian (vertical script) with table of contents also in English
710			‡a Mongolian Grasslands collection
985			‡a stor

Before

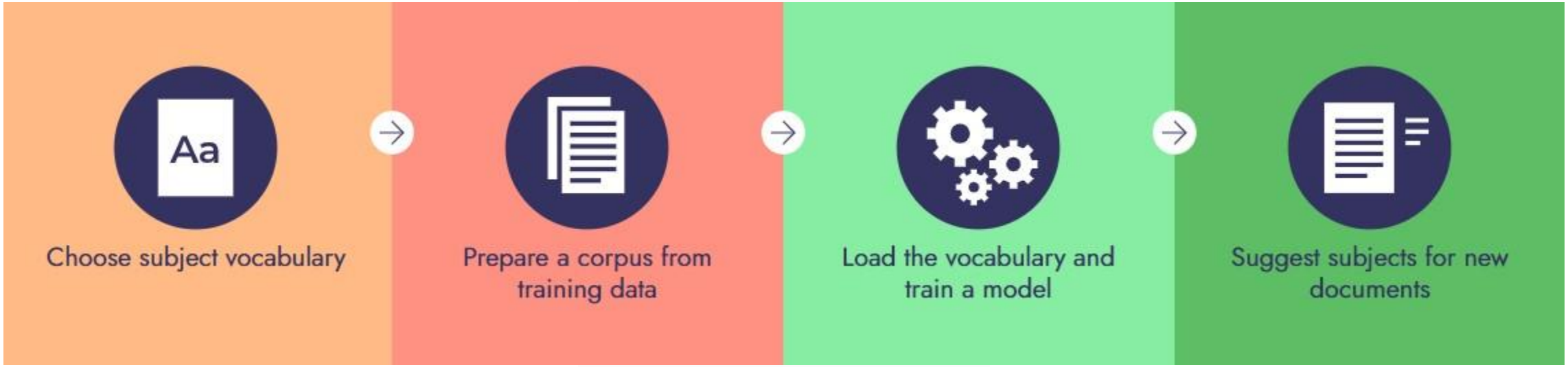
040			‡a XFF ‡b eng ‡c XFF ‡d OCLCF ‡d OCLCO ‡d DLC ‡d OCLCO ‡d OCLCQ ‡d OCLCO ‡d OCLCL
041	0		‡a mon ‡f eng
042			‡a lccopycat
043			‡a a-cc-im
050	0	0	‡a DS793.M7 ‡b O174
245	0	0	‡a Öbör Mongɣol-un Baǧsi-yin Yeke Surɣaǧuli-yin erdem sinjilegen-ü sedkül. ‡p Neyigem-ün sinjilekü uqaǧan-u keblel.
246	1	3	‡a Journal of Inner Mongolia Teachers' University ‡f 1985-
246	1		‡i English title varies, <2012, 1->: ‡a Journal of Inner Mongolia Normal University. ‡p Philosophy & social sciences
246	1	3	‡a Nei Menggu shi da xue bao. ‡p Zhe xue she hui ke xue ban ‡f 1982, 4-1988, 1.
246	1	3	‡a Erdem sinjilegen-ü sedkül (Öbör Mongɣol-un Baǧsi-yin Yeke Surɣaǧuli). ‡p Neyigem-ün sinjilekü uqaǧan-u keblel
246	1		‡i Mongolian part title varies, <2012, 1->: ‡a Gün uquǧan, neyigem-ün sinjilekü uqaǧan-u keblel
246	1	3	‡a Nei Menggu shi da xue bao. ‡p Meng wen zhe xue ban ‡f 1989, 2-1990, 1 7.
246	1	3	‡a Nei Menggu shi da xue bao. ‡p Zhe she Meng wen ban ‡f 1992-
260			‡a Kökeqota : ‡b Öbör Mongɣol-un Baǧsi-yin Yeke Surɣaǧuli-yin erdem sinjilegen-ü sedkül-ün Nayiraǧulqu Keltes, ‡c 1982-
300			‡a volumes : ‡b illustrations ; ‡c 26 cm
310			‡a Quarterly
336			‡a text ‡b txt ‡2 rdacontent
337			‡a unmediated ‡b n ‡2 rdamedia
338			‡a volume ‡b nc ‡2 rdacarrier
362	0		‡a 1982, 4-
500			‡a Issues have various subtitles in Chinese on p. [4] of cover.
500			‡a Some special issues (nemelte sedkül) have distinctive titles.
515			‡a Issues also numbered consecutively: 1982, 4 called also no. 15.
546			‡a In Mongolian (Mongolian script); table of contents also in English.
651	0		‡a Inner Mongolia (China) ‡v Periodicals.
650	0		‡a Social sciences ‡v Periodicals.
650	0		‡a Humanities ‡v Periodicals.
650	0		‡a Education ‡z China ‡z Inner Mongolia ‡v Periodicals.
651	6		‡a Mongolie-Intérieure (Chine) ‡v Périodiques.
650	6		‡a Sciences sociales ‡v Périodiques

After

Penn Libraries
University of Pennsylvania

Annif

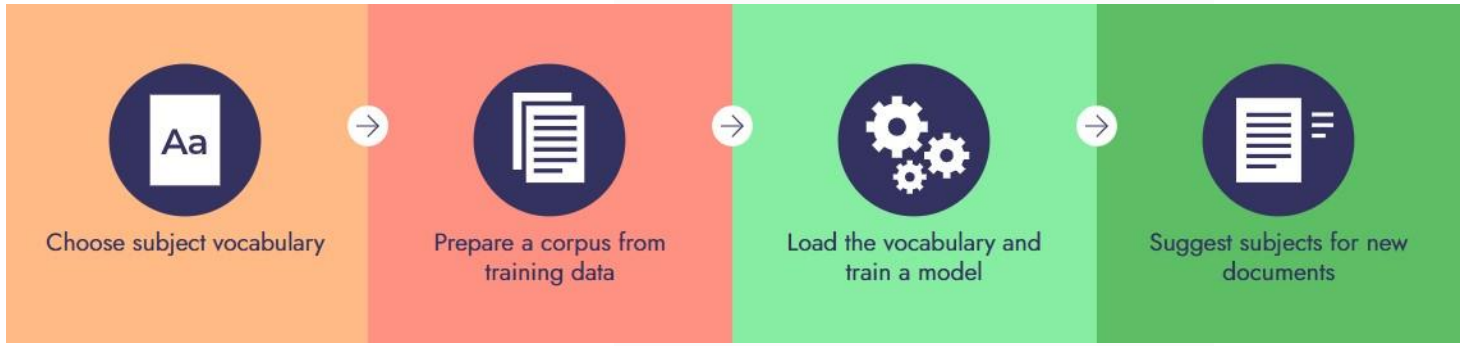
- **Annif** is an open-source tool for automated subject indexing. It is developed at the National Library of Finland.
- Annif uses a combination of existing **natural language processing** and **machine learning** tools including TensorFlow, Omikuji, fastText and Gensim. It is **multilingual** and can support **any subject vocabulary*.**



*https://annif.org/

Our test with Annif at Penn libraries

- **Jim Hahn**, Head of Metadata Research, is leading the pilot test.



Step 1 Use FAST subject headings as our primary subject vocabulary.

Step 2 Prepare a corpus from enriched MARC records

Step 3 Load RDF** files from OCLC to train a model

Step 4 Load a list of titles to generate a report with FAST subject headings.

annif

*https://pod.stanford.edu/
**https://www.oclc.org/research/areas/data-science/fast/download.html

Penn Libraries
University of Pennsylvania

We upload the list of titles to a shared drive on our local server

IPC (Information Processing Center) / **Using Annif-Airflow for FAST Subject Recommendations** Verified Share

Using Annif-Airflow for FAST Subject Recommendations

Owned by Jim Hahn, created with a template •••
Last updated: Feb 05, 2025 • 2 min read • 19 people viewed

This guide explains how to use our Penn Annif-Airflow service for FAST subject recommendations.

To make use of the service, format a spreadsheet as described below. The system is now configured to add to Alma any subjects recommendations over a .70 threshold. A report of the added subjects is generated in Box.

Note: it is up to the cataloger to check the records that get updated in Alma.

Instructions

1. First **format a spreadsheet** with the following columns: **mmsid** and **titleauthor**:

	A	B
1	mmsid	titleauthor
2	9979364596803681	마른 가지에 바람처럼 1~4 (리커버) / 달새울 지음 9788950994273
3	9979364596103681	전지적 독자 시점 Part 5 1,2 / 싱숑 지음 9788934967491
4	9978016352403681	Saturnario.
5	9978007249303681	Die Karl May Filme / Reinhard Weber ; Vorwort Rudolf Worschech.
6	9978093710603681	The Age of Anxiety : [exposition], Toronto, The Power Plant - Contemporary Art Gallery, september 22 - november 26, 1995 / Louise Dompierre Chief Curator.

mmsid, titleauthor

Quickstart

annif

Sample report with FAST subject headings

	A	FAST subject headings	Confidence Index
1	row_number,group_summary,refined_llm_agreed_subjects,annif_score		
2	1,The text describes the history and development of Qigong an ancient Chinese practice that combines physical movement breathir	['Qi gong']	[0.912917971611023]
3	20,Chinas grasslands account for approximately of the countrys total land area playing a crucial role in livestock farming and enviro	['Grassland ecology']	[0.7829089760780334]
4	26,The book Ethnodemographics by Zhang Tianlu is a comprehensive summary of decades of research on the establishment and d	['Puppet plays, Chinese']	[0.7196585536003113]
5	69,The text describes the life and work of Na Saiin a prominent contemporary Chinese poet and founder of contemporary Mongolia	['Mongolian language']	[0.7279444932937622]
6	70,The text describes the Donghu a tribal confederation of Hu nomadic people that lived in northern Hebei southeastern Inner Mon	['Xiongnu (Asian people)']	[0.9429646730422974]
7	90,The text appears to be a collection of abstracts and descriptions related to Mongolia and Inner Mongolia touching on topics such	['Shamanism']	[0.7850553393363953]
8	92,The text appears to be a collection of references and citations from various academic journals and publications likely from a bibl	['Citation of legal authorities']	[0.7194916009902954]
9	94,The text appears to be a collection of linguistic resources and research papers related to the Mongolian language including dictic	['Dagur language']	[0.7436279058456421]
10	103,assistant The text appears to be a passage written in Mongolian with some English words and names interspersed and it seems	['Mongols--Kings and rulers']	[0.7601145505905151]
11	139,The relationship between language and cognitive development is a complex and multifaceted one with research suggesting tha	['Cognition']	[0.753873348236084]
12	139,The relationship between language and cognitive development is a complex and multifaceted one with research suggesting tha	['Bilingualism']	[0.7441017031669617]
13	139,The relationship between language and cognitive development is a complex and multifaceted one with research suggesting tha	['Cognition']	[0.7481814622879028]
14	146,A study analyzing the spatial and temporal changes of grassland resources in the Xilinguole region of Inner Mongolia from to fo	['Grassland ecology']	[0.8428817987442017]
15	147,The Silin Guelleh League in Inner Mongolia plays a crucial role in Chinas energy security and ecological environment but its coal	['Grassland ecology']	[0.7032566070556641]
16	148,The text describes Xilingol in Inner Mongolia known for its beautiful grasslands savannahs and unique winter landscapes with	['Grassland ecology']	[0.7445220947265625]
17			
18			

Step 5: Human Evaluation

Penn Libraries
University of Pennsylvania

annif

LDR			00875nam#a22001935#4500
001			9979515473403681
008			241025s1987||||xx#||||||||||||#||mon||
005			20241025195332.0
245	0	0	‡a 锡林郭勒草地资源 = : ‡b Xilinguolecaodiziyuan ‡c Nei Menggu Xilinguole Meng cao yuan gong zuo zhan ; Gereltū [Geriletu] zhu bian /
246			‡a Xilinguole cao di zi yuan
264		1	‡a Xilinguole : ‡b Nei Menggu Xilinguole Meng Caoyuan Gongzuozhan [Inner Mongolia Xilingol League Grassland Workstation], ‡c 1987
300			‡a 336 pages, 9 unnumbered pages of plates
336			‡a text ‡b txt ‡2 rdacontent
337			‡a unmediated ‡b n ‡2 rdamedia
338			‡a volume ‡b nc ‡2 rdacarrier
500			‡a "全国北方重点牧区"
546			‡a In Chinese
710			‡a Mongolian Grasslands collection
985			‡a stor

Before

LDR			01546nam#a22002775i#4500
001			9979515473403681
005			20250219091928.0
008			241025s1987||||cc#abf||||||||||#||chi||
043			‡a a-cc-im
245	0	0	‡6 880-01 ‡a Xilinguole cao di zi yuan = ‡b Xilinguolecaodiziyuan / ‡c Nei Menggu Xilinguole Meng cao yuan gong zuo zhan.
	0	0	‡6 245-01/$1 ‡a 锡林郭勒草地资源 = ‡b Xilinguolecaodiziyuan / ‡c 内蒙古锡林郭勒盟草原工作站.
246			‡a Xilinguolecaodiziyuan
264		1	‡6 880-02 ‡a Xilinguole : ‡b Nei Menggu Xilinguole Meng Caoyuan Gongzuozhan, ‡c 1987.
		1	‡6 264-02/$1 ‡a 锡林郭勒 : ‡b 内蒙古锡林郭勒盟草原工作站, ‡c 1987.
300			‡a 8 unnumbered pages of plates, 336 pages : ‡b illustrations (some color), maps ; ‡c 27 cm
336			‡a text ‡b txt ‡2 rdacontent
337			‡a unmediated ‡b n ‡2 rdamedia
338			‡a volume ‡b nc ‡2 rdacarrier
500			‡a "全国北方重点牧区".
546			‡a In Chinese.
650		7	‡a Grassland ecology. ‡2 fast ‡0 (OCoLC)fst00946801
651		0	‡a Xilin Gol Meng (China)
710	2		‡6 880-03 ‡a Nei Menggu Xilinguole Meng cao yuan gong zuo zhan, ‡e editor.
	2		‡6 710-03/$1 ‡a 内蒙古锡林郭勒盟草原工作站, ‡e editor.
710			‡a Mongolian Grasslands collection
985			‡a stor

After

Penn Libraries
University of Pennsylvania

annif

Some useful links about Annif:

1. The machine learning model that Jim used for this project is a NN-ensemble (https://github.com/NatLibFi/Annif/wiki/Backend%3A-nn_ensemble) of Omikuji (https://github.com/NatLibFi/Annif/wiki/Backend%3A_Omikuji) and TF-IDF (https://github.com/NatLibFi/Annif/wiki/Backend%3A-TF-IDF)
2. Annif and Finto AI: DIY automated subject indexing from prototype to production https://www.youtube.com/watch?v=nzK97hzPMNE (November, 2020)
3. Annif tutorials: https://www.youtube.com/watch?v=qqSrVXEzKjw&list=PLa9kvrI3VLf5K-bjvVDaIWMi5CACGjPUM
4. Insight into the machine-based subject cataloguing at the German National Library https://www.youtube.com/watch?v=CM2eneIf-HU (December, 2022)
5. Annif users' mailing list https://groups.google.com/g/annif-users?pli=1

(information provided by Jim Hahn)