Overview

Jonathan Hooper, Leeds. May 2022

Contents

Introduction

This software was written to collect country affiliation data on authors of reading list citations. This is an extension of work done by Imperial College, London. *Reference 1*

Imperial's work used a single source, the Web of Science API Expanded, and retrieved data using DOIs. In our extension, we also included the sources VIAF (stronger than WoS for books) and Scopus (like WoS, article-focused, but also including some books and other resources). We also searched for resources by author and title where no DOI was available.

Approach in brief

For this phase of the project, we are not providing a simple end-user application, nor any website or database, and we are not harvesting data routinely for all reading lists.

We are providing a set of individual scripts for the various steps in the process, which are run by an operator in IT, against a relatively small set of reading lists. The output data is shared with the Library and Project staff.

The scripts are written in PHP. API interaction is over https with JSON the preferred format (although some APIs only offer XML). JSON is used as the temporary intermediate storage format and UTF-8-encoded CSV as the output format for the Library and Project staff.

The process could be modified or extended: e.g. the WoS integration could be left out if an organisation does not have a subscription to the API; additional integrations with other data sources could be coded in an analagous way to the existing ones; and data could be output in different formats.

Installing and using the software

The software repository is: https://dev.azure.com/uol-support/Reading%20Lists/_git/Citation%20enhancement

See 1_1_install_configure_and_run_the_software.pdf for instructions.

Although written for Leeds, this software could be used by other institutions, provided they use the Alma and Leganto, and have subscriptions to Elsevier's Scopus and the Web of Science and the WoS

*API Expanded (a WoS subscription alone is not sufficient). Configuration for different organisations is a question of populating a file **config.ini** with the organisation's own API keys for the various data sources.*

Possible further development work

- Obtain additional country information by looking up non-country affiliation data (e.g. "Toronto") in a suitable service (VIAF?) to identify the country (e.g. "Canada")
- Explore other data sources for authors e.g. Wikipedia, Google Scholar
- Improve error handling and reporting
- Store results in a database, and surface data in an interactive website for Library and academic staff

Further information

Information about specific aspects of the software is in separate documents:

- 1_install_configure_run/
  - 1_1_install_configure_and_run_the_software.pdf
  - 1_2_API_keys.pdf
  - 1_3_the_individual_scripts.pdf
  - 1_4_the_batch_script.pdf
  - 1_5_error_handling.pdf
- 2_data_integrations/
  - 2_1_interaction_with_Alma_and_Leganto.pdf
  - 2_2_searching_the_data_sources.pdf
  - 2_3_author_title_similarity_scores.pdf
  - 2_4_data_structure_in_citation_files.pdf
  - 2_5_encoding_and_special_characters
- 3_reports/
  - 3_1_export_scripts.pdf
  - 3_2_columns_in_the_reports.pdf
  - 3_3_the_short_report_export_script.pdf
  - 3_4_the_long_report_export_script.pdf
  - 3_5_rerun_export_without_refetching_data.pdf

References

1. Price, R., Skopec, M., Mackenzie, S. *et al.* A novel data solution to inform curriculum decolonisation: the case of the Imperial College London Masters of Public Health. *Scientometrics* **127,** 1021–1037 (2022). https://doi.org/10.1007/s11192-021-04231-3 *plus supplemental materials: https://osf.io/cyj2x/*

2. Country code and name mapping uses data from http://country.io/data/