# The long report export script

Jonathan Hooper, Leeds. May 2022

## Contents

Setup Collect options, config etc ..... 1
Stage 1: Loop over each citation in the input JSON data ..... 1
Stage 1a: Citation-level data ..... 1
Stage 1b: Scopus records ..... 2
Stage 1c: WoS records ..... 2
Stage 1d: VIAF records ..... 4
Stage 2: Write the output files ..... 4

## Setup

## Collect options, config etc

**longExport.php circa lines 87 - 121**

- Set threshold for including data in report (60% similarity)
- Get config from config.ini - output CSV or TXT, whether to include byte-order-mark (BOM)
- Collect country name/code mappings
- Define column headings to include in report (i.e. the keys of $outputRecord to include)

## Stage 1:

## Loop over each citation in the input JSON data

**longExport.php circa lines 123 - 576**

In this loop:

- The input citation is **$citation**
- In each iteration zero, one or more output records (**$outputRecord**) are created and appended to **$outputRecords** (the list of output records)
  *In the long report, unlike the short, we include each individual bit of geographical data for a citation in a separate row, and we do not include citations without any geographical data*

## Stage 1a:

## Citation-level data

**longExport.php circa lines 141 - 275**

The script collects Leganto and Alma data i.e. data which will be constant for this citation, regardless of specific data from Scopus, WoS and VIAF.

It is saved in a variable **$outputRecordBase** e.g.: **$outputRecordBase["CIT-AUTHORS"]** is an array of authors found in Leganto or Alma.

## Stage 1b:
## Scopus records

**longExport.php circa lines 280 - 371**

Where there is data from Scopus, we will have an array **$citation["Scopus"]["first-match"]["authors"]** which contains an entry for each author in the citation.

The code loops over this array:

- Creating a working record **$outputRecordScopus** from the citation-level $outputRecordBase and populating it with data from the Scopus citaton
  *e.g. putting the various forms of this particular author's name in the array* ***$outputRecordScopus["SOURCE-AUTHORS"]***
- Preparing a new output record **$outputRecord** from the working record
- If this author has a contemporary affiliation:
  - Adding location details to the output record
  - Appending the output record to the list of output records $outputRecords
  - Preparing a new output record **$outputRecord** from the working record
- If this author has a current affiliation:
  - Adding location details to the output record
  - Appending the output record to the list of output records $outputRecords
  - Preparing a new output record **$outputRecord** from the working record

*At the end of this process, we've added zero, one or more output records with* ***$outputRecord["SOURCE"] = "SCOPUS"****. How many we add depends on how many Scopus authors there are, and whether they have contemporary and/or current affiliations.*

## Stage 1c:
## WoS records

**longExport.php circa lines 280 - 371**

Where there is data from WoS, we will have an array **$citation["WoS"]["first-match"]["metadata"]["authors"]** which contains an entry for each author in the citation.

The code loops over this array:

- Creating a working record **$outputRecordWoS** from the citation-level $outputRecordBase and populating it with data from the WoS citaton
  *e.g. putting the various forms of this particular author's name in the array* ***$outputRecordWoS["SOURCE-AUTHORS"]***
- Preparing a new output record **$outputRecord** from the working record

- If this author has an address:
    - This particular address is flagged as "seen" so we don't also include it below
    - Adding location details to the output record
    - Appending the output record to the list of output records $outputRecords
    - Preparing a new output record **$outputRecord** from the working record

*Unlike Scopus, Wos also includes various bits of location information not tied to an individual author:*

For each **address**:

- If we've not already seen it when processing authors above:
- Add location details to the output record
- Source authors is empty for this record (because we don't have a named author)
- Append the output record to the list of output records $outputRecords
- Prepare a new output record **$outputRecord** from the working record

*This is required because most pre-2008 and some post-2008 WoS records store addresses without tying them to particular authors.*

For each **reprint address** (also called corresponding address):

- Add location details to the output record
- If we have a named author for this reprint address include it in source authors otherwise leave empty
- Append the output record to the list of output records $outputRecords
- Prepare a new output record **$outputRecord** from the working record

*NB the short export does not look at reprint addresses.*

For each **publisher address**:

- Add location details to the output record, also including any publisher name
- Source authors is empty for this record (because we don't have a named author)
- Append the output record to the list of output records $outputRecords
- Prepare a new output record **$outputRecord** from the working record

*NB the short export does not look at publisher addresses.*

*At the end of this process, we've added zero, one or more output records with **$outputRecord["SOURCE"] = "WOS"**. How many we add depends on how many WoS authors there are, whether they have addresses, and what other (non-author) geographical information is in WoS.*

## Stage 1d:
## VIAF records

**longExport.php circa lines 534 - 561**

Where there is data from VIAF, we will have an array **$citation["VIAF"]** which contains an entry for each author we searched for.

We loop over this array, and if we got usable data back there will be a value **$citation["VIAF"][N]["best-match"]**

If present, the code:

- Creates a working record **$outputRecordViaf** from the citation-level $outputRecordBase and populating it with data from the VIAF citaton
  *e.g. putting the various forms of this particular author's name in the array* ***$outputRecordWoS["SOURCE-AUTHORS"]***
- Prepares a new output record **$outputRecord** from the working record
- The best-match may have geographical data in "nationalities", "countriesOfPublication", "locations", "affiliations" - for each one the code:
  - Adds location details to the output record
  - Appends the output record to the list of output records $outputRecords
  - Prepares a new output record **$outputRecord** from the working record

*NB the short export does not look at countries of publication, or locations, or affiliations (the latter two are not directly machine-translatable to countries).*

*At the end of this process, we've added zero, one or more output records with* ***$outputRecord["SOURCE"] = "WOS"****. How many we add depends on how many WoS authors there are, whether they have addresses, and what other (non-author) geographical information is in WoS.*

## Stage 2:
## Write the output files

**longExport.php circa lines 580 - 663**

*This is simpler than for the short report because there is no summary report and no additional country-count table*

Looping over $outputRecords:

- A new report file is triggered by the function outFilename($outputRecord) returning a different a value from the previous iteration *(i.e. in practice this is from a new reading list)*
- When a new report file is required, any existing file handle is closed and a new one opened *(taking account of CSV|TXT and BOM options in config.ini)*
- Fields from **$outputRecord** that will be output are specified in **$rowHeadings** (see Setup above)
- In this process, scalars (e.g. 100 or "Y") are copied as-is

- Missing values are replaced with the empty string
- Arrays are flattened - at the top-level with pipe character, and then with semi-colon e.g.:
  **[ "A", "B", "C" ]** becomes **"A|B|C"** and
  **[ [ "GB", "IE" ], [ "DE" ], [ "AU", "DK" ] ]** becomes **"GB;IE|DE|AU;DK"**
- Rows with a similaraity below threshold (60%) are not output (threshold is set in Setup above)
    - *There is a clumsy fudge here because for Scopus and WoS the similairty we are interested in is the* best *one anywhere in this citation, not the particular one in this row, and we have stashed that in* ***$outputRecord["SIMILARITY-MAX-INDEX"]***

Finally we close off any open file handles.