Export scripts

Jonathan Hooper, Leeds. May 2022

Contents

Overview

The citation enhancement scripts save a large quantity of data in relatively non-human-readable JSON files. It is necessary to produce briefer, more human-readable reports summarising the key data.

CSV format is used because files can readily be opened in Excel for further processing by non-technical staff.

There are two scripts:

- **simpleExport.php** writes a short report, with a line-per-citation, for headline analysis
- **longExport.php** writes a longer report which may have multiple rows-per-citation

The primary goal of the short report is to produce a list of countries for each citation, and for the list as a whole: countries lend themselves to further numeric analysis, and visuals like pie-charts. The long report includes all geographical and affiliation data found in the sources, not just countries.

Most tutors are likely to be principally interested in the short report. Some may be interested to look at the long report, and a very few may be interested to see the raw JSON data from which the reports are generated.

Basic process

Both the two scripts (for the short and the long export) take JSON-encoded citations from STDIN, and loop over them collecting the data they are interested in. They then assemble and total this and write the output files directly (unlike the other scripts in the process, they don't write output to STDOUT).

There is a lot of business-logic coded in these scripts - selecting fields for the report; mapping country names and codes in source data onto standard values; collating data at citation- and list-level; and counting country occurrences.

The short report vs the long report

- The short report does not include as much citation-metadata as the long report *e.g. it only includes one title for each citation where the long report includes every title found*
- The short report always has one and only one row for each bibliographic citation in the list, whether or not any data was found in the external sources. The long report may have zero, one or more rows for each citation: zero if no data was found in any external source, and one for each bit of geographical data found for that citation
- The short report groups and summarises data at the citation-level, whereas the long report does not *e.g. the short report gives a citation-level author-title-similarity; the long report gives a specific similarity score for each row*
- The short report script also produces a summary report, which contains list-level data which cannot easily be incorporated into the (citation-level) short report
- The short report (optionally) also includes a second table below the main row-per-citation table: this lists countries found in the citations and gives an "author count" for each - this table could be directly visualised as a pie-chart
- The short report has a higher threshold for author-title similarity (80%) meaning data included is very likely to be a correct match to the reading list citation. The long report with a lower threshold (60%) may include a higher number of incorrect matches *(although it may also include correct matches that happen to have a low similarity score).*

Identifying citations in the reports

- Both the short and the long reports have a column CIT-NUMBER which identifies the citation and is consistent between the two reports, and the reading list
- In the short report:
  - Each CIT-NUMBER will only occur once
  - Where a number in sequence is missing, it is because the citation in the reading list is a Note rather than a bibliographic citation
- In the long report:
  - Each CIT-NUMBER may appear multiple times

    - Where a number in sequence is missing, it is **either** because no geographical data could be retrieved for that citation, **or** because the citation in the reading list is a Note and so not being analysed

- Citation numbers should always correspond to the order of the citation in the reading list in Alma/Leganto provided it is sorted **in tutor order**

Overview of data in the short report

- The short report has one and only one row per bibliographic citation
- This may include data from zero or one source (Scopus/WoS/VIAF)
- The order of preference for sources is:
    - *For article-like citations* (Leganto material type is CR/E_CR/JR):
        1. **Scopus**
        2. **WoS**
        3. **VIAF**
    - *For other citations*:
        1. **VIAF**
        2. **Scopus**
        3. **WoS**
- The first source in the preference list *either* with a similarity >= 80% *or* resulting from a DOI search is the one used
- Country data included from each source is:
    - Scopus:
        - **Contemporary affiliations** *(i.e. contemporary to time of publication)*
        - **Current affiliations** *only if no contemporary affiliation*
    - WoS:
        - **Addresses** *(whether tied to an individual author or not)*
        - *Reprint addresses and publisher addresses are ignored*
    - VIAF:
        - **Nationalities**
        - *Countries of publication are ignored*
        - *Locations and affiliations are ignored because they cannot be directly machine-translated to countries*
- Countries in the short report are standardised to English country names by passing through mapping tables in **Config/CountryCodes**

The long report contains richer data - from all sources, with all available fields from each source, and the similarity cut-off is 60% rather than 80%.

See also "The short report export script" for a detailed description of the generation of the short report.

Configuration

config.ini

- To export tab-delimited text files rather than CSV files, set Export.Format to TXT
- To remove the byte-order-mark in the output files, remove Export.BOM
- To omit the second table from the short reports (country counts), set Export.CountryCounts to 0

Country files

These are in the directory **Config/CountryCodes/**

Two key configuration files that can be edited are:

- **nameAlias.json**
  This maps non-standard names that are known to occur in the source data onto standard names *e.g. WoS lists "England", "Scotland", "Wales" and these are mapped in this file to "United Kingdom"*
- **iso2Alias.json**
  This maps non-standard or non-current 2-letter country codes onto suitable alternatives *e.g. "UK" occurs in some data and should be converted to "GB", and historical codes like "DD" (East Germany) can be converted to current ones like "DE" (Germany)*

It is recommended that the other files in this directory are not edited, but the alias files above are used instead where needed. These other files were downloaded from http://country.io/data/:

- **continent.json**
  Maps 2-letter country codes to 2-letter continent codes
  Not currently used in the process but might be useful in future?
- **iso3.json**
  Maps 2-letter country codes to 3-letter codes
  Reversed, and used in the export scripts to convert 3-letter codes to 2-letter codes
- **names.json**
  Maps 2-letter country codes to standard names
  Reversed, and used in the export scripts to convert names to 2-letter codes

World Bank GNI data

This is not currently used by the export scripts

*World Bank Gross National Income (GNI) data is in the directory **Config/WorldBank**, downloaded from https://data.worldbank.org/indicator/NY.GNP.PCAP.CD*

It was originally included to replicate the work of Imperial College where GNI rankings were used to calculate a score for each citation, allowing numerical comparison of reading list diversity https://osf.io/cyj2x/

Possible future improvements

How the reports are presented

CSV is not a perfect format for this, because it doesn't group hierarchical data (lists->citations->authors->affiliations) in a natural way:

- *Multiple affiliations require individual rows, with all of the parent data repeated (if we are to take advantage of sorting and filtering in Excel)*
- *A single report cannot easily contain module-level, list-level, and citation-level totals*

I haven't been able to find any suitable alternative format that doesn't require specialist knowledge or software to use.

The best alternative would probably be to surface the results in a web interface, allowing users to interactively sort, filter, expand and contract the data. If doing this, the harvested results could be stored in a MySQL database behind the website.

Making business-logic more transparent

Too much of the business logic is present in (opaque) PHP code - it would be better if mappings and decisions were stored in human-readable configuration files which could be shared with non-technical staff.

More detailed information

Columns in the reports

These are described in a separate document, 3_2_columns_in_the_reports.pdf

What the scripts do in detail

The scripts are described in separate documents:

- “The short report export script”
- “The long report export script”