



Batch Authority Searching with Python

A work-in-progress

Kelsey George
Cataloging & Metadata Strategies Librarian
University of Nevada, Las Vegas

Project

UNLV received over 3,000 retrospective ETD records from ProQuest which didn't meet our standards.

- Varying capitalization across author and title fields
- Needed to be upgraded to RDA (missing carrier fields; abbreviations; and contributor information)
- Superfluous ProQuest codes in 590, 690, and 790 fields.

Project

One goal was to upgrade the minimal subject headings found in the records, all ProQuest supplied 650 #4 fields, and replace them with LCSH terms for the MARC and FAST headings for our institutional repository (IR), Digital Scholarship@UNLV.

DIGITAL SCHOLARSHIP@UNLV

Step 1: *Isolate unique subject heading values provided by ProQuest*

=650 \4\$aGeology.
=690 \a0372

=LDR 02274nam a2200301 4500
=001 AAI1331529
=005 20171019073145.5
=008 171019s1986\\\$a|eng\\\$d
=035 \\\$a(MiAaPQ)AAI1331529
=040 \\\$aMiAaPQ\$cMiAaPQ
=100 1\\\$aHARDY, JOSEPH KIRK.
=245 10\\\$aSTRATIGRAPHY AND DEPOSITIONAL ENVIRONMENTS OF LOWER AND MIDDLE CAMBRIAN STRATA IN THE LAKE MEAD REGION, SOUTHERN NEVADA AND NORTHWESTERN ARIZONA.
=260 1\\\$aAnn Arbor : \$bProQuest Dissertations & Theses, \$c1986
=300 \\\$a324 p.
=500 \\\$aSource: Masters Abstracts International, Volume: 26-01, page: 9100.
=502 \\\$aThesis (M.S.)--University of Nevada, Las Vegas, 1986.
=520 \\\$aThe stratigraphy and depositional systems revealed within the Lower to Middle Cambrian Tapeats Sandstone-Pioche Shale-Lyndon Limestone-Chisholm Shale succession were studied between the easternmost limit of Cretaceous overthrusting in southern Nevada and the western edge of the Colorado Plateau in northwestern Arizona. Nine intergradational lithofacies were distinguished based on a rich suite of sedimentary structures, including lenticular-, wavy-, and flaser-bedding, wrinkle marks, herringbone cross-stratification, convolute bedding, primary current lineations, interference ripple marks, bird's eye structures, thrombolites, intraformational conglomerates, oncoids, cryptalgal laminites, and Arenicoloides, Corophioides, and Skolithos ichnofossils. The collective presence of these structures indicates that Lower and Middle Cambrian rocks in the study area were deposited in shallow-subtidal to supratidal settings.
=520 \\\$aThe proposed depositional model incorporates an east-west oriented Cambrian shoreline and asserts that the Lower and Middle Cambrian sequence represents a series of sedimentological responses to fluctuating rates of relative sea-level rise.
=590 \\\$aSchool code: 0506.
=650 \4\$aGeology.
=690 \a0372
=710 20\\\$aUniversity of Nevada, Las Vegas.
=773 0\\\$tMasters Abstracts International\$g26-01.
=790 \\\$a0506
=791 \\\$aM.S.
=792 \\\$a1986
=793 \\\$aEnglish
=856 \\\$uhttp://gateway.proquest.com/openurl?url_ver=Z39.88-2004

Step 1: *Isolate unique subject heading values provided by ProQuest*

O
SUBJ DESC
Geology
Mass communication Law
Early childhood
Archaeology American history Religious history
Geology

AA	AB
SUBJ GROUP DESC	SUBJ CODE
Earth Sciences	372
Communication and the Arts Social Sciences	0708 0398
Education Education	0518 0529
Social Sciences Social Sciences Philosophy, Religion and Theology	0324 0337 032
Earth Sciences	372
Earth Sciences	372

Step 1: *Isolate unique subject heading values provided by ProQuest*

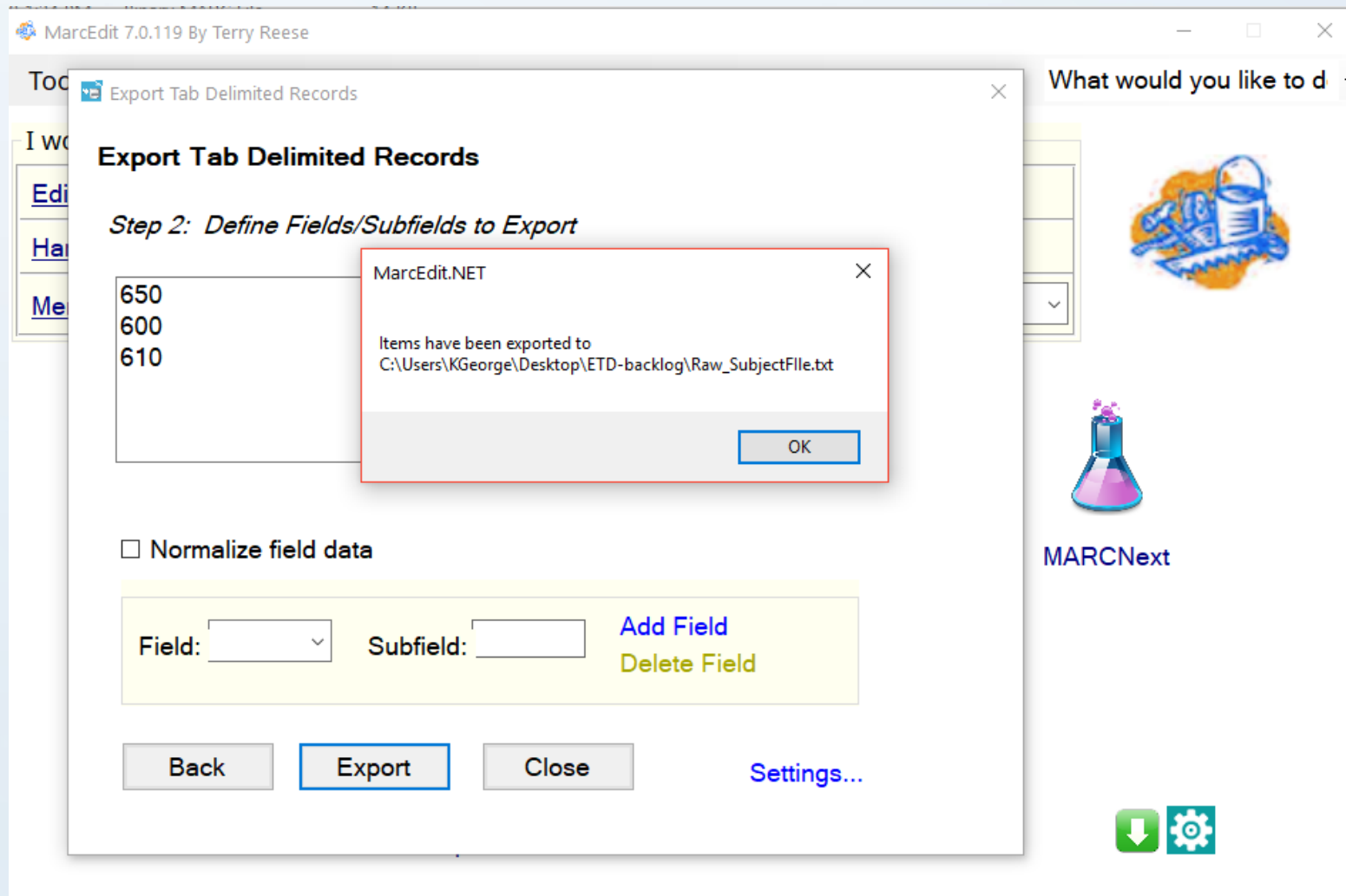
Details

Subject
Geology

Classification
0372: Geology

It became apparent that ProQuest assigned our MARC records' headings based on their own taxonomy. The 690 field was a code referencing their subject.

Step 1: *Isolate unique subject heading values provided by ProQuest*



From MARCEdit, I exported just the 6xx fields into an Excel spreadsheet via a tab delimited text file.

Step 1: *Isolate unique subject heading values provided by ProQuest*

Uploaded that
Excel into
OpenRefine in
order to
deduplicate the
values.

Refine^{OPEN} Raw_SubjectFile.xlsx [Permalink](#)

Facet / Filter Undo / Redo 1

Refresh Reset All Remove All

3124 rows

Show as: rows records Show: 5 10 25 50 rows

▼ All ▼ 650.0 1 ▼ 650.0 2 ▼ 650.0 3 ▼ 650.0 4 ▼ 650.0 5 ▼ 650.0 6

☆	1.	Geology.				
☆	2.	Mass communication.	Law.			
☆	3.	Early childhood education.	Special education.			
☆	4.	Archaeology.	American history.	Religious history.		
☆	5.	Geology.				
☆	6.	Geology.				
☆	7.	Animal Physiology.				
☆	8.	Geology.				
☆	9.	Computer science.				
☆	10.	Hydrologic sciences.				

650.0 1 change

193 choices Sort by: name count Cluster

Accounting. 5
Adult education. 13
Aerospace engineering. 3
Agricultural economics. 2
Agricultural education. 1
Agronomy. 4
American history. 50
American literature. 83
American studies. 2
Analytical chemistry. 34
Animal diseases. 1
Animal Physiology. 20

650.0 2 change

229 choices Sort by: name count Cluster

Statistics. 5
Systems science. 2
Teacher education. 23
Theater. 9
Transportation. 18
Urban planning. 17
Veterinary science. 1
Vocational education. 9
Women's studies. 52
Zoology. 8
(blank) 991

Facet by choice counts

Step 1: *Isolate unique subject heading values provided by ProQuest*

A
Proquest Subject Description
Geography
Geology
Geophysics
Geotechnology
German literature
Gerontology
Health care management
Health education
Health sciences
Health Sciences, Education
High energy physics
Higher education

SubjDesc.txt - Notepad
File Edit Format View Help
Accounting
Acoustics
Adult education
Aerospace engineering
African history
African literature
Agricultural chemistry
Agricultural economics
Agricultural education
Agricultural engineering
Agriculture
Agronomy
American history
American literature
American studies
Analytical chemistry
Ancient history
Animal diseases
Animal Physiology
Aquatic sciences

Went back into Excel so I had one sheet with 257 unique values for subject headings that I put into a .txt file.

Step 2: Search for corresponding LCSH terms

I decided to use the Batch subject heading search tool in OCLC Connexion.

The screenshot shows the 'Enter Authority Batch Search Keys' dialog box. At the top, there's a 'Local File:' label and a 'Local File Manager...' button. Below this is a text field containing the file path 'C:\Users\KGeorge\AppData\Roaming\OCLC\Connex\Db\ETD.retrospective.subj.auth.db'. The 'Search Keys' section includes a 'Query:' label, a 'Total Entered: 257' status, and an 'Enter Diacritics...' button. A 'Use default index:' dropdown is set to 'LCSH (su:)'. A list box displays the following terms: 'su:Accounting', 'su:Acoustics', 'su:Adult education', 'su:Aerospace engineering', and 'su:African history'. To the right of the list box are buttons for 'Add', 'Delete', 'Replace', 'Import...', 'Copy...', and 'Print'. At the bottom, there are checkboxes for 'Retrieve all records from online save file' and 'Delete downloaded records from online save file', along with a 'Limit by Review Status:' section containing 'Non-Submitted' and 'Submitted' options. The bottom right corner features 'Save', 'Close', and 'Help' buttons.

Enter Authority Batch Search Keys

Local File: Local File Manager...

C:\Users\KGeorge\AppData\Roaming\OCLC\Connex\Db\ETD.retrospective.subj.auth.db

Search Keys

Query: Total Entered: 257 Enter Diacritics...

Use default index: LCSH (su:)

su:Accounting
su:Acoustics
su:Adult education
su:Aerospace engineering
su:African history

Add
Delete
Replace
Import...
Copy...
Print

☐ Retrieve all records from online save file

Limit by Review Status:

☐ Non-Submitted
☐ Submitted

☐ Delete downloaded records from online save file

Save Close Help

Step 2: *Search for corresponding LCSH terms*



Errors (20)
Too Many Matches (226)
Successful Searches (11)

“Wouldn’t
it be
nice?”



“Have to
meet a
deadline”

A	B	C	D	E	F	G	H
Proquest Subject Description	Error Message	650 Field	010 Field	651 Field	010 Field	Records corrected	
Geography		Geography	sh 85053986			14	
Geology		Geology	sh 85054037			125	
Geophysics		Geophysics	sh 85054185			12	
Geotechnology		Geotechnology	sh2013000289			16	
German literature		German literature	sh 85054380			2	
Gerontology		Gerontology	sh 85054694			28	
Health care management		Health services administratio	sh 85059600			16	
Health education		Health education	sh 85059537			26	
Health sciences		Medical sciences	sh 85083022			11	
Health Sciences, Education		Medical sciences\$Study and	sh2009010083			15	
High energy physics		Particles (Nuclear physics)	sh 85098374			1	
Higher education		Education, Higher	sh 85041065			83	
	No records found for your search. Please change or simplify your search and try again						
Hispanic American studies		Hispanic Americans\$Study				10	
History, Church		Church\$History of doctrines				3	
Home economics		Home economics	sh 85061673			2	
Home economics education		Home economics\$Study and	sh 85061679			1	
Hydrologic sciences	No records found for your search	Hydrology	sh 85063458			73	
Immunology		Immunology	sh 85064579			2	
Individual & family studies	No records found for your search	Social psychology	sh 85123994			3	
Industrial engineering		Industrial engineering	sh 85065864			13	
Information science		Information science	sh 85066150			25	
Inorganic chemistry		Chemistry, Inorganic	sh 85023017			11	

Find

Replace

Find:

=650 \4\$aGeology.

Replace

Replace:

=650 \0\$aGeology.

Replace All

☐ Perform Find/Replace If..

Close

Search Options:

☒ Match case☐ Exact Word Match☐ Use regular expressions☐ MultiLine Evaluation☐ Use External Search/Replace Criteria

C	D	E	F	G	H
650 Field	010 Field	651 Field	010 Field	Records corrected	
Geography	sh 85053986			14	
Geology	sh 85054037			125	
Geophysics	sh 85054185			12	
Geotechnology	sh2013000289			16	
German literature	sh 85054380			2	
Gerontology	sh 85054694			28	
Health services administratio	sh 85059600			16	
Health education	sh 85059537			26	
Medical sciences	sh 85083022			11	
Medical sciences\$xStudy and	sh2009010083			15	
Particles (Nuclear physics)	sh 85098374			1	
Education, Higher	sh 85041065			83	



FAST Converter

Convert LSCH Subject Headings to FAST Subject Headings

ENTER LIBRARY OF CONGRESS SUBJECTS:

600 \$aWashington, George, \$d 1732-1799
650 \$aTrenton, Battle of, Trenton, N.J., 1776 \$v
Juvenile literature.
651 \$aDelaware River (N.Y.-Del. and N.J.).

Enter LCSH subject headings here, using \$ for field separators and including the 6xx field number. Clicking Convert will start the conversion, and the results will appear to the right. This converter has a 20 heading limit.

FILE UPLOAD:

Choose File No file chosen

Browse for a Marc file with LCSH headings (MARC-8 or UTF-8). By clicking Convert, a resulting file with FAST headings will be available for download to the right. Limit approximately 500 records.

CONVERT

Delimiter selection ☒ \$ Dollar sign ☐ ‡ Double Dagger ☐ † Double-barred Pipe

FAST Result

600 17 \$a Washington, George, \$d 1732-1799 \$2 fast \$0 (OCoLC)fst00178100
611 07 \$a Trenton, Battle of (New Jersey : 1776) \$2 fast \$0 (OCoLC)fst0140429
648 7 \$a 1776 \$2 fast
651 7 \$a New Jersey \$z Trenton. \$2 fast \$0 (OCoLC)fst01207908
651 7 \$a United States \$z Delaware River. \$2 fast \$0 (OCoLC)fst01310316
655 7 \$a Juvenile works. \$2 fast \$0 (OCoLC)fst01411637

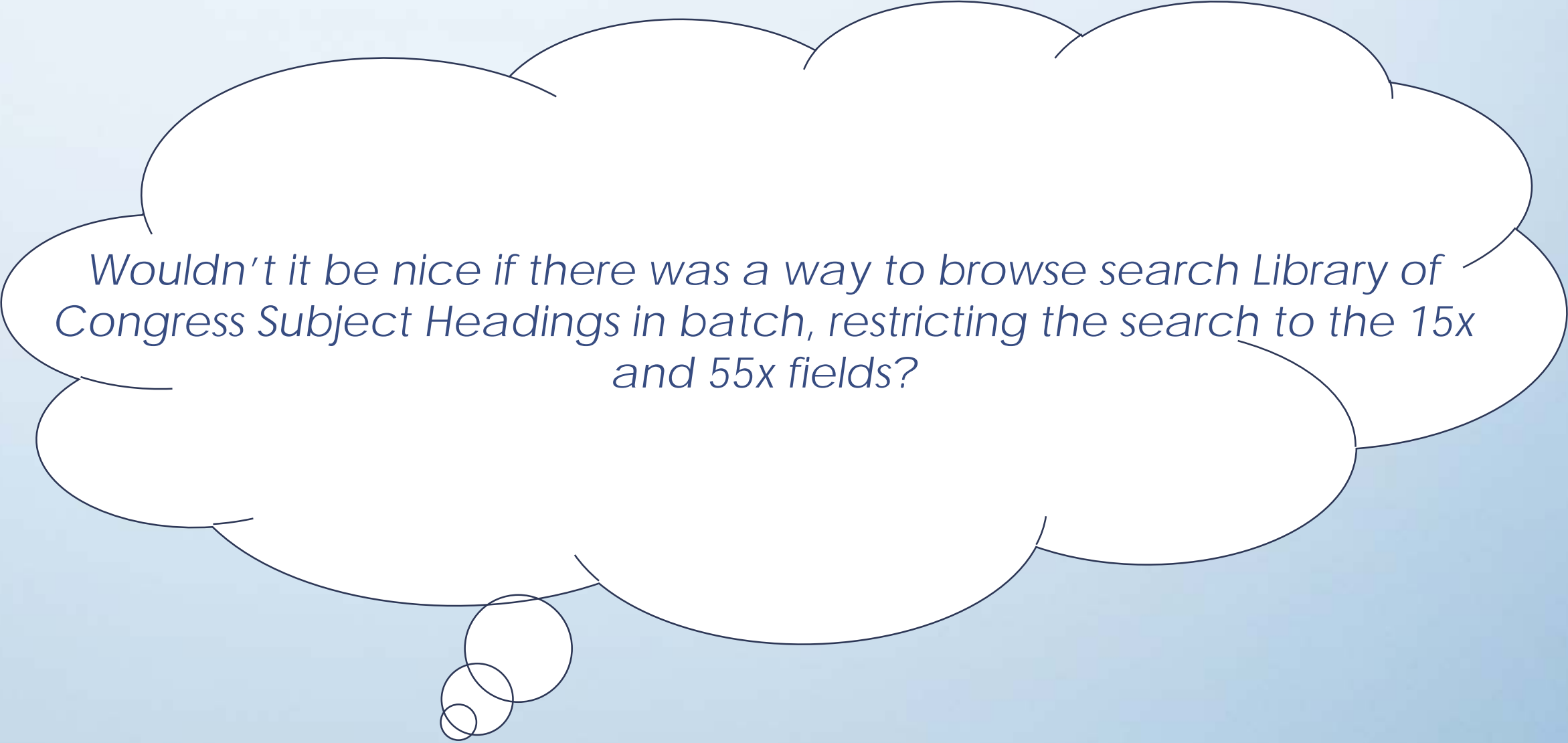
DOWNLOAD

A	W	X
title	keywords	subject_controlled
Stratigraphy And Depositional Environments Of Lower And Middle Cambrian Strata In The Lake Mead Region, Southern Nevada And Northwestern Arizona.	Arizona; Cambrian; Depositional; Environment; Lake; Lake Mead; Lower; Mead; Middle; Nevada; Northwestern; Region; Southern; Strata; Stratigraphy	Geology
Attorney advertising: The effect on juror perceptions and verdicts.	Advertising; Attorney; Effects; Juror; Perceptions; Verdicts	Mass media; Law
The effects of early intervention on young handicapped children who are nonverbal or have limited expressive language skills.	Children; Early; Effects; Expressive; Handicapped; Interventions; Language; Limited; Nonverbal; Skills; Young	Early childhood education; Special education
A matter of faith: A study of the Muddy Mission.	Faith; Matter; Mission; Muddy; Mormon; Nevada; Study	Archaeology; Religion; History
Provenance and tectonic significance of the Lower Paleozoic Douglas Conglomerate, northern Churchill Mountains, Antarctica.	Antarctica; Churchill; Conglomerate; Las Vegas; Lower; Mountains; Northern; Paleozoic; Provenance; Significance; Tectonic	Geology
The Muddy Creek Formation: Depositional environment, provenance, and tectonic significance in the western Lake Mead area, Nevada and Arizona.	Area; Arizona; Creek; Depositional; Environment; Formation; Lake; Mead; Muddy; Nevada; Provenance; Significance; Tectonic; Western	Geology
The effect of wetsuit leg coverage on swimming speed and selected physiological measures.	Coverage; Effect; Leg; Measures; Physiological; Selected; Speed; Swimming; Wetsuit	Physiology
Morphology and development of the Red Rock Canyon alluvial fan, Clark County, Nevada.	Alluvial; Canyon; Clark; County; Development; Fan; Morphology; Nevada; Red; Rock	Geology
Grasp--a language to facilitate the synthesis of parallel programs.	Facilitate; Grasp; Language; Parallel; Program; Synthesis	Computer science
Hydrology of Bishop Creek, Inyo County, California: An isotopic analysis.	Analysis; Bishop; California; County; Creek; Hydrology; Inyo; Isotopic	Hydrology
Inversion of input/output map, sliding mode and nonlinear flight control system design.	Control; Design; Flight; Input; Inversion; Map; Mode; Nonlinear; Output; Sliding; Systems	Electrical engineering; Computer science

“Wouldn’t
it be
nice?”



“Have to
meet a
deadline”



Wouldn't it be nice if there was a way to browse search Library of Congress Subject Headings in batch, restricting the search to the 15x and 55x fields?

Is it possible to scrape <https://authorities.loc.gov/> and use regular Python to find matches?

Is it possible to use Python to search a downloaded "LC Subject Headings (MADS/RDF or SKOS/RDF only)" file?



MARCEdit has a validate headings tool, is there room for interoperability?

Questions to answer:

- Would the development of a Python-based tool be useful for our work long-term?
- What kind of features would it need to meet the needs of a wide audience?
 - What level of knowledge/coding skills would it require users to operate?
 - What contexts could it be used in?
- How would it improve interoperability between MARC, existing metadata, and linked data?

Thank you!

Email: kelsey.george@unlv.edu

