Data Quality Days

[[User:Mike Peel]]

14 September 2021

Cross-checking on-wiki: visibility, duplication, migration

WIKIDATA

If you want good quality data,
you have to use it, and
you have to show it to others.

(Otherwise you might never see what's wrong!)

Visibility

- Use information on:
  - Wikipedias (infoboxes, authority control, etc.)
  - Commons (infobox, SDC on files)
  - Wikisource
  - ...
- With clearly marked edit links!
- Can use all info, or only selected parts, e.g., referenced information

Duplication

- Projects might not be willing to give up having a local copy
- Can do automatic comparison between local values and Wikidata values
- Tracking categories can flag differences
- Automatic bot imports can help import new data
- Manual checks of discrepancies (painful!)

Migration

- Just use Wikidata information directly!
- Much simpler than having multiple copies!
- Local maintenance easier -> more work done on Wikidata (and less duplication)
- Fix a problem in one place, it's fixed everywhere
- Vandalise in one place, it appears everywhere...

Example 1: Identifiers

Subcategories

This category has the following 195 subcategories, out of 195 total.

0–9

- 90minut template with ID different from Wikidata (empty)

A

- Admiralty number different from Wikidata (142 P)
- AHL profile different from Wikidata (empty)
- AlloCiné name template with ID different from Wikidata (empty)
- AOC athlete ID different from Wikidata (13 P)
- ARKive ID different from Wikidata (13 P)
- ARRS name ID different from Wikidata (empty)
- Articles using sports-reference citation with different Wikidata (514 P)
- ATP template with ID different from Wikidata (empty)
- Australian Statistical Geography Standard 2011 ID different from Wikidata (75 P)
- Australian Statistical Geography Standard 2016 ID different from Wikidata (484 P)
- Australian Wetlands Code different from Wikidata (empty)
- Avibase ID different from Wikidata (12 P)
- FIG template with ID different from Wikidata (empty)
- FIG template with licence number different from Wikidata (18 P)
- FIL template with ID different from Wikidata (1 P)
- FINA template with ID different from Wikidata (empty)
- Find a Grave template with ID different from Wikidata (4 P)
- Finnish MP ID different from Wikidata (empty)
- FIS alpine skier template with ID different from Wikidata (2 P)
- FIS cross-country skier template with ID different from Wikidata (empty)
- FIS freestyle skier template with ID different from Wikidata (empty)
- FIS Nordic combined skier template with ID different from Wikidata (empty)
- FIS ski jumper template with ID different from Wikidata (empty)
- FIS snowboarder template with ID different from Wikidata (empty)
- FISA ID different from Wikidata (empty)
- Flora of Australia ID different from Wikidata (8 P)
- NFL.com player template with ID different from Wikidata (empty)
- NFT template with ID different from Wikidata (189 P)
- No local image but image on Wikidata (5,379 P)
- NZOC profile ID different from Wikidata (4 P)

O

- Official website different in Wikidata and Wikipedia (58,765 P)
- Olympedia template with ID different from Wikidata (empty)
- Olympic Channel template with ID different from Wikidata (12 P)
- Open Library ID different from Wikidata (428 P)

P

- Plantarium ID different from Wikidata (empty)
- Power of 10 ID different from Wikidata (empty)
- ProCyclingStats race template with ID different from Wikidata (8 P)
- ProCyclingStats team template with ID different from Wikidata (30 P)
- ProCyclingStats template with ID different from Wikidata (686 P)

https://en.wikipedia.org/wiki/Category:Wikipedia_categories_tracking_Wikidata_differences

Challenges & Solutions

- ID can be:
  - Missing on Wikidata
  - Wrong on Wikidata
  - Wrong locally
  - Misformated
  - Duplicated on source website

- Can fix by:
  - Importing to Wikidata (bot?)
  - Fixing on Wikidata/locally (manual?)
  - Migrating to Wikidata
  - Reporting to source website

Example 2: Commons categories

On enwiki, populated using Template:Commons_Category

Can install this on your local wiki if there's consensus.

https://en.wikipedia.org/wiki/Category:Commons_category_Wikidata_tracking_categories

Challenges

- Commons category link can be:
  - Wrong
  - Moved
  - Missing
- Wikidata can be:
  - Missing link between topic and category items (P910/P301)
  - Missing link between list and category item (P1753/P1754)
  - Linked to a gallery item (category item needs to be created)
  - Misplaced link (in another item)
  - Missing the sitelink to Wikipedia

Solutions

- Pi bot automatically updates:
  - Missing categories (remove link)
  - Redirects (update link)
  - New links (if matching pagename)
- Manual/semi-auto edits:
  - Everything else!
  - New links with different names
  - Missing Wikipedia sitelinks
  - Better category (not linking to a redirect)
  - Multiple commons links (normally only need one!)
  - New Commons category needed

Editing tasks

Clean up existing mis-matches

- Commons links
- Local authority control
- Others?

Add more checks of local info to Wikidata

- Match link templates with Wikidata properties
- Update local templates (e.g., commonscat) to compare with WIkidata
- Propose migrating templates to Wikidata

Useful links:

- https://en.wikipedia.org/wiki/Template:Commons_category
- https://commons.wikimedia.org/wiki/Module:WikidataIB
- https://en.wikipedia.org/wiki/Category:Wikidata_tracking_categories