

User Manual for Library Metadata Harvester Application

I. Introduction

The Library Metadata Harvester (LMH) project accepts a list of either International Standard Book Numbers (ISBNs) or Online Computer Library Center (OCLC) Control Numbers (aka OCNs) to perform a priority search of the internet sources for related metadata. The retrieved metadata is stored in an SQLite database and output as a tab-delimited file. The internet sources include a variety of Application Programming Interfaces (APIs), Blacklight-based library catalogues, and WorldCat-based library catalogues. LMH also uses Z39.50 for searching through databases over a TCP/IP network which the Library of Congress supports. It also has an optional web-scraping that could be set up. Check the web-scraping section for further information.

II. Installation

- Prerequisites

LMH uses various Python libraries that are provided in the **requirements.txt** file on the GitHub page. PyInstaller installs all the requirements on its own, if you are not using PyInstaller refer to the requirements.txt for the libraries that you will need.

- Installation Steps

1. Make sure you have Python installed on your computer. You can test this by running the **'python -V'** command in your terminal
2. Make sure you have pip installed onto your computer. You can test this by running the **'pip --version'** command in your terminal

3. Navigate to where you downloaded the folder containing the `lmh.py` file in the terminal
4. Run the following command **'pip install pyinstaller'**
5. After PyInstaller has finished installing, run the following command while being in the directory that contains the `lmh.py` file **'python -m PyInstaller lmh.py --onefile --windowed --name LibraryMetadataHarvester'** this will generate a **dist** directory which will contain the Library Metadata Harvester executable.
6. Click on the **LibraryMetadataHarvester** executable to run the application

- Setting up Yaz-client for Z39.50 Search System

- Install the Yaz-client through the indexdata.com website
- The version used for this project is:

<code>yaz_5.34.0.exe</code>	2023-01-12	12:42	29M
-----------------------------	------------	-------	-----
- Run the installer and make sure to select the following options:
 - YAZ Runtime
 - YAZ Development
 - YAZ4J
 - YAZ Path
- Restart your system

III. Getting Started

- Overview of the GUI

ISBN / OCN - Select which input type you are using

Choose File - Select your input file

Output File - Name your output file (note that whether you do this or not the data will be saved in the LMH_database.db file)

Sources - Click on sources that you would like to use and move the sources with the highest priority to the very top one at a time, then press the Save Order button. The highlighted sources will be used in the search

Settings - In settings, you can select what data type to retrieve, change your API key, and add and remove YAZ and Web-scraping sources

- Setting up YAZ client

- In **Settings** click on **Change YAZ Client Path** and select the **yaz-client.exe** file that is located in **/YAZ/bin/yaz-client.exe**
- You can add/remove Z39.50 sources by clicking on **Change Z39.50 Settings** and then entering the displayed name of the source and the link to it

- Setting up Web Scraping

- Go to **Settings** then to **Change Web Scraping Settings**
- Enter the name of the source (how it will appear in the list)
Ex. PSU
- You can find the Query URL by searching for the ISBN/OCLC number in the search bar of the website and replacing the searched numbers in the URL with **?q={number}**
Ex. <https://catalog.libraries.psu.edu/catalog?q={number}>
- For Base URL enter the URL for the catalogue of the library
Ex. <https://catalog.libraries.psu.edu/catalog>
- Then click on **Add Source**
- Make sure the desired sources are selected (highlighted) before searching

- Setting up API Keys for Google API

- Note: Without the Google API key you are limited to 100 requests per minute and 1000 per day hence we disabled the search without the API key
- Please refer to the official Google Books API website to set up your API key.
The process is free and only requires a Google account

<https://developers.google.com/books/docs/v1/using>

- Understanding ISBN/OCLC Input File Format

Make sure your input file is properly formatted in the following manner:

1306059035

1298391601

1298393059

...

IV. Searching and Harvesting Metadata

- Choosing Search Priority

To choose and select the search priority click on the desired API and move it up and down based on the search priority with the top API being first and the bottom API being the last

- Initiating Metadata Search

Select your input file from the GUI and click on **Start Search**

- Monitoring Search Progress

Initially, there will be a handshaking process to check whether the requested sources are properly functioning. Once that step has been completed the search

will start. It might take up to a few hours depending on the sample size to finish the retrieval.

V. Troubleshooting And Best Practices

- Make sure the handshaking has been completed and the search has started before moving away from the screen
- If you think the progress is frozen check the `lmh_log.txt` file and also check the database to see when the last entry was added
- Make sure you are using OCN codes only for sources that work with OCN numbers such as LOC and Open Library API. Note that Z39.50 and Harvard API do not accept OCNs.