Developing a linked data workflow using Wikidata

Will Kent
Wiki Education

SWIB - September 2023
Agenda

- Program background
- Course details
- Sharing outcomes
Background

- Wiki Education courses started in 2019*
  - Intro course; no prior experience necessary
  - Six sessions, each an hour long
  - Seven slide-based training sessions and some exercises as homework^
  - ~3 hour commitment/session
  - All meetings virtual
  - Slack channel to supplement
  - Particular emphasis on community integration, querying, manual editing, tools, and batch edits

*https://meta.wikimedia.org/wiki/Wiki_Education_Foundation/Wikidata_Program_Evaluation
^https://dashboard.wikiedu.org/training/wikidata-professional
Outcomes

• Based on survey results we know the course is a success!
Participants are satisfied with the curriculum
Understanding of Wikidata fundamentals
There was one common piece of feedback...
Let’s keep going

- Is there an advanced course?
  Can we have space to practice with these tools?
- How do we convince stakeholders to run Wikidata projects?
- Complex questions about more queries, tools, and policies
- How can we develop a linked data workflow with Wikidata?
Some challenges

- Different goals
  - Metadata roundtripping
  - Metadata enrichment
  - Modeling/schemas
  - Translation
  - Data visualization
  - Querying
  - Property creation
  - Wikipedia work

Intuition can only take you so far
Sample questions

- What are best practices for translating labels, descriptions, and aliases?
- How do I highlight the work our researchers do in Scholia?
- Everything about batch edits - how do I make them? How can I be precise? How can I build them into a workflow for my staff?
- How do I create a property? How do hierarchies work on Wikidata? Do they work?
- How do I develop visualizations with Wikidata?
- How do you create a workflow?
A project-based course

- In November 2022 we piloted a new Wikidata curriculum, requiring a project and accompanying dataset to participate.

To date we have run two of these courses with another two on the horizon.

Twenty-five participants so far.

Sharing the outcomes of these courses is important as Wikidata becomes more common in libraries, museums, and across the internet.
Course details

Same setup as the beginner course - six sessions, hour-long, all virtual
No training modules, more live demos
Aim for 10-20 participants per course
Emphasis on the cohort learning together
1:1 meetings encouraged
Specific case studies and resources
Meetings loosely themed, but greater emphasis on participant-driven sessions
Curriculum

Some experience with Wikidata is assumed
Meetings themed around data modeling,
querying/data visualization, general Wikidata tools,
batch editing, documentation and community integration
Specific tools include Cradle, Tabernacle,
SQID/Reasonator, Mix’n’match, SourceMD, Merge tools, Graph builder, Wikidata/Wikipedia spreadsheet extension, PAWS, PetScan, Listeria, etc.
Batch tools include Quickstatements, Open Refine, Author Disambiguator
Wikibase - how and when you may want to use it
Ample time left for discussion, troubleshooting, and individual questions/participant-led tutorials

https://dashboard.wikiedu.org/courses/Wiki_Education/Wikidata_Project_Course_(Spring_2023)/timeline#week-1
## Wikidata:Linked open data workflow

There are many considerations when contributing data, media or other assets to Wikimedia projects. This chart attempts to list some of them and track tools and scripts in the linked data workflow. This is based on the [data and media partnerships chart on Outreach Wiki](https://www.wikidata.org/wiki/Wikidata:Linked_open_data_workflow).

<table>
<thead>
<tr>
<th>PREPARE and tidy source data and images</th>
<th>RECONCILE with Wikimedia modeling and coverage</th>
<th>INGEST data and media</th>
<th>ANALYZE, correct, and enrich</th>
<th>RE-USE content intra-wiki and externally</th>
<th>REPORT and measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Wikidata Data Donation</td>
<td>• OpenRefine</td>
<td>• QuickStatements</td>
<td>• Wikidata Query</td>
<td>• Wikidata infobox / Infobox templates</td>
<td>• GLAMorgan</td>
</tr>
<tr>
<td>• Creative Commons</td>
<td>• Wikidata Mix’n’match</td>
<td>• OpenRefine</td>
<td>• TABELnacle</td>
<td>• Creator templates</td>
<td>• BaGLAMa</td>
</tr>
<tr>
<td>• Flickr Image hosting</td>
<td>• Wikidata Query</td>
<td>• Wikidata-CLI</td>
<td>• Listeria</td>
<td>• Template:Artifact, Template:Book, etc</td>
<td>• GLAM Wiki Dashboard</td>
</tr>
<tr>
<td>• Google Spreadsheets, Wikipedia and Wikidata tools for Google Spreadsheets</td>
<td>• WikiCommons Query Service</td>
<td>• Pattypan</td>
<td>• Maintenance query scripts</td>
<td>• Template:Label, Template:WrapW, etc</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Wikipedia Graph Builder</td>
<td>• url2commons</td>
<td>• Wikidata Distributed Game</td>
<td>• Listeria / Template:Wikidata_list</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• SQID</td>
<td>• PyWikiBot / upload.py</td>
<td>• Cat-a-lot</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Petscan</td>
<td>• Mediawiki API</td>
<td>• VisualFileChange.js</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• PAWS</td>
<td>• Commons Upload Wizard</td>
<td>• SPARQL RC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Mediawiki API</td>
<td>(from Flickr)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Notes

- Try finding a similar project or collection set on Wikidata or Commons to see how it has been done in the past.
- For Wikidata, usually a "crosswalk database" is needed to map terms from the uploading data set (a CSV file or records from an API). In addition, try uploading small test batches before doing large data sets. When ingesting collection data, be aware that you may be importing duplicates or conflicts with other editors.
- Depending on the success of the import and uploading process, you may need to deal with duplicates or conflicts with other editors.
- Scripts and templates can generate on-wiki content such as tables and infoboxes from Wikidata.
- Show the impact of contributions by tracking metrics on files used or impressions over time. For partnerships, this can help.

Impact

www.wikidata.org

<table>
<thead>
<tr>
<th>General</th>
<th>Items</th>
<th>Claims</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.84K</td>
<td>1.69K</td>
<td>17.9K</td>
<td>1.74K</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>984</td>
<td>513</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>314</td>
<td>23</td>
</tr>
<tr>
<td>Descriptions</td>
<td>Aliases</td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>1.74K</td>
<td>1.94K</td>
<td>1.58K</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>268</td>
<td>Qualifiers added</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>References added</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Redirects created</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reverts performed</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Restorations performed</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other updates</td>
<td></td>
</tr>
</tbody>
</table>

- Items Created: 1.69K
- Items Edited: 2.27K
- Total Edits: 5.84K
- Editors: 11
- References Added: 18.7K
- Item Views: 20.6K
- Commons Uploads: 1

https://dashboard.wikiedu.org/courses/Wiki_Education/Wikidata_Project_Course_(Spring_2023)/home
Why is this important?

Structured courses with live instructors remain rare
Without continued work, projects stop
Creating a space for sharing allows for more support,
   exchange of ideas, and recommendations for best practices
Since there are so many moving parts with a workflow, if
   one person leaves, projects can slow or stop outright
The impact project-based courses have on Wikidata can be BIG!
Not to mention impact on local collections
New tools and systems will appear - having courses,
   spaces, and people familiar with this will improve a lot
Building a workflow is complicated - there is no manual for any of this
Course feedback

People want more! Longer courses, sustained contact, more time to spend on projects with mentors, coaches, other practitioners.

There is no one-size-fits-all course and that’s okay.

Content that was review for some was helpful and not annoying or a waste of time.

The size was particularly helpful - there are many Wikidata offerings - videos, slide tutorials, etc, but the combination of size, live sessions, and timing worked well for many.
Pedagogical style

Participant-driven (when possible)
Inquiry and action-based
Open pacing, leave room for conversation to develop
Curriculum pushes editing and becoming integrated into the community; encouraging these is central to the fabric of the course
Assume Wikidata is right and wrong; frame as opportunities for improvement
What’s next?

Build the curriculum out to better engage with more specific needs
For popular topics, create training modules
Document everything - failures, successes, and workflows
Wikidata in the classroom
Share findings at conferences to encourage more Wikidata work and convince a broader audience to let their staff run Wikidata projects
Thank you!

Learn more / stay in touch

will@wikiedu.org
learn.wikiedu.org