Managing Metadata Overload:

Automating E-Resource Workflows with Computer Scripts

Benjamin Bradley
Discovery Librarian
University of Maryland Libraries

UNIVERSITY OF MARYLAND 1856
UNIVERSITY LIBRARIES

Check E-Resource Access with the E-Resource Access Checker

JRuby

What?

The E-Resource Access Checker is a **JRuby script** developed by Kristina Spurgin that enables librarians to automate link-checking for electronic resources.

- **Used for:**
  - Checking if links work
  - Checking if publisher is providing access to entitlements
- **Powerby by:**
  - JRuby

GitHub repository:
github.com/UNC-Libraries/Access-Checker

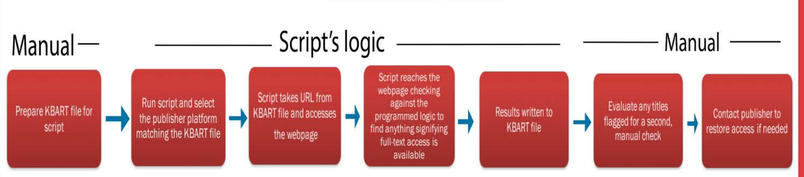
Read Kristina's article, "Getting What we Paid for: a Script to Verify Full Access to E-Resources," here:
https://journal.code4lib.org/articles/9684

How?

The script checks a batch of titles for individual platforms. As long as the file you run the script on is a CSV (comma separated values) file, the script should work. The file does not need to be a KBART file or other format; you just need to ensure that the URLs are moved to the last column.

The script iterates through the CSV file, following each URL and reading the HTML on the page. It checks the HTML against a set of programmed conditions to evaluate if the publisher is providing access to the title. If access is not provided, the script will report that the title should be checked manually, and then the librarian can follow-up with the provider to resolve any access problems.

Workflow and script's logic




```
elsif package == "igi"
  sleeptime = 10
  if  page.match(/title=:"Owned"/)
      access = "owned"
  elsif page.include?("Institution Prices")
      access = "not owned"
  else
      access = "check manually"
  end
```

Code for IGI Global, illustrating the logic for determining if access is available

Screenshot demonstrating the output of the script (highlighted column)

publication_title	title_url	access
Starer/Thorne	http://www.	Full access
Vinny Golla Quintet: Nation Of Laws	http://www.	Full access
Descansos, past	http://www.	Full access
Joelle Leandre/Phillip Greenlief: That Overt Desire Of O	http://www.	Full access
Bob Nell - Why I Like Coffee	http://www.	Full access
Andrei Eshpai Edition, Vol. 3	http://www.	Full access
Electronic Pioneers	http://www.	Full access
Andrew D'Angelo: Skadra Degis	http://www.	Full access
Nicholas Anthony Ascioti: Creation's Voice	http://www.	Full access
Night Watch	http://www.	Full access
Bonnie Barnett and Ken Filiano: Trio For Two	http://www.	Full access
Morton Subotnick: Electronic Works Vol. 2	http://www.	Full access
El Motivo	http://www.	Full access
Old-Country Music in a New Land: Folk Music of Immig	http://www.	Full access
Alessandro Scarlatti: La Giuditta	http://www.	Full access
Polish Romantic Violin Music	http://www.	Full access
Tribute to Soprano Bethany Beardslee	http://www.	Full access
Dello Joio: Family Album, Piano Works, Vol. 3	http://www.	Full access
Tom Johnson - Music for 88	http://www.	Full access
Larry Polansky: The World's Longest Melody	http://www.	Full access
Birds of Maya: Volume 1	http://www.	Check access
David Behrman - Unforseen Events	http://www.	Full access
American Clarinet	http://www.	Full access
Makiko Nishikaze: pianopera I & II	http://www.	Full access
Chas Smith: Nikko Wolverine	http://www.	Full access
Cassatt String Quartet	http://www.	Full access
Phonographic Yearbook: 1912	http://www.	Full access

Create KBART files to supplement publisher data

python

What?

The MARCDownloader is a Python script that uses the **WorldCat Search API** to find and then transform MARC records into KBART.

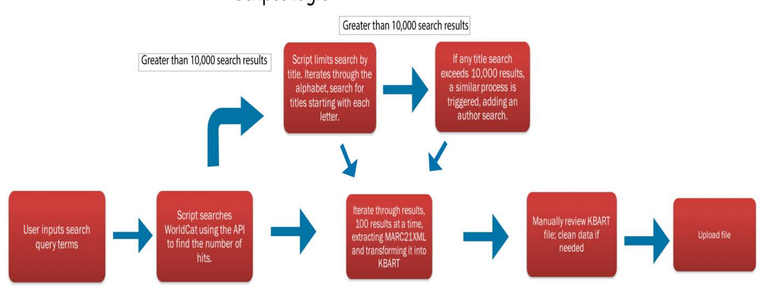
- **Used for Improving discovery and access of resources by:**
  - Supplementing knowledge base collections in **WorldCat Collection Manager** that are missing important data such as URLs or OCLC numbers
  - Creating collections when one is unavailable for subscription or open access platforms and collections.
- **Powered by:**
  - Python
  - WorldCat Search API

Under-development GitHub Repository:
https://github.com/bradley-benjamin26/WCSearchAPIMARCHarvester

How?

The script uses a query to search WorldCat, most often searching URLs to find a set of records for titles in a database or platform. It then iterates through the search results converting the relevant MARC data into KBART. If desired, the script can be edited to clean the MARC21XML data, especially for URLs in 856 fields. Searches should be targeted and concise because the API only provides access to the first 10,000 results. The script has a workaround to add title and author searching, but it is not effective and is a last resort. When completed, the output can be uploaded into Collection Manager or added to another collection as required.

Workflow and script's logic



Screenshot of a KBART created by the script

publication_title	print_identi	online_iden	date_first_i	num_first_v	num_first_i	date_last_i	num_last_v	num_last_i	title_id	first_author	title_id	embargo_in	coverage_de	coverage_n	publisher_n	location	title_notes	staff_notes	vendor_id	oclc_collecti	oclc_collecti	oclc_entry_i	oclc_linksch	oclc_number	ACTION
The relationship between bike produc		1975			1975				https://ear	Shenn, Brian Raymond,		ebook			ProQuest Dissertations Publishing					Proquest DI customer ID	12785195		12785195	row
Energy exchange on a melting glacier		1975			1975				https://ear	Munro, Donald Scott MacC,		ebook			ProQuest Dissertations Publishing					Proquest DI customer ID	12818844		12818844	row
Santos: the household wooden saints		1975			1975				https://ear	Lange, Yvonne Marie,		ebook			ProQuest Dissertations Publishing					Proquest DI customer ID	12844857		12844857	row
A technique for measuring food intak		1975			1975				https://ear	Warrigan, Anthony Paul,		ebook			ProQuest Dissertations Publishing					Proquest DI customer ID	12718317		12718317	row
The feasibility of rewritten subject ma		1975			1975				not found	Arnold, Jeffrey Kent,		ebook			ProQuest Dissertations Publishing					Proquest DI customer ID	12824402		12824402	row
Proteins in development: two studies		1975			1975				https://ear	Hubbard, Daniel Lawrence,		ebook			ProQuest Dissertations Publishing					Proquest DI customer ID	12785205		12785205	row
Cytogenetic studies on Peromyscus m		1975			1975				https://ear	Murray, James D,		ebook			ProQuest Dissertations Publishing					Proquest DI customer ID	12785199		12785199	row
Reproductive endocrine interrelations		1975			1975				https://ear	Smith, Robert Wyatt,		ebook			ProQuest Dissertations Publishing					Proquest DI customer ID	12785201		12785201	row
The plan-makers and the city: architects, engineers, surveyors and urban planning in Canada, 1880-1939									https://ear	Van Nus, Walter,		ebook			ProQuest Dissertations Publishing					Proquest DI customer ID	41858385		41858385	row

Some code for normalizing URLs

```
for url in urls:
    docViewUrl = re.search(r'docview/(\d+)', url)
    if docViewUrl is None:
        gatewayURL = re.search(r'gateway\.proquest\.com/(openurl.+)', url)
        if gatewayURL is None:
            umiURL = re.search(r'www\lib.+)', url)
            if umiURL is None:
                titleURL = "not found"
            else:
                titleURL = umiURL.group(1)
        else:
            titleURL = "http://gateway.proquest.com/" + gatewayURL.group(1)
    else:
        titleURL = "https://search.proquest.com/docview/" + docViewUrl.group(1)
```

When I initially created the script, it was just to download MARC records that I then cleaned up manually with regular expressions and then transformed into a KBART file using MarcEdit. After improving my coding knowledge, I was able to clean the data returned by the API and convert it into a KBART automatically, dramatically increasing the workflow's efficiency.

Collect Coverage Data and License Terms with KBQuery

python

What?

KBQuery is a Python script that runs automated batch searches, combining data from the **WorldCat knowledgebase API** and the **WorldShare License Manager API.** The script outputs the report as a tabbed separated value (TSV) file.

- **Used for:**
  - Finding different sources your library has to access titles (electronic subscription, aggregator database, print, etc.)
  - Compiling coverage and licensing data to support collection development activities
  - Checking if entitlements are all selected in the knowledge base to support e-resource maintenance work
- **Power by:**
  - Python
  - WorldCat knowledge base API
  - WorldShare License Manager API

Under-development GitHub repository:
https://github.com/bradley-benjamin26/WCKBQuery

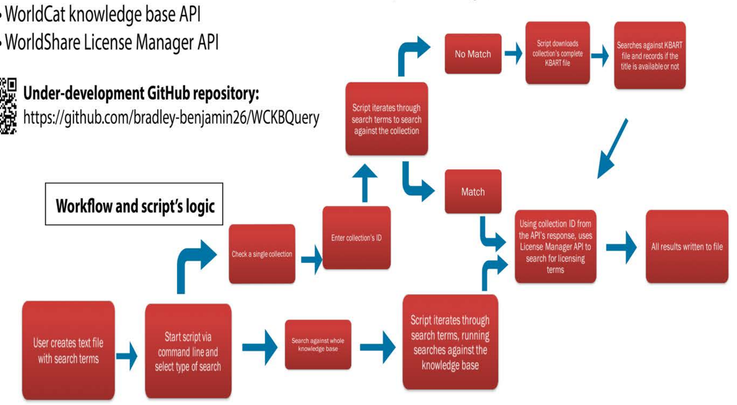
How?

KBQuery reads a text file, created by the user, containing a list of search terms (ISBN, ISSN, OCLC Number or title). The script then runs a query against the knowledge base using each search term. If a match is found, it pulls out a set of data, including the available coverage, and writes it to a file. It also then takes the Collection ID from the knowledge base API's response to search for a corresponding license using the License Manager API. If a license is found, it searches for perpetual access rights an archival access rights and adds that to the output file.

The script has two primary modes:

1) Search within a single collection (ideal for maintenance work to ensure your entitlements are selected)
2) Search against the whole knowledge base (ideal for supporting collection development work by identifying your complete coverage).

Workflow and script's logic



Example of a KBQuery Report

Number field groups together the search terms with the search results. The format is: (search number).(result number)

Status represents if the title is selected or not

Coverage contains the format and coverage date/publication date as listed in the knowledge base

number	status	title	ISBN or ISSN	ocn	Collection Name	KB ID	coverage	Perpetual Access	Archival Copy	Search Term
1.1	selected	012 IEEE 20th Symposium on Mass Storage Sy	9781467317450	812607995	IEEE Xplore All Conference Proceedings	11067293	ebook@2012	no or silent	yes	812607995
1.2	selected	012 IEEE 20th Symposium on Mass Storage Sy	9781467317450	812607995	IEEE/IET Electronic Library (IEL)	no KB ID	ebook@2012	no or silent	yes	812607995
1.3	selected	Mass Storage Systems and Technologies (MSS'	No Standard numb	812607995	IEEE Proceedings (UMC)	812607995	ebook	no or silent	yes	812607995
2.1	selected	Energy Journal	0195-6574	563150979	Business Source Complete	514991	fulltext@1990-01-01	no or silent	silent	563150979
2.2	selected	Energy Journal	0195-6574	563150979	International Association for Energy Econom	806402	fulltext@1982	no or silent	silent	563150979
2.3	selected	Energy Journal, The	0195-6574	44393051	JSTOR Archive Collections Complete	9332124	fulltext@1980-01-01^P4Y	no or silent	yes	563150979
2.4	selected	Energy Journal	0195-6574	563150979	Academic Search Ultimate (UMC)	customer.12	fulltext@1990-01-01	no license found	no license found	563150979
2.5	selected	Energy Journal, The	0195-6574	563150979	Materials Science & Engineering Database	27567894	fulltext@1998-01-01^2013-07-01	no license found	no license found	563150979
2.6	selected	The energy journal /International Association o	0195-6574	5585856	UMC Print Journals (UMC)	no KB ID	print@1980	no license found	no license found	563150979
3.1	selected	#Democracy : the internet and support for der	No Standard numb	896155769	Proquest Dissertations and Theses Global (U	896155769	ebook@2014	no license found	no license found	896155769
4.1	selected		291 2330-5983	448035264	Blue Mountain Project	448035264	fulltext@1915-04^1915-11	no license found	no license found	448035264
4.2	selected		291 1054-7193	448035264	JSTOR Arts & Sciences VIII Collection	3238533	fulltext@1915-03-01^1916-02-01	no or silent	yes	448035264
5.1	selected	1,001 Ways to Get Promoted	9781564144300	44955812	All EBSCO eBooks	251017	ebook@2000	no or silent	silent	44955812
6.1	selected	SYMPOSIUM: CANADIAN JOURNAL OF CONTIN	2154-5278	971917295	Portico Journals (UMC)	c1284.73	fulltext@2014^2016	yes	silent	971917295
7.1	selected	European Journal of Endocrinology	0804-4643	40801501	BioScientifica	806679	fulltext@1948	no or silent	yes	40801501
7.2	selected	European journal of endocrinology / European	0804-4643	29970781	UMC Print Journals (UMC)	no KB ID	print@1994	no license found	no license found	40801501
8.1	selected	Reproduction	1470-1626	60638423	BioScientifica	836325	fulltext@1996	no or silent	yes	60638423