Summary

**Problem:** While we have quite some knowledge about the Wikidata community, particularly about the very active members, we lacked a general and overview of the community and its structure.

**Goal:** Create a short survey to gather data on community members concerning their demographics and activities. Use the insights as baseline for future metrics and to guide activities.
Summary

- Most participants are
  - binary male
  - young
  - from the global north
  - employed
  - have between 1000 and 100,000 edits
- A lot of participants started editing Wikidata 2012/2013

- The tenure or edit count of participants seems to have no meaningful relation with the diversity of gender or place of living.
- Some activities on Wikidata go together: There is a cluster of collaborating, events and outreach activities and a weaker cluster of adding and cleaning data.

The results are based on 872 answers to the survey. The survey was advertised via Wikidata project chat and via Central Notice banners on Wikidata.
Survey Questions
Survey

Implemented on Lamapoll

We are running this survey to understand who edits Wikidata and to discover if there are any groups that currently might have difficulties in participating. Wikidata attempts to cover human knowledge as broadly as possible. What people know and care for also depends, in practice, on their cultural background, gender, age and other factors.

- Takes about 1-2min
- Anonymous
- No mandatory questions
- Survey platform (Lamapoll, hosted in Germany) and data handling are GDPR compliant.
  - [Data Privacy Statement Wikimedia Deutschland e.V.](#)
  - Data will be published in aggregated form

More information on the background of the questions

What is your age group?

- [ ] less than 18 years
- [ ] 18-24 years
- [ ] 25-34 years
- [ ] 35-44 years
- [ ] 45-54 years
- [ ] 55-64 years
- [ ] 65-74 years
- [ ] 75-84 years
- [ ] 85 years and up
- [ ] Prefer not to disclose
Survey

The Wikidata Community in 2021

What is your Gender?
- Woman
- Man
- Non-Binary
- Prefer not to disclose
- Prefer to self-describe

In which country do you live?

How many years did you spend in education?
- Years of primary and secondary school
- Years of vocational education
- Years of university

I am currently...
- Employed
- Self-Employed/Freelancer
- Seeking employment
- In education
- Retired
- Primary caregiver for another person, e.g. a child or elderly relative
- other:

I work in...

Job title or field of work you identify the most with

For background on this way of asking for gender information:
Survey

Since which year have you been active on Wikidata?

How many edits did you make on Wikidata so far?

In which of the following data related activities are you typically involved?

- Adding data to Wikidata
- Editing, cleaning or maintaining data on Wikidata
- Finding, sharing or reusing data from Wikidata for Wikimedia Projects
- Finding, sharing or reusing data from Wikidata for external, non-Wikimedia projects
- Collaborating with other members on Wikidata and/or Wikimedia projects
- Attending and participating in Wikidata and/or Wikimedia events and workshops
- Reaching out to increase community awareness about Wikidata
- Something else

If you had to give your role a name (based on the activities you are typically involved in), what would it be?

Name of my role: Your answer
Is there anything else you would like to let us know about your participation in the Wikidata community?

Comments / Criticis / Suggestions ...
Setup
Having the banner shown via central notice was no problem per se, but it needed some communication to get the banner activated after all questions were answered. We are unsure if one needs to contact specific people to get it going, if we oversaw a needed process step or if our request has gotten lost.
Coordination with Community Communication Team

The survey and its goals were communicated by the Community Communication Team. We also prepared a page documenting the goals as well as the survey questions and their purpose.
Timeframe

- **Online: 10th of February to the 21nd of March.**
  - For the review process on central notice, people wanted to see the survey the banner would lead to.
  - 31 people participated before we showed the banner
- **Banner campaign: 10th of March to the 19th of March 2021.**
  - Most participants (943) participated while the Banner was shown. Seven participants participated after the period of showing the banner ended; this is most likely due to us sharing the links to the survey in our communication with the community.
Data cleaning, structuring and visualization in google sheets and R

The country names were normalized in google sheets and supplemented with data on their continent and whether they are part of the “global south” via ITU’s classification. Data was cleaned and re-coded in R. We excluded responses that did fill free form fields with obviously provocative answers (3 participants)
Possible Changes in the future
Countries and Gender as open fields

**Gender**: Checkmarks Man/Woman/non binary/self-defined.

**Countries**: Open text input. We normalized the country names to their English-speaking equivalent.
Back to a more constrained input?

- Have a (binary) male / (binary) female / non-binary / prefer to self define/not disclose
  - similar to as I summarized the data in the end
  - but it loses some expressiveness (e.g. non-binary-female)
- Have a list of countries if possible
  - met some technological difficulties;
  - some people also want to put in countries like “basque country”
Activities

- Ranking by importance could be helpful
- Include a “connecting Wikipedia Articles” as it is quite possible that many participants do not see themselves as Wikidata Editors by more or less willingly using Wikidata for the sitelinks.
Work and Carework

- We included the question for being a primary caregiver in the work section. This might run counter to the expectations and might be better in a separate section...
- ...on the other hand: “Care work is work”
Education

Shall we go to the standard “highest qualification”?
- Easier to analyze and thus easier to show over/underrepresentation
- ...but also reinforces hierarchy.
Some participants remarked that they did not know what “Vocational Education” is.
We had some minor problems here:

- A copy paste error in Spanish, which included one option twice and not another.
- One complaint about the quality of the Arab translation and one on the German translation.
- One person wanted to take the survey in Russian.
Multilingual Approach

How many answers in which translations?

Catch 22, here?
Results
### Gender

<table>
<thead>
<tr>
<th>Wikidata</th>
<th>All Wikimedia Projects in 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Count</strong></td>
<td><strong>Percent</strong></td>
</tr>
<tr>
<td><strong>Percent</strong></td>
<td><strong>Percent</strong></td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td>138</td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td>656</td>
</tr>
<tr>
<td><strong>nonbinary</strong></td>
<td>25</td>
</tr>
<tr>
<td><strong>No Data /NA</strong></td>
<td>53</td>
</tr>
</tbody>
</table>

**Chart Description:**
- The chart shows the gender distribution across Wikidata and all Wikimedia Projects in 2020.
- **Female:** 138 (16% of the total), 12% of Wikidata.
- **Male:** 656 (75% of the total), 87% of Wikidata.
- **Nonbinary:** 25 (2.9% of the total), less than 1.8% of Wikidata.
- **No Data /NA:** 53 (6.08% of the total), unknown proportion across projects.
When interpreting the diagram, keep in mind that the first age bracket spans only 6 years and not 10 as the rest, as minors were not all allowed to participate.

The most frequent age bracket was 25-34 years. However, the previous bracket also has a high count, despite spanning only 6 years.
Countries where participants live

- Other
- Netherlands
- India
- Canada
- United Kingdom
- Spain
- Italy
- Czechia
- France
- United States
- Germany
Countries where participants live

We have a strong representation of the Global North in our participants: 76% of the people who participated are from the Global North, only 16% are from the Global South*. (We mapped countries according to the mappings of the ITU)

The term “Global South” is not merely a geographical term. It is referring to countries and regions of the world less powerful and less wealthy than the countries and regions of the “Global North”. The “Global North” still includes e.g. Australia.
Possible interpretations:
In the beginning years of the project, 2012 and 2013, many people started to edit Wikidata. Many of them were probably Wikipedia editors who also started editing Wikidata (e.g. as part of their work in Wikipedia).

Also, we seem to have more respondents starting to edit Wikidata in the recent years. However, this is likely to be caused by regular dropouts over time (proportional to the entry date, see Lindy Effect), a hypothesis also supported by data independent of the survey (see issue T282563).

For the sake of clarity, the diagram does not include 2012 and 2021 as both years could only span a few months: Wikidata started late 2012 and the survey was done early 2021.
In a survey on all Wikimedia Projects in 2020, there was a higher gender diversity among newer contributors. In our data there was no clear connection.

For the sake of clarity, the diagram does not include 2012 and 2021 as both years could only span a few months: Wikidata started late 2012 and the survey was done early 2021.
Year of first editing Wikidata

It is unclear if we get more diverse in regards to participants from the global south – there is a strong but varying tendency towards contributors from the global north.
Work participants do aside of Wikidata

Any combination of types of work was permitted, we only can show the most frequent combinations here (the long tail of other combination is in “Other”)

![Bar chart showing work participation categories with Employed being the most frequent, followed by Retired, in Education, Freelancing, and so on, with a notable “Other” category at the end.]
Work participants do aside of Wikidata

Very few people in our sample said they are the primary caretaker for another person: 2.41%.

Caregiving is a gendered activity in our sample: 3% (binary gendered) woman and 1.4% of (binary gendered) man said they are primary caretakers of another person.

The number of caretakers are very small in total – even in the our sample with over 800 responses the rates mean that just 6 woman and 10 man are being primary caretakers. So any correlation is brittle.
To read the diagram, be aware that the brackets are logarithmically scaled: The next bracket is always 10 times larger than the one before.
We have two peaks at the 0-10 and at the 1001-100 000 groups.
If there is a correlation between gender and number of edits, it is weak. Limiting factors most likely happen before.
If there is a correlation between gender and number of edits, it is weak. Limiting factors most likely happen before.
Amount of edits

If there is a correlation between place of living and number of edits, it is weak. Limiting factors most likely happen before.
Amount of edits

(same data as on the previous slide, but using a stacked form)

If there is a correlation between place of living and number of edits, it is weak. Limiting factors most likely happen before.
Activities on Wikidata

Like the work activities, these could have any combination of factors. Here are the most frequent combinations, with less frequent ones lumped together in “other combination”

Activities that were typically combined (see appendix for details):
  - events/collab/outreach
  - cleaning/adding
next steps
Next Steps

- Publish and share with community and public
- Use as baseline for coming activities
- Possibly to kickstart smaller initiative for improvement
Appendix
Activities on Wikidata – Are there clusters?

(Yes!)
There seem to be two plausible clusters of correlating activities:

1) Cleaning and Adding data
2) Collaborating, events and outreach

A principal component analysis corroborates the collaborating, events and outreach cluster; Depending on the number of assumed factors, adding and cleaning data is attributed to one component (5 Components) or grouped with sharing Data in Wikimedia Projects/“aIsShareWM” (3 Components)
Activities on Wikidata – Principal Component Analysis

**KMO** = 0.76 with individual values being all >0.62
(>0.5 is recommended). The lowest values were 0.62 for “other”, 0.64 for “adding data”.
**Scree plot** was ambiguous without a clear inflection point.
Activities on Wikidata
– Principal Component Analysis
(five factors)

Going with 5 components based on the scree plot and the correlation plot.

```r
pca_activities3pm <- principal(justActivities_numeric, nfactors = 5, rotate = "promax")
```

...results in this Factor loading (<0.3 is displayed as empty cell)
Activities on Wikidata – Principal Component Analysis (three factors)

Going with 3 components based on the scree plot and the correlation plot:
```r
cpa_activities3pm <- principal(justActivities_numeric, nfactor = 3, rotate = "promax")
```
...results in this Factor loading (<0.3 is displayed as empty cell):

<table>
<thead>
<tr>
<th></th>
<th>RC1 &lt;S3: AsIs&gt;</th>
<th>RC2 &lt;S3: AsIs&gt;</th>
<th>RC3 &lt;S3: AsIs&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>alsOutreach</td>
<td>0.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alsEvents</td>
<td>0.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alsCollab</td>
<td>0.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alsShareExt</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alsAddingData</td>
<td></td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>alsCleaning</td>
<td></td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>alsShareWM</td>
<td>0.52</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>alsOther</td>
<td></td>
<td></td>
<td>0.94</td>
</tr>
</tbody>
</table>
Area of Work and Diversity

exploratory results
The field of work people identified most with has a relation to diversity of gender or place of living. This is based, however, on the coding of the data of a single person in the categories of “Tech”, “Research”, “GLAM” and other/NA. So the data should be treated as exploratory.
Area of Work and Diversity – Gender

Binary female participants more often work in a field of work coded as GLAM.
Area of Work and Diversity – Place of living

Participants form the Global South work a bit less frequently in a GLAM-coded field, but do, almost to equal proportions, work more in a Research-coded field.
Authors and License

Research and Report was conducted and documented by Jan Dittrich, Wikimedia Deutschland e.V. with support and feedback by the UX-, community communications- and product management-team at Wikimedia.

This research would not have been possible without the participation of several hundred Wikidata Community members, who filled out our survey.

June 2021, Wikimedia Deutschland e.V.
licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.