

Querying for ACM Articles
by Forum Modi






This document serves as a guide on how to use *XML_scraping.py* and *queries.py* when trying to obtain an excel sheet of articles using keywords, conference venues, and year ranges in the ACM database. These scripts will only work with ACM's XML files which can be obtained through Craig Rodkin (rodkin@hq.acm.org), the current ACM Publications Operations Manager.

XML_scraping.py: Takes a folder of ACM's XML files and inserts the articles into a database

SKIP IF YOU ALREADY HAVE A DATABASE FILE






1. You will need to start with this script: Enter `python .\XML_scraping.py`
2. Enter the folder name with your XML files
3. Enter 0 if the files are proceedings or Enter 1 if the files are periodicals. (MAKE SURE TO ENTER THIS CORRECTLY!!)
4. You will be prompted to enter whether your XML files are in nested folders
 - a. Open up the folder with your XML files
 - b. If they look like this (not in nested folders), Enter 0

› This PC › Desktop › Research - LGBTQ Social Media › data

Name	Date modified	Type
 PROC-CHI00-2000-332040	7/11/2022 10:45 AM	XML Document
 PROC-CHI01-2001-365024	7/11/2022 10:45 AM	XML Document
 PROC-CHI02-2002-503376	7/11/2022 10:45 AM	XML Document
 PROC-CHI02-2002-507752	7/11/2022 10:45 AM	XML Document
 PROC-CHI03-2003-642611	7/11/2022 10:45 AM	XML Document

- c. If they look like this (nested folder), Enter 1

› This PC › Desktop › Research - LGBTQ Social Media › periodicals

Name	Date modified	Type
 JOUR-AJCD-V24I1-330409	7/9/2022 1:43 PM	File folder
 JOUR-AJCD-V24I2-337271	7/9/2022 1:43 PM	File folder
 JOUR-AJCD-V24I3-344599	7/9/2022 1:43 PM	File folder
 JOUR-AJCD-V24I4-353927	7/9/2022 1:43 PM	File folder
 JOUR-AJCD-V25I1-383948	7/9/2022 1:43 PM	File folder

5. Press Enter to fill database
 - a. It may take some time for the database to fill based upon the amount of files you have. I would recommend letting this run and working on something else for a while.
 - b. Should output a .db file with the same name of folder that will be used with *queries.py*

```
> python .\XML_scraping.py
<---Hello, welcome to the ACM database XML parsing script!--->

<---Please enter the folder name that includes your XML files--->
periodicals

<---Do you have proceedings or periodicals?--->
Selection Menu
0 - proceedings
1 - periodicals
1

<---Are your XML files in nested folders?--->
Selection Menu
0 - No(default)
1 - Yes
1
Press Enter to Start!
```

queries.py: Takes database file (named after folder) filled with ACM articles and outputs an excel file of article data with many user options such as searching for specific conferences, year ranges, and keywords

START HERE IF YOU ALREADY HAVE DATABASE FILE

1. Enter *python .\queries.py*
2. Enter database name (same as folder name in XML_Scraping.py) without the .db ending
 - a. Make sure it is in the same directory as queries.py
3. Enter the name of what you want the excel file to be called without the .xlsx ending

```
> python .\queries.py
<---Welcome to the database parsing script!--->

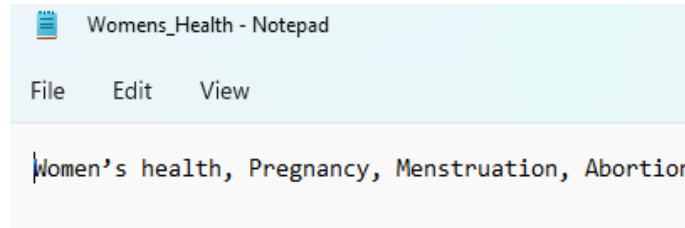
What is the database name (do not include .db)?
periodicals
Trying to enter database...
...Connected to database!

What would you like to call the excel file?
periodicals_example
```

4. Enter a numeric menu option

- a. **Enter 1 - Keyword Search:** Search for articles with keywords in the abstract, title, full text, and keyword list

- i. Enter the name of a text file without the .txt ending with keywords separated by commas. Keep in mind, searches include plurals

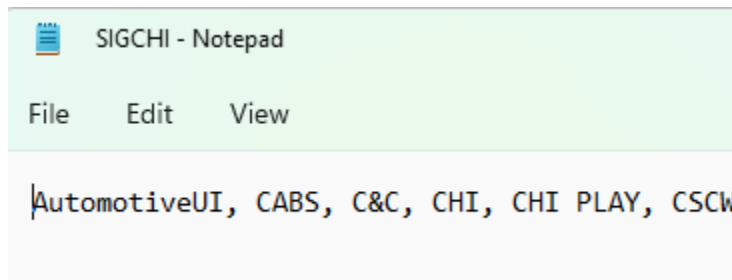


- b. **Enter 2 - Year Range:** Enter starting and ending year range

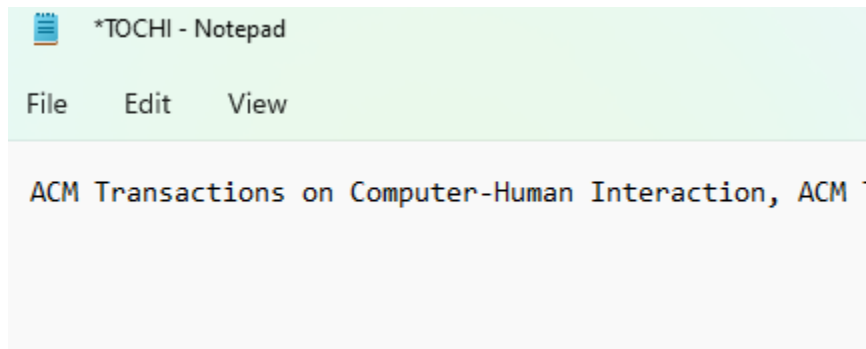
- c. **Enter 3 - Conference Search:** Search for articles from specific conference venues

- i. Enter 0 for proceedings, Enter 1 for periodicals

1. *Proceedings:* Enter the name of a text file without the .txt ending with **publication venue acronyms** separated by commas.



2. *Periodicals:* Enter the name of a text file without the .txt ending with **publication venue full names** separated by commas.



- d. **Enter 4 - Start Search:** Outputs excel file that meets searching criteria

- i. This may also take awhile depending on how large the database is and how intensive the search is. I would recommend working on something else while it runs.

```

<--Query Menu-->
Please enter the corresponding menu option!
1- Keyword Search
2- Year Range
3- Conference Search
4- Start Search
4
Press Enter to start the search!
Please be patient... starting the search
Creating periodicals_example...
...finished creating periodicals_example

```

5. You can input any combination of the three menu options before starting to search.

```

<---Search by Keywords--->
Please enter a text file with each keyword seperated by a comma.
Ex: hci, binary trees, algorithm

What is the textfile name?
Womens_Health
Keyword Search Set!

<--Query Menu-->
Enter 4 to start search or set another menu option.
1- Keyword Search
2- Year Range
3- Conference Search
4- Start Search
2
<--- Search by Year -->

What is the starting year?
2007
What is the ending year?
2017
Year Range Set!
Keyword Search Set!

<--Query Menu-->
Enter 4 to start search or set another menu option.
1- Keyword Search
2- Year Range
3- Conference Search
4- Start Search
3
<--- Search by Conference -->

Are you searching for proceedings or periodicals?
Enter the corresponding menu option
1 - Proceedings
2- Periodicals
2

Please enter the name of text file with each conference venue seperated by a comma.
Ex: ACM Transactions on Mathematical Software, ACM Transactions on Computer-Human Interaction

What is the textfile name?
TOCHI
Year Range Set!
Keyword Search Set!
Conference Search Set!

```