Data structure in citation files

Jonathan Hooper, Leeds. May 2022

Contents

Overview

Citation enhancement consists of running a number of separate scripts, which progressively enhance the original reading list data with data from external sources.

The format used to temporarily save this data in-between steps in the process is JSON.

In general, a JSON file will exist in the directory **Data/** for each module being analysed. This file lists the citations in that module's reading lists (the module may have zero, one or more reading lists).

Description

The data structure in the JSON citations files is not centrally defined or enforced, and is simply based on what is convenient when extracting data from the various sources. Defining a standard data format, and a standard for the various data integrations, might be a future improvement.

Each JSON file contains an array of citation elements i.e. the file looks like:

{ *Citation*, *Citation*, *Citation*, ... }

Citation

Normally a **Citation** element represents an individual citation in a reading list:

- The exceptions to this are:
    - Where we try to collect information for a module that does not have a corresponding course in Alma
    - Where any corresponding course(s) in Alma do not have any associated reading lists
    - Where a associated reading list has zero citations
    - In those cases, a **single** place-holder **Citation** element will be present for the module, without any citation-level data
- A citation may simply be a note, without any bibliographic reference
    - In this case there will be a Citation element for that note, but without any data enhancement from Alma, Scopus, WoS or VIAF
    - *Notes can be distinguished from the value of Citation.Leganto.secondary_type.value ("NOTE")*

A citation element is an associative array with possible keys **Course**, **Leganto**, **Alma**, **Scopus**, **WoS** and **VIAF** i.e.:

```
{ "Course":Course, "Leganto":Leganto, "Alma":Alma, "Scopus":Scopus, "WoS":WoS, "VIAF":VIAF
}
```

- **Course** will always be present
- **Leganto** will *usually* be present (if a reading list with citations exists)
- The other keys will be present once a search has been attempted against that source for this citation

These top-level values in the Citation associative array are described below:

Citation.Course

Added to citation by: **getCitationsByModule.php**

Holds information about the **module**

Properties:

Property	Type	Present?	Meaning
modcode	String	Always	The short module code (e.g. **DEPT1234**) this citation is associated with *One JSON file may in principle contain citations from more than one module code.* *But in practice, we tend to keep to one file per module, and the batch script will always use one file per module.*
course_code	String	If a course exists in Alma for modcode	The course code this citation's reading list is associated with, as used in Alma *If not present, this means no course exists in Alma for this*

			modcode - and no further enhancement can or will be done
list_code	String	Always	The code for the associated reading list *This duplicates the entry in Citation.Leganto (below) because of the case where a list has zero citations, and so has no Citation.Leganto element*

Citation.Leganto

Added to citation by: **getCitationsByModule.php**

Holds information about an individual **reading list citation**. Will only be present if there is a course in Alma matching the modcode, **and** if that course has a reading list.

Properties:

Property	Type	Present?	Meaning
id	Integer	Always	The system-generated ID for this reading list
list_code	String	Always	The code for this reading list
list_title	String	Always	The title of the reading list. Unlike id and list_code, this can be changed by instructor/library and displays to users, so this may be a better way of referring to a reading list in user-facing reports
citation	Integer	Always	Position of citation in list (starting at 1) This should always correspond to ordering of citations in Alma and Leganto, when viewing in instructor-order. *Notes count as citations for the purposes of this numbering.*
section	String	Always	The name of the reading list section this citation belongs to *May be helpful in reports, especially for long reading lists*
secondary_type	Associative array	Always	The citation's type code and description as used in Alma e.g. **{"value": "BK", "desc": "Book"}** *Types are listed in Ex Libris's Reading List Loader documentation.*
metadata	Associative array	Always	Selected fields from the citation data returned by the Alma Courses API e.g. **title**, **author**, **mms_id**

			getCitationsByModule.php could be modified to fetch additional fields if required

Citation.Alma

Added to citation by: **enhanceCitationsFromAlma.php**

Holds information about a **bibliographic record** from Alma. Will only be present if there is a value in **Citation.Leganto.metadata.mms_id** with a corresponding bibliographic record in Alma.

Properties:

Property	Type	Present?	Meaning
titles	Array of associative arrays	Usually	Title-like data from the bibliographic record *(defined by Marc tags in enhanceCitationsFromAlma.php)*. May be more than one value e.g. title plus alternate title. And may be alternative representations (e.g. in Chinese characters) from linked 880 field. For each title: **title.tag** holds the Marc (or linked Marc) tag *e.g. "245"* **title.originalTag** may hold "880" if the data comes from a linked field **title.a** holds the value of the $a subfield if present *etc* **title.collated** holds the values of the $a $b subfields, joined with spaces
creators	Array of associative arrays	Usually	Author-like data from the bibliographic record (defined by Marc tags in enhanceCitationsFromAlma.php). May be more than one value e.g. author plus additional author. And may be alternative representations (e.g. in Chinese characters) from linked 880 field. For each creator: **creator.tag** holds the Marc (or linked Marc) tag *e.g. "100"* **creator.originalTag** may hold "880" if the data comes from a linked field **creator.a** holds the value of the $a subfield if present *etc* **creator.collated** holds the values of the $a $b $c $d $q subfields, joined with spaces
ids	Associative array	Often	Identifiers (ISBNs, ISSNs, LCCNs) found in the bibliographic record

			e.g. ***ids.isbn*** *holds an array of ISBN values*

Citation.Scopus

Added to citation by: **enhanceCitationsFromScopus.php**

Holds information about **Scopus searches and results** for a citation.

Properties:

Property	Type	Present?	Meaning
rate-limit	Associative array	Usually	Rate-limit data from the API, with an entry for each type of query attempted with this citation. For each query-type: **limit** = maximum number of allowed requests per week **remaining** = number of allowed requests remaining this week **reset** = timestamp of start of next week period *NB when any of the* ***remaining*** *figures reach zero, we will no longer be able to harvest data until the start of the next week (i.e. the time in the* ***reset*** *timestamp)*
searches-no-results	Array	Yes (may be empty)	Scopus API queries that returned no results for this citation, in the order they were attempted *The script tries progressively broader searches until it gets a result - see 2_2_searching_the_data_sources.pdf*
search-active	String	If successful search	The first Scopus API query that retrieved at least one result
search-pref	Integer	If successful search	The position of the first successful query in the order-of-preference – a lower number indicates a narrower search
result-count	Integer	If successful search	The number of records returned by the first successful query *NB currently the code is only analysing the first result - it does not attempt to determine whether a result lower-down the set may be a better match to the reading list citation*
results	Array of associative arrays	If successful search	Brief summary of all the records returned by the first successful search, using selected data from the API response

first-match	Associative array	If successful search	Data from the API response for the first record in the result set from the first successful search - this is the record that will be used in subsequent processing. *See Citation.Scopus.first-match description below*

Citation.Scopus.first-match

Added to citation by: **enhanceCitationsFromScopus.php**

Holds information about the **Scopus record** that will be used in subsequent processing.

Properties:

Property	Type	Present?	Meaning
summary	Associative array	Always	Core metadata for this record, selected by the script from the full Scopus record (id, title, creator, publication, doi, document type).
self	String [URL]	Yes	A URL identifying this record in the Scopus API
authors	Array	Always (?)	Data on the authors of this citation i.e. **[Author, Author, Author, ...]** *See Citation.Scopus.first-match.Author description below*

Citation.Scopus.first-match.Author

Added to citation by: **enhanceCitationsFromScopus.php**

Holds information about an individual **author** from the Scopus record that will be used in subsequent processing. Typically a single citation may contain several authors.

Properties:

Property	Type	Present?	Meaning
@auid	String [numeric]	Always	Author ID in Scopus
author-url	String [URL]	Yes	A URL identifying this author in the Scopus API
preferred-name	Associative array	Always (?)	The preferred author-name *The similarity calculation may try various forms of name against the reading list authors, and the export script will decide which form(s) to use in the report*
ce:initials, ce:surname, ce:indexed-name	String	Always (?)	Author-name data *The similarity calculation may try various forms of name against the reading list authors, and the export script will decide which form(s) to use in the report*

affiliation	Associative array	Often	The author's **contemporary** affiliation (i.e. at the time of the publication), broken into subfields ***affiliation.country*** *generally contains standard English-language country names, but other forms may appear.* *No conversion/translation is done in saving this data from Scopus, it is up to the export scripts to handle this.*
affiliation-current	Associative array	Often	The author's **current** affiliation (i.e. at the time of data extract), broken into subfields *NB the subfield structure of affiliation and affiliation-current are slightly different, because of the different way the data is receved from the API.* ***affiliation-current.address.@country*** *generally contains ISO-3-letter country codes, but other forms may appear.* *No conversion/translation is done in saving this data from Scopus, it is up to the export scripts to handle this.*
similarity-title	Integer	Always	Scores 0-100 giving string-similarity between title as appears in reading list and title as appears in Scopus record Value will be identical for all authors in a given citation (because all Scopus authors belong to the same Scopus citation). *See 2_3_author_title_similarity_scores.pdf for details of Similarity calculation*
similarity-author	Integer	Always	Scores 0-100 giving string-similarity between author as appears in reading list and author as appears in Scopus record Unlike similarity-title, will vary between authors in a citation. *See 2_3_author_title_similarity_scores.pdf for details of Similarity calculation*
search-parameters	Associative array	Always	The data values (DOI, TITLE etc) extracted from Leganto/Alma, to be used in searching Scopus
extra-parameters	Associative array	Always	Data values extracted from Leganto/Alma that are **not** used in searching Scopus, but *may* be used by the script to calculate similarity

Citation.WoS element

Added to citation by: **enhanceCitationsFromWos.php**

Holds information about WoS searches and results for a citation.

Properties:

Property	Type	Present?	Meaning
rate-limit	Associative array	Usually	Rate-limit data from the API. **x-rec-amtperyear-remaining** = searches remaining this year **x-req-reqpersec-remaining** = searches remaining this second *NB if either figure reaches zero, we will no longer be able to harvest data in that period.* *I assume the per-year quota will be reset when the annual subscription is renewed (c. April 2023?)*
searches-no-results	Array	Yes (may be empty)	Searches using the WoS API Expanded query syntax that returned no results for this citation, in the order they were attempted *The script tries progressively broader searches until it gets a result - see 2_2_searching_the_data_sources.pdf*
search-active	String	If successful search	The first search that retrieved at least one result, using the WoS API query syntax
search-pref	Integer	If successful search	The position of the first successful search in the order of preference of search strategies – a lower number indicates a narrower search
result-count	Integer	If successful search	The number of records returned by the first successful search *NB currently the code is only analysing the first result - it does not attempt to determine whether a result lower-down the set may be a better match to the reading list citation*
results	Array of associative arrays	If successful search	Brief summary of all the records returned by the first successful search, using selected data from the API response
first-match	Associative array	If successful search	Data from the API response for the first record in the result set from the first successful search - this is the record that will be used in subsequent processing. *See Citation.WoS.first-match description below*

Citation.WoS.first-match

Added to citation by: **enhanceCitationsFromWos.php**

Holds information about the WoS record that will be used in subsequent processing.

Properties:

Property	Type	Present?	Meaning
metadata	Associative array	Always	Core metadata for this record, selected by the script from the full WoS record (id, source, title, year, doctype, publisher, authors, addresses etc). Particular individual entries to note: • **authors** is an array of individual authors for this citation i.e. **[Author, Author, Author...]** *See Citation.WoS.first-match.metadata.Author description below* • **addresses** is an array of individual addresses for this citation i.e. **[Address, Address, Address...]** *See Citation.WoS.first-match.metadata.Address description below* • **reprint_addresses** is an array of individual reprint (corresponding) addresses for this citation i.e. **[Reprint_address, Reprint_address, Reprint_address...]** *See Citation.WoS.first-match.metadata.Reprint_address description below*

Citation.WoS.first-match.metadata.Author

Added to citation by: **enhanceCitationsFromWos.php**

Holds information about an individual author from the WoS record that will be used in subsequent processing. Typically a single citation may contain several authors. Where an author is linked to a specific address (generally post-2008) also includes author's address - otherwise all addresses are in CITATION.WoS.first-match.metadata.addresses (see below).

Relevant properties:

Property	Type	Present?	Meaning
addr_no	Integer | String [space-separated Integers]	Often (post-2008)	Identifies the address(es) in CITATION.WoS.first-match.metadata.addresses for this author *If Integer identifies a single address e.g. 12 -> Address # 12*

			If String list of Integers identifies a list of addresses e.g. e.g. "2 3 5" -> Addresses # 2, 3, 5
last_name, display_name, full_name, wos_standard, first_name	Strings	Usually	Various forms of author's name *The similarity calculation may try various of these against the reading list authors, and the export script will decide which form(s) to use in the report*
preferred_name	Associative array	Always (?)	full_name, last_name, middle_name, first_name *The similarity calculation may try various of these against the reading list authors, and the export script will decide which form(s) to use in the report*
addresses	Array of associative arrays	Often (post-2008)	Copy (for conveinience) of the address from Citation.WoS.first-match.metadata.addresses which matches this author (where available). Array of addresses, each one an associative array of different elements *address.country generally contains standard English-language country names but other forms may appear - no conversion/translation is done in saving this data from WoS, it is up to the export scripts to handle this*
similarity-title	Integer	Always	Scores 0-100 giving string-similarity between title as appears in reading list and title as appears in WoS record Value will be identical for all authors in a given citation (because all WoS authors belong to the same Scopus citation). *See 2_3_author_title_similarity_scores.pdf for details of Similarity calculation*
similarity-author	Integer	Always	Scores 0-100 giving string-similarity between author as appears in reading list and author as appears in WoS record Unlike similarity-title, will vary between authors in a citation. *See 2_3_author_title_similarity_scores.pdf for details of Similarity calculation*

Citation.WoS.first-match.metadata.Address

Added to citation by: **enhanceCitationsFromWos.php**

Holds information about an individual address from the WoS record. Typically a single citation may contain several addresses: Citation.WoS.first-match.metadata.Address always contains **all** addresses in the citation, whether or not individual authors are linked to them.

Properties:

Property	Type	Present?	Meaning
names	Array of associative arrays	Usually	Authors associaited with this address *NB We do not refer to this data and instead come in the other direction, from the authors array, and from there to linked addresses*
address_spec	Associative array	Always	Details of this address (including country) *NB address_spec.addr_no forms the link with Author.addr_no.* *There may be a many-to-many mapping between authors and addresses. There may also be addresses that are not linked to any individual author.*

Citation.WoS.first-match.metadata.Reprint_address

Added to citation by: **enhanceCitationsFromWos.php**

Holds information about an individual reprint address (corresponding address) from the WoS record.

We are currently not making use of these addresses in the reports we generate, because we understand they do not add to the information in Addresses (unless the reprint address is a purely administrative one).

Properties:

Property	Type	Present?	Meaning
names	Array of associative arrays	Usually	Authors associaited with this address
address_spec	Associative array	Always	Details of this address (including country)

Citation.VIAF element

Added to citation by: **enhanceCitationsFromViaf.php**

Holds information about VIAF searches and results for a citation.

Important differences between VIAF, and Scopus/WoS results

- *Scopus and WoS are **citation**-databases where VIAF is a **person**-database:*
- *With Scopus and WoS, a **single** search is done at the **citation-level***
- *With VIAF, a **separate** search is done **for each author** in the reading list citation*

- *Citation.VIAF is an **array** of datasets, each one for a separate search for a different author*
- *Citation.Scopus and Citation.WoS are each single datasets for a citation - but each of these single datasets may contain data for multiple authors*
- *Different VIAF authors for a citation will typically have **different title-similarity scores***
- *Different Scopus and WoS authors will have the **same** title-similarity*
- *With Scopus and WoS, the code currently just takes the first record from a result set - with VIAF, the code finds the record which best matches by title the reading list citation*

Properties of each individual entry in the VIAF element:

Property	Type	Present?	Meaning
searches	Associative array	Always	How the VIAF search is constructed - author and title will be filled in from search-term-au and search-term-ti *(see below)*
search-pref	Integer	If successful search	The position of the first successful search in the order of preference of search strategies – a lower number indicates a narrower search *e.g. a search-relation "any" in searches.AU or searches.TI is a **broader** search than a search-relation "all", and will have a **higher** search-pref*
data-source	String	Always	Where the author/title comes from to carry out the search against VIAF *i.e. Alma | Leganto | Scopus | WoS*
records	Integer	If successful search	The number of records returned by the first successful search *NB For VIAF, unlike Scopus and WoS, the code looks at all the returned records and decides which is the best match on the basis of title-similarity*
results	Array of associative arrays	If successful search	List of author names returned by the first successful search, using selected data from the API response *Only one of these will be the author selected by the script, on the basis of title-similarity*
best-match	Associative array	If successful search	Data from the API response for the record in the result set from the first successful search which best matches (by title) the reading list citation. This is the record that will be used in subsequent processing. *See **Citation.VIAF.best-match** description below*

Citation.VIAF.best-match

Added to citation by: **enhanceCitationsFromViaf.php**

Holds information about a VIAF record that will be used in subsequent processing.

Properties:

Property	Type	Present?	Meaning
type	String	Always	The integration script filters this to only include entries with value "Personal" *Some searches will include results for institutions etc, and these need filtering out*
heading	String	Always	A definitive form of the personal name *The script prefers the form from Library of Congress, otherwise it takes the first one found*
headings-all	Array of Strings	Always	All the forms of the personal name present in the record *The $a $b $d $q sub-fields of the VIAF data, joined with spaces*
about	String [URL]	Always	A URL identifying this record in the VIAF API
similarity-title	Integer	Always	Scores 0-100 giving string-similarity between title as appears in reading list, and a title associated with this author in VIAF *VIAF lists works associated with an author, and we compare each entry in this list with each form of the title in Leganto/Alma* *See 2_3_author_title_similarity_scores.pdf for details of Similarity calculation*
similarity-author	Integer	Always	Scores 0-100 giving string-similarity between author as appears in reading list and author as appears in VIAF record *See 2_3_author_title_similarity_scores.pdf for details of Similarity calculation*
best-matching-title	String	Usually	The title associated with this author in VIAF that is the most similar to a title in our reading list
titles	Array of Strings	Usually	All the titles associated with this author in VIAF
nationalities	Array	Often	Nationalities of this author

			Generally 2-letter ISO codes (e.g. "GB") but other forms appear, including foreign-language country names *No conversion/translation is done in saving this data from VIAF, it is up to the export scripts to handle this*
countriesOfPublication	Array	Often	Countries of publication for this author *Not currently used by export scripts*
affiliations	Array	Sometimes	Author affiliations (taken from 551$a fields in VIAF) *Does not include countries, and so currently not used by export scripts - a future improvement to the process would be to make use of this data*
locations	Array	Sometimes	Author locations (taken from 510.2_$a fields in VIAF) *Does not include countries and so currently not used by export scripts - a future improvement to the process would be to make use of this data*