How to extract thesis data from DSpace for entry into Wikidata via OpenRefine

Extract from DSpace

Log in to your DSpace instance with an administrator account.

Browse to your thesis community or collection(s).

In the sidebar, click “Export Metadata” and wait a little. When the CSV is ready, Save to disk (don’t open directly in Excel).

Import into Excel

If you open the CSV directly in Excel, it may not recognise Unicode characters. (**Note:** This may be fixed in more recent versions, or you may be using something like LibreOffice which is better behaved. If so, go for it and skip to the next section.)

If you import it in the Approved Way, it will break on every line break in every thesis abstract.

So you’re going to need a little pre-processing. Unless you have a preferred other text editor, I recommend downloading Notepad++ from https://notepad-plus-plus.org/ and carrying out the following steps:

- Open your CSV file in Notepad++ and Save As a new name
- Go to Encoding > Convert to UTF-8-BOM
- Search > Replace and tick the “Wrap around” option and “Extended” search mode (**Pro tip:** if you have to do this for multiple documents, open them all in separate tabs and use the “Replace all in all opened documents” button!)
    - Replace `\r\n` with ` ` ie a space
    - Replace `\n\n` with ` ` ie a space
    - Replace `\n""` with ` ""` ie a space plus ""
    - Replace `\n"` with something that won’t appear anywhere else, eg `%^&`
    - Replace `\n` with ie nothing
    - Replace `%^&` with `\n"`
- Save!

Now in Excel:

1. Create a new blank workbook
2. Data > Get External Data > From Text
3. Select your edited CSV file > Import
4. Tick “Delimited”, “65001: Unicode (UTF-8)”, and “My data has headers”
5. On the next page untick Tab and tick Comma
6. On the final page, select all the columns in the data preview, then tick Text
7. Finish
8. Save!

Your data should now be in beautiful columns in an Excel spreadsheet. Skim through to make sure that nothing looks wonky, eg text in the id field.

Tidy the data

General notes:

- If you select all and go to Home > Sort & Filter > Filter, then it's easy to use the dropdown menu on each column to sort alphabetically and/or scan the values to see if there's any highly unexpected data, eg a number in an author column, or Journal Article in a type column that should only contain Thesis and Dissertation.
- Where you've got multiple columns eg dc.title, dc.title[en], dc.title[en_US] they should be merged into a single column. Eg create a new column D with formula along the lines of =CONCATENATE(A2,"||",B2,"||",C2)
- **Save early, save often.** ☺

At a minimum you'll need to:

- Delete columns with confidential data, eg authors' email addresses, processing notes etc
- Delete rows referring to confidential items, eg a thesis so confidential you can't even make the title public

Then do as much/as little other tidying as you have time for. As a guide for the ideal case:

Required columns

- **id**
  - The Wikidata folk don't need this but keeping it in the dataset makes it possible for you to re-import back the Wikidata "Q" number when we're done as eg dc.identifier.wikidata. You can then use this to link to Wikidata and/or to identify which records have(/n't) been uploaded if you do future updates
- **dc.title**
  - This should be in sentence case – if you have titles in all-caps try running them through https://titlecaseconverter.com/ but watch out for proper nouns
- **thesis.degree.level**
  - (or other field that shows whether it's eg a master's, doctoral, etc)
- **dc.contributor.author**
- **dc.date.***
  - Pick the most sensible date for when the thesis was first submitted/deposited. For born digital ones this might be dc.date.accessioned, or it might be dc.date.issued or dc.date.copyright
- **dc.identifier.uri**
  - If this isn't a handle, then include the handle in a separate column
  - If you have a DOI, include that in a separate column
- **Current institution**
  - current name of institution. Add this as a new column and fill down if you don't have it already.

Strongly desirable

- **dc.language / dc.language.iso**
- **ORCID numbers**
- **Original institution**
  - if you have it / want it - name of institution at time of publication. This might be in thesis.degree.grantor, or dc.publisher

Optional / nice to have

- **Subject keywords** eg dc.subject.*
    - Controlled vocabulary is especially nice to have! This could include ANZSRC fields of research
- **Number of pages** eg dc.format, dc.format.extent
- **Copyright/Creative Commons statuses** eg dc.rights.*
- **Degree name and/or department** eg thesis.degree.name, thesis.degree.discipline etc
- **Supervisor** eg dc.contributor.advisor, lu.thesis.supervisor (this doesn't go in the thesis record itself but can be added to an author record)
- **Embargo end dates**

Not wanted - delete

- Abstracts - Wikidata doesn't have a place for it, there are potential copyright issues, and it also avoids potential embargo issues
- Any other fields that aren't relevant.

And check again that you've deleted:

- Columns with confidential data, eg authors' email addresses, processing notes etc
- Rows referring to confidential items, eg a thesis so confidential you can't even make the title public

1 August 2024