Searching the data sources

Jonathan Hooper, Leeds. May 2022

Contents

Overview

We attempt to find resources in the external sources (Scopus, WoS, VIAF) that match our reading list citation.

The search query run against the external source may use data from the various fields in the Leganto record, and any linked Alma bib record, and resources already retrieved from other sources.

The different data sources have different query syntaxes, and offer different options. But broadly, searching is carried out with some combination of DOI, author, title, journal, year, ISBN/ISSN. The general approach is to start with a DOI-search, and then to try progressively broader searches using the other fields, until a match is found.

Data used to construct the search queries

Scopus

- DOI from Leganto citation
- Article title from Leganto citation
- Author/chapter author/editor from Leganto citation
- ISBN/ISSN from Leganto citation
- Material type from Leganto citation

No data from other sources (e.g. Alma bib record) is used to construct the search, although the author and title from the Alma bib-record are included in the similarity calculation.

WoS

- DOI from Leganto citation
- Article title from Leganto citation
- Author/chapter author/editor from Leganto citation
- ISBN/ISSN from Leganto citation
- Publication date from Leganto citation

No data from other sources (e.g. Alma bib record) is used to construct the search, although the author and title from the Alma bib-record are included in the similarity calculation.

VIAF

We use the first of the following groups that can be found:

From a linked Alma bib record:

- Creators (100 and 700 fields plus any 880 linked to one of them)
- Either the 245 field, or in its absence the first of the other title fields found (210, 240, 242, 243, 245, 246, 247, 730, 740 plus any 880 linked to one of them)

Typically there may be more than one creator. Only one title will be used in the search. By preference this will be the 245 field. Additionally, for each author and for the title, two versions are kept, called "a" and "collated": "a" contains just the $a subfield and "collated" contains the $a $b subfields (titles) or $a $b $c $d $q subfields (creators) joined with spaces.

From a previously found Scopus record:

- Authors
- Title

*Typically there may be more than one author. For each, two versions are kept, called "a" and "collated" - "a" generally contains the **indexed-name** from Scopus, and "collated" generally contains **surname, given-name**, but see the source code of **enhanceCitationsFromViaf.php** for all possible scenarios.*

There will be only one title, and it will only be kept as a "collated" version.

From a previously found WoS record:

- Authors
- Titles

*Typically there may be more than one author. For each, two versions are kept, called "a" and "collated" - "a" generally contains the **display_name** from WoS, and "collated" generally contains the **wos_standard** name from WoS, but see the source code of **enhanceCitationsFromViaf.php** for all possible scenarios.*

There will be only one title, and it will only be kept as a "collated" version.

From Leganto:

- Author and editor
- Either article title (for articles) or title

There will only be one author and one editor, so at most two creators, and they will only be kept as "collated" versions. There will only be one title, and it will only be kept as a "collated" version.

Meta-characters and reserved words in queries

Some special characters and reserved words may cause problems in some queries. Some of these are being escaped or removed, but further work may improve the quality of results. Changes currently made to the data fields before constructing search query are:

Scopus

- Double-quote characters " are removed from titles

WoS

- Double-quote characters " are removed from all search fields
- Parentheses ( ) are removed from all search fields except DOI

VIAF

- Double-quote characters " are removed from authors and titles

Start search narrow, then broaden

In the Scopus, WoS and VIAF integrations, the scripts start by attempting an exact match on DOI. After that, they look for an exact match on multiple fields in the original citation metadata, and then try progressively broader searches until they get at-least one result from the API.

This approach is used to maximise the chance that correct matches are identified, while still allowing for matches despite data-discrepancies:

- For example with a common author like "Smith, J" we want to start with narrow search, with confirmation from multiple data fields, rather than initially trying a broad search.
- But with less common names and titles we want the flexibitlity to find a match, even if the data in our reading list does not quite match the external source.

The order of searches attempted is:

Scopus

1. Exact match on a DOI found in the Leganto citation
2. Exact match on title and author surname and doctype and ISBN/ISSN
3. Exact match on title and doctype and (author surname or ISBN/ISSN)
4. Exact match on title and (author surname or ISBN/ISSN)
5. Approximate title search and exact match on author surname and doctype and ISBN/ISSN
6. Approximate title search and exact match on author surname and ISBN/ISSN
7. Approximate title search and exact match on (author surname or ISBN/ISSN)

WoS

1. Exact match on a DOI found in the Leganto citation
2. Exact match on title and author surname and publication year and ISBN/ISSN
3. Exact match on title and author surname and ISBN/ISSN
4. Exact match on title and (author surname or ISBN/ISSN)
5. Exact match on title and publication year

VIAF

1. All-words match in mainHeadingEl for ("collated") author, and all-wards match on ("collated") title
2. All-words match in mainHeadingEl for ("a") author, and all-words match on ("a") title
3. All-words match in mainHeadingEl for ("collated") author, and any-words match on ("collated") title
   NB there will be no need to also try an any-words match on "a" title, since all the words in "a" have already been tried in "collated"
4. All-words match in mainHeadingEl for ("a") author, and any-words match on ("collated") title
5. All-words match in personalNames for ("collated") author, and any-words match on ("collated") title
6. All-words match in personalNames for ("a") author, and any-words match on ("collated") title

NB Unlike Scopus and WoS, separate searches will be done against VIAF for each author in the citation.

Multiple results from search

If a search returns more than one record:

Scopus and WoS

We just take the first record in the result set (although we do record brief details of the others in the downloaded data)

VIAF

VIAF returns person-records rather than citations, and here we go through them all and find the entry with the highest similarity between a title in the reading list and a title of a work by that author in VIAF

Similarity scores

In order to estimate whether a record in an external source is a correct match to a reading list citation, the scripts calculate and record similarity scores for author and title (see “Author-title similarity scores”).

This calculation may be based on additional data, not used to construct the search. e.g. in the Scopus integration, a good match between Alma bib record title and Scopus title will be recorded rather than a poorer match between Leganto title and Scopus title, even though the Alma title is not part of the original Scopus search. And in the VIAF search, all the title fields in the Alma record (245, 210, 240, 242, 243, 246, 247, 730, 740, linked-880) will be tried against the VIAF title data to find the best match - not just the 245 that was used for the initial search of the VIAF API.

For Scopus and WoS, this calculation has no impact on the search strategy - we simply record the first record from the first search that yields results, regardless of similarity. *A future enhancement might use the similarity to decide which of a set of results is best, or whether and how to modify the search.*

For VIAF, the title similarity is used to decide which of the set of returned author records is the preferred match to the author in our reading list citation.

These scores will be used in the export step to decide whether to include that data.