Data Quality in Wikidata

Wikidata provides data to Wikimedia projects and beyond. The data can be queried, visualized, presented in an infobox on Wikipedia, answer a question in a digital personal assistant and much much more. Data quality is crucial for Wikidata because only if our data is in good shape can we really serve our purpose of giving more people more access to more knowledge.

There are many tools to help you monitor and improve the quality of the data in Wikidata.

**Watchlist and Recent Changes integration**: On Wikipedia and the other sister projects you can show changes from Wikidata in your watchlist and Recent Changes. This way whenever someone makes a change on Wikidata that affects an article you are watching you see that in our watchlist.

**Listeria**: This bot lets you create lists based on a query to Wikidata and puts the result in a wiki page. You can put this page on your watchlist and be notified whenever anything changes in the query result based on an edit someone made on Wikidata.

**Constraint checks**: Constraint checks allow editors to define rules on the data (like “someone’s date of birth and date of death should not be apart more than 120 years” — unless they are a time traveler, of course). When you then come across a statement that violates one of those constraints you will see a little note alerting you to the issue so you can easily spot and correct it.

**sparql-rc**: This tool allows you to monitoring changes in areas you care about. For example this can be useful for a wiki project that wants to keep items about their topic of interest in good shape.

**Wikidata vandalism dashboard**: This dashboard lets you get a list of changes to descriptions, labels, aliases and sitelinks for a specific language. It lets you easily find and remove vandalism and errors specific to a language you speak.

**Page information**: In the page information of every article on Wikipedia and the other sister projects you can find a list of all items that contribute to the article, making it easier to see where data is coming from.

**Edit groups**: This tool lets you more easily see and understand a group of edits like a bot-run, and if necessary revert all the changes it made.

We are far from done when it comes to creating tools to help everyone keep data quality in Wikidata high. Here are some highlights we will be working on in the future:

- **ShEx**: Shape Expressions will make it easier to define what certain classes of items should look like and find the ones that don’t fit the desired shape. For example we could define a shape that says that all movies should have a link to IMDB and a release date.

- **Signed statements**: We will enable organisations who contribute larger amounts of data to Wikidata to cryptographically sign the statements they contribute. This way it will be easy to spot when one of these statement was changed and we can evaluate if this was a beneficial change or not.

- **Checks against other databases**: Comparing Wikidata’s data with other databases will allow us to find mismatches and flag them. This way we could for example more easily find the different date of birth of a musician in Wikidata and MusicBrainz and then investigate further which one is backed up by reliable sources.

Contact: Lydia Pintscher (lydia.pintscher@wikimedia.de) / @wikidata cc-by-sa