Increasing Visibility and Discoverability of Electronic Theses and Dissertations Using Linked Open Data:

A Simple Process for Uploading Metadata to Wikidata

***Steven J. Baskauf**, Data Science and Data Curation Specialist, Vanderbilt University Libraries, ORCID: 0000-0003-4365-3135 and*
***Shenmeng Xu**, Librarian for Scholarly Communications, Vanderbilt University Libraries, ORCID: 0000-0001-8475-0746*

NUTRITION INFORMATION

Wikidata is a freely available knowledge graph that, like Wikipedia, can be edited by anyone. It is multilingual, supports Linked Open Data as a mechanism for discovery and exploration, and is in common use worldwide. Because Wikidata is now commonly used as a data source by data aggregators like Google, it can be a tool for making metadata about electronic theses and dissertations (ETDs) more open and easily discoverable. Wikidata was developed in part to make it easier to provide citations for Wikipedia, so including theses in Wikidata also lowers the barriers for citing them as sources in Wikipedia articles.

Wikidata has grown beyond its original purpose and is now widely used by a well-developed user community committed to open data. That community is continually developing new products based on data from Wikidata and is working towards making it easier for new members of the community to contribute. This recipe describes a way to lower the barrier for entry to increase participation in the Wikidata community.

One reason for Wikidata's popularity is its very easy-to-use human interface. However, uploading large numbers of items can be very time consuming and labor intensive. This recipe will walk cooks through an alternative mechanism for creating Wikidata items using spreadsheets and a tool called VanderBot. Unlike similar tools for uploading from spreadsheets, VanderBot also makes it possible to delete or change large numbers of statements and references after they have been uploaded.

LEARNING OUTCOMES

By following this recipe, cooks will:

- gain practical experience with Linked Open Data by incorporating spreadsheet data into a well-known knowledge graph; and
- learn general principles of how Wikidata models and describes publications through preparing and uploading ETD metadata.

NUMBER SERVED

This recipe can be used to upload metadata for ETDs to serve Internet users around the world and provide greater exposure for thesis authors —with no expiration date! The number of ETDs uploaded can range from less than a hundred into the thousands. The upward limit is really based on the number of rows of a spreadsheet with which a human can reasonably interact.

COOKING TIME

Depending on how many software applications need to be downloaded and your familiarity with installing software, plan for fifteen minutes to an hour to prepare those ingredients.

Getting a Wikimedia account and preparing the bot password should take about fifteen minutes for someone familiar with using a text editor. Cooks that are unfamiliar with Wikidata should probably spend at least an hour exploring it before undertaking the recipe.

After the necessary software is installed, the cooking time will depend largely on the amount of time necessary to prepare the data. If the system that manages the ETD content provides clean CSV exports, the time will mostly depend on how long it takes to copy and paste the exported data into the spreadsheet (see Cooking Method section, step 3). That may take fifteen minutes to an hour depending on how clean the data are and the cook's familiarity with using spreadsheets.

The time required to do the actual upload is only 1.25 seconds per thesis.

DIETARY GUIDELINES

Wikipedia has become increasingly strict about requiring citations to support assertions made in its articles. Once a thesis is included in Wikidata with its supporting references, it becomes easy to cite in Wikipedia. Each item in Wikidata has a unique Q ID, an identifier that can be used in Wikipedia's {{Cite Q}} template to automatically generate a citation of the thesis. Wikidata is also commonly used as a data source by large information aggregators—to support Google's Knowledge Graph, for example—so including theses in Wikidata improves their discoverability.

INGREDIENTS AND EQUIPMENT

For help with any of the tasks in the following sections, see the Chef's Notes section. Refer to the figure to see how the ingredients blend together.

To be successful in completing this recipe, cooks should be familiar with their computer's file system and should have a basic knowledge of giving commands in their computer's console.

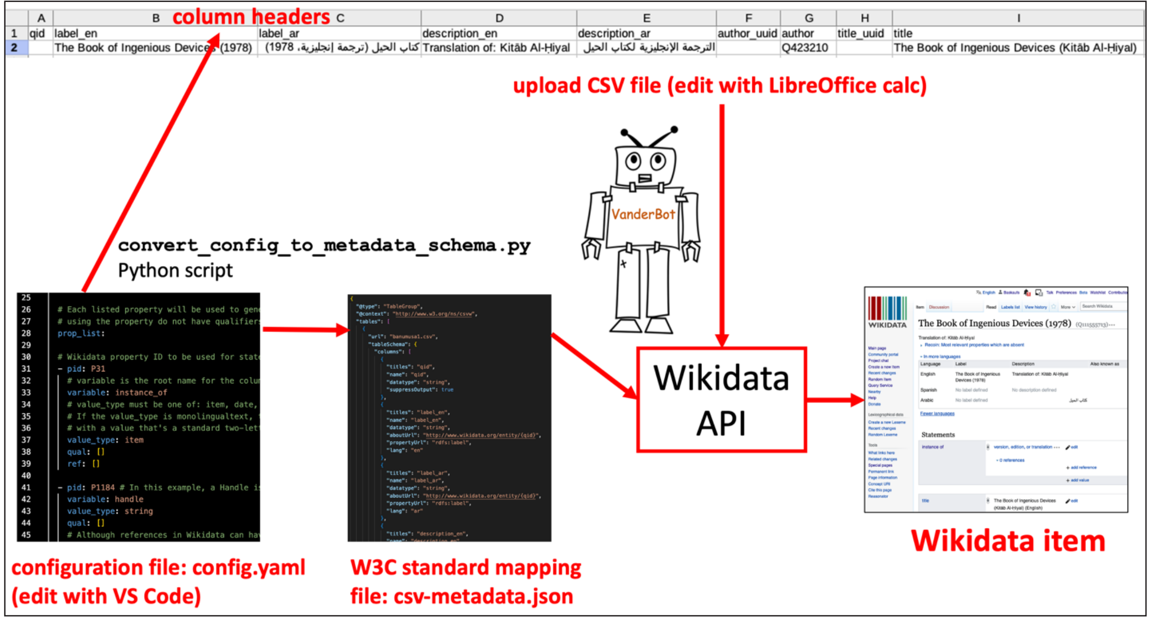
Before starting this recipe, make sure that the computer has Python 3 installed. Although the recipe does not involve coding in Python, the Python application is necessary to run the VanderBot script. Cooks may also need to install the Python requests library.

A spreadsheet editor that works well with CSV files is also required. Although Excel may be used, cooks are likely to eventually run into problems with automatic text conversion and encodings of non-Latin characters. The open source LibreOffice application is recommended for reliably editing CSV files, particularly those that will be accessed by scripts. To set up the configuration of the spreadsheet, cooks will need to have a good code editor. The free Visual Studio Code (VS Code) application is recommended.

The last required ingredients are the VanderBot and convert_config_to_metadata_schema.py scripts. They should be downloaded into the directory where the data will be kept (the working directory).

PREPARATION

Cooks that don't already have a Wikimedia account will need to create one. The same account works across Wikipedia, Wikidata, and Wikimedia Commons, so an account on any of these platforms will be sufficient. Writing to the Wikidata API also requires having a bot username and password, which need to be saved in the home directory of the computer.



	A	B	C	D	E	F	G	H	I
1	qid	label_en	label_ar	description_en	description_ar	author_uuid	author	title_uuid	title
2		The Book of Ingenious Devices (1978)	كتاب الحيل (ترجمة إنجليزية، 1978)	Translation of: Kitāb Al-Ḥiyal	الترجمة الإنجليزية لكتاب الحيل		Q423210		The Book of Ingenious Devices (Kitāb Al-Ḥiyal)

```
# Each listed property will be used to gene
# using the property do not have qualifiers
prop_list:

# Wikidata property ID to be used for state
- pid: P31
  # variable is the root name for the colu
  variable: instance_of
  # value_type must be one of: item, date,
  # If the value_type is monolingualtext,
  # with a value that's a standard two-lett
  value_type: item
  qual: []
  ref: []

- pid: P1184 # In this example, a Handle i
  variable: handle
  value_type: string
  qual: []
  # Although references in Wikidata can hav
```

```
{
  "@type": "TableGroup",
  "@context": "http://www.w3.org/ns/csvw",
  "tables": [
    {
      "url": "banumusa1.csv",
      "tableSchema": {
        "columns": [
          {
            "titles": "qid",
            "name": "qid",
            "datatype": "string",
            "suppressOutput": true
          },
          {
            "titles": "label_en",
            "name": "label_en",
            "datatype": "string",
            "aboutUrl": "http://www.wikidata.org/entity/{qid}",
            "propertyUrl": "rdfs:label",
            "lang": "en"
          },
          {
            "titles": "label_ar",
            "name": "label_ar",
            "datatype": "string",
            "aboutUrl": "http://www.wikidata.org/entity/{qid}",
            "propertyUrl": "rdfs:label",
            "lang": "ar"
          },
          {
            "titles": "description_en",
            "name": "description_en"
```

column headers
upload CSV file (edit with LibreOffice calc)
VanderBot
convert_config_to_metadata_schema.py
Python script
configuration file: config.yaml
(edit with VS Code)
W3C standard mapping
file: csv-metadata.json
Wikidata
API
The Book of Ingenious Devices (1978)
Wikidata item

Figure 1. Relationship among files involved in the workflow for uploading to Wikidata using VanderBot

It is also helpful for the cook to be familiar with the user interface on the Wikidata website. An example of a doctoral thesis item that is similar to the ones created through this recipe can be found at https://www.wikidata.org/wiki/Q111636714 .

NOTE: to avoid creating duplicate items, the cook should ensure that no one from their institution has already uploaded the dissertations and theses being considered for upload. It is probably a good idea to search for several using the Wikidata search box (upper right corner) before undertaking a mass upload.

COOKING METHOD

1. **Extract data about the theses into a spreadsheet.** This step will be easiest if the ETDs are managed using a content management system (e.g., DSpace) that can export the metadata as a spreadsheet. The preferred export format is CSV, although if an Excel export is possible, the .xlsx file can be opened in Excel and saved as a CSV. The exported data should have a row for each thesis or dissertation and contain columns with the following metadata: title, publication date in ISO 8601 format (YYYY-MM-DD; may contain only the year or the year and month), thesis author name (one author only), and a unique URL identifier for the thesis. Typically, a DOI or Handle will be assigned, and these can be expressed as URL identifiers. Otherwise, provide a stable URL.
2. **Modify the configuration file to include the desired properties.** The columns in the CSV file used in the upload are mapped to Wikidata properties using a simple YAML configuration file. Download the example file from https://github.com/HeardLibrary/linked-data/blob/master/etd/config.yaml into the working directory. Open the downloaded file using a code editor. Most of the settings can be left as they are. The property ID in line 41 is set for P1184 (Handle ID). If you are using a DOI, change this to P356 (DOI) and change the value in line 42 from "handle" to "doi". If using some other kind of persistent URL identifier, delete lines 41 through 53.
3. **Generate column headers for the upload CSV file.** After saving the modified configuration file, use the convert_config_to_metadata_schema.py script downloaded as an ingredient to generate column headers of an empty CSV. In the terminal application, navigate to the working directory and issue the following command (replace "python" with "python3" if your system requires it):

```
python convert_config_to_metadata_schema.py
```

The script will generate two files. The file csv-metadata.json is a W3C standard description file used by VanderBot to understand the structure of the upload CSV. The second file is a file named htheses.csv that contains the header row for the upload CSV file that will be created. The "h" at the start of the filename is to prevent overwriting any existing theses.csv file, so remove the "h" from the filename before using the file.

4. **Adding the thesis data to the upload CSV.** Open the upload CSV file (theses.csv) and the CSV containing the extracted thesis metadata using the spreadsheet editor software. The upload CSV contains a lot of columns that are used by VanderBot for record keeping, so many do not need to be filled in by the cook. These are the ones that need to filled:
   - **label_en**. Paste in the titles of the theses copied from the extracted data CSV. (For a language other than English, this column name would have a different language tag, e.g. "label_fr".)
   - **title**. Paste in the titles of the theses. Note: this system is limited to one title language per spreadsheet (the one designated in line 94 of the configuration file).
   - **description_en**. Paste in either "doctoral dissertation" or "master's thesis" as appropriate for the work.
   - **instance_of**. Paste in "Q187685" for doctoral dissertations or "Q1907875" for master's theses.
   - **handle** (or **doi** if you changed this in the configuration). Paste in the raw Handle strings (or raw DOI strings). That is, use "1803/11165" rather than the URL version "https://hdl.handle.net/1803/11165".
   - **full_text_available**. Paste the URL versions of the Handle, DOI, or permanent URL, e.g. "https://hdl.handle.net/1803/11165".
   - **author_string**. Paste the author

names. The convention is to put the given name first, surname last for languages where that is the norm.

- **published_val**. Paste the publication date (must be in ISO 8601 format; can be the year only).
- **language**. Paste in "Q1860" for English. For other languages, search for the Q ID in Wikidata.
- **dissert_submit_to**. Look up the Q ID for the university to which the thesis was submitted. For example, use "Q29052" for Vanderbilt University.

In each reference date column (columns whose name ends in "_ref1_retrieved_val"), paste the full ISO 8601 date for when the metadata was exported. For example: "2022-02-18" for February 18, 2022. There should be seven of these retrieved date columns if no columns were removed from the configuration file.

In each reference URL column (columns whose name ends in "_ref1_referenceUrl"), paste the same URL used in the full_text_available column, e.g. "https://hdl.handle.net/1803/11165".

5. **Sandbox upload.** If the qid column in a row of the spreadsheet is empty, VanderBot creates a new item. If that column in a row contains an item Q ID, the statements are added to that existing item. To ensure that everything is working as planned, it's best to try uploading statements to one of the existing Wikidata sandbox items. Save a copy of the upload spreadsheet under a different name. In the original copy, delete all of the lines except for the first one, then enter "Q4115189" in the qid column. Be sure to save and close the CSV file after you edit it. View the sandbox item at https://www.wikidata.org/wiki/Q4115189. In the terminal application, issue the command:

```
python vanderbot.py
```

If the script finishes successfully, refresh the sandbox item web page to verify that the changes made are what was expected. (To get rid of the test changes, click on the "View history" link at the top of the sandbox item page, then click "undo" by the test revision.) After reopening the upload spreadsheet, the identifiers VanderBot added after the upload will be visible.

If the script does not finish successfully, look at any error messages to see what needs to be corrected.

6. **Test upload.** Now try creating a new item for the first thesis in the spreadsheet. Repeat what was done in the previous step (make a copy of the full data CSV, delete all but the first line, and name it theses.csv), but this time leave the qid column empty. Save and close the CSV. Run the VanderBot script again. If the upload was successful, look at the qid column of your spreadsheet to see what Q ID was assigned to the new item. Use that Q ID to look up the new item and verify that everything is as you expected.
7. **Upload the actual data.** Open the copy of the upload data that was saved under another name and open the theses.csv file just used for the test upload. Copy and paste the first line of data from theses.csv to the CSV copy so that the assigned identifiers will be in the full copy. Save the CSV copy, delete the single-line theses.csv file and rename the copy to theses.csv. The next time VanderBot is run, it will upload the data for all of the theses (skipping the first line since its identifiers are already filled in). Note: "newbie" users are users who have accounts less than four days old and who have done fewer than fifty edits. If a cook is a newbie user, they will be subject to a slower edit rate (eight edits per minute). In order to avoid getting blocked, newbies should use the apisleep option to add a longer value of eight seconds between edits when they run VanderBot:

```
python vanderbot.py
--apisleep 8
```

Cooks that are not new users can run the script without the apisleep option to write at the normal rate of fifty edits per minute.

CHEF'S NOTES

Test uploads help to avoid a mess in the kitchen. A small number of errors can be corrected using the Wikidata user interface. However, if a cook discovers that they've made many mistakes in their uploads, they can delete statements or references by providing the identifier information stored by VanderBot in the upload CSV to a related script called VanderDeleteBot. (For information about the VanderDeleteBot script, visit https://github.com/HeardLibrary/linked-data/blob/master/vanderbot/vanderdeletebot.md.) Corrected statements can then be added using VanderBot.

If cooks discover that they have created duplicate thesis items by mistake, they can't delete them, but they can merge the duplicate items by using the "Merge with…" tool under the "More" menu at the top of the Wikidata page of the duplicate item.

In this upload process, the name string for the author was used. If the author has a Wikidata item, it is best to link to it instead. Author Disambiguator (https://author-disambiguator.toolforge.org/) is a good tool for making those links.

Notes on Ingredients and Equipment

- If you are unfamiliar with your computer's file system or with giving commands in your computer's console, see Lessons 1, 2, and 6 at http://vanderbi.lt/computer.
- If you don't know whether you already have Python 3 or if you need to install it, there are detailed instructions in the "Before starting" section at http://vanderbi.lt/ld4vb. That page also has instructions for installing the "requests" library.
- To download LibreOffice, visit https://www.libreoffice.org/. The Visual Studio Code (VS Code) application is available for free at https://code.visualstudio.com/download.
- The VanderBot script can be downloaded from https://github.com/HeardLibrary/linked-data/blob/master/vanderbot/vanderbot.py. The convert_config_to_metadata_schema.py script can be downloaded from https://github.com/HeardLibrary/linked-data/blob/master/vanderbot/convert_config_to_metadata_schema.py. If you haven't downloaded a file from GitHub before, you can see a video explaining how in the "Downloading the VanderBot script" section at http://vanderbi.lt/ld4vb.

Notes on Preparation

- To create a Wikimedia account, visit https://en.wikipedia.org/wiki/Special:CreateAccount. There are instructions (with videos) for creating a bot username and password in the Preparation section at http://vanderbi.lt/ld4vb.
- For some beginner video lessons on Wikidata (available in English, Spanish, and Chinese), visit https://www.learn-wikidata.net/.