Multilingual Structured (Climate) Research Data in Wikidata

The Community Perspective

Cristina Sarasua & Daniel Mietchen

DOI: <u>10.5281/zenodo.3994272</u>

View on YouTube

Hurricane Katrina with 10 years hindsight



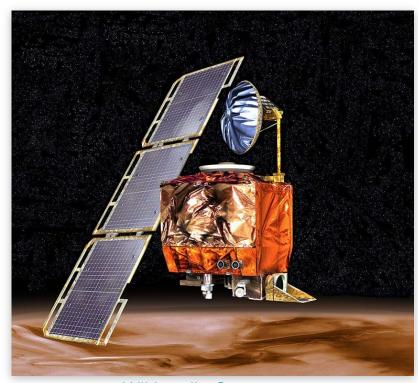
"Open data matters most when the stakes are high

[....]

One of the great lessons we learned through the experience was the power of data to illuminate our path to recovery."

—Denice Ross (2015), Presidential Innovation Fellow "<u>Ten Years After Katrina:</u> <u>New Orleans' Recovery, and What Data</u> Had to Do with it"

NASA Mars Climate Orbiter



source: Wikimedia Commons

"The problem here was not the error; it was the failure of NASA's systems engineering, and the checks and balances in our processes, to detect the error. That's why we lost the spacecraft."

—Edward Weiler, NASA associate administrator for space science, IEEE Spectrum: Why the Mars Probe went off course" source: Wikipedia

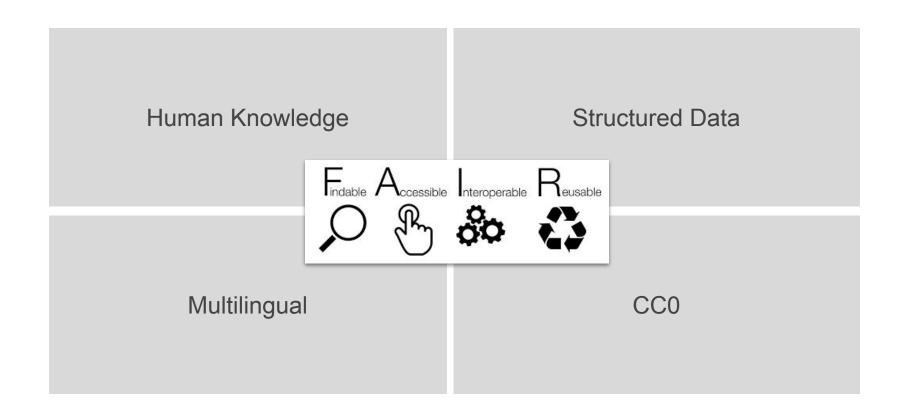
NASA Mars Climate Orbiter

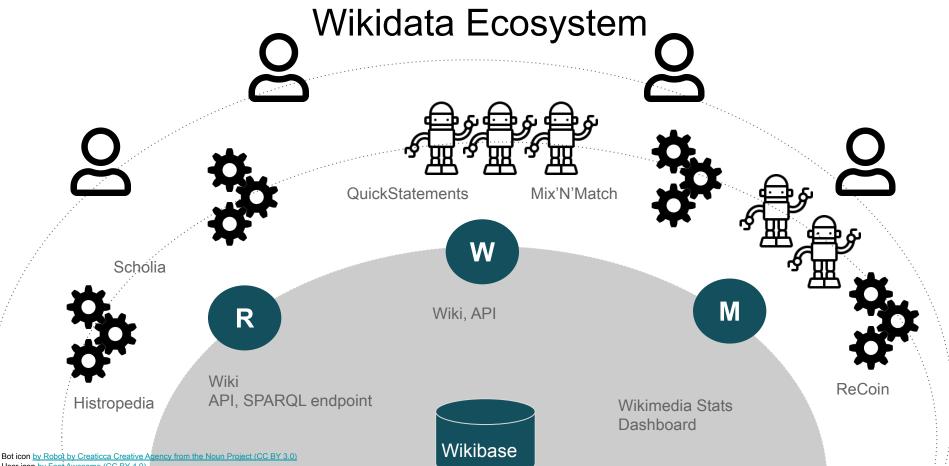
Toward More Effective Data Quality Management

- ★ Many different automated data quality checks
- ★ Many diverse human developers and reviewers
- ★ Contextualizing the data

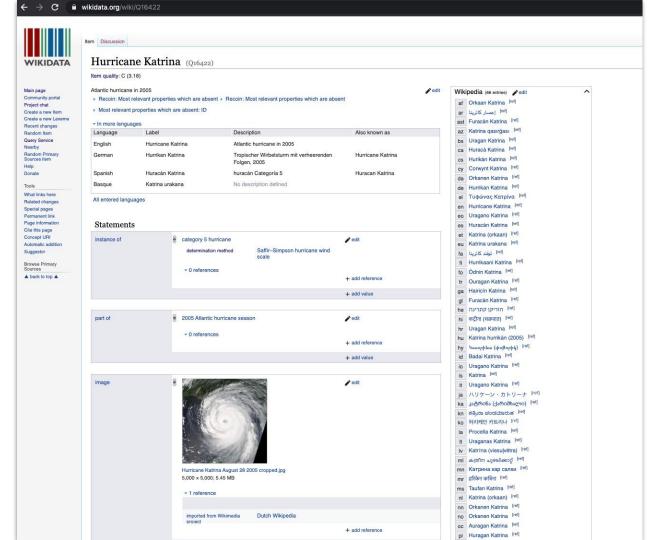


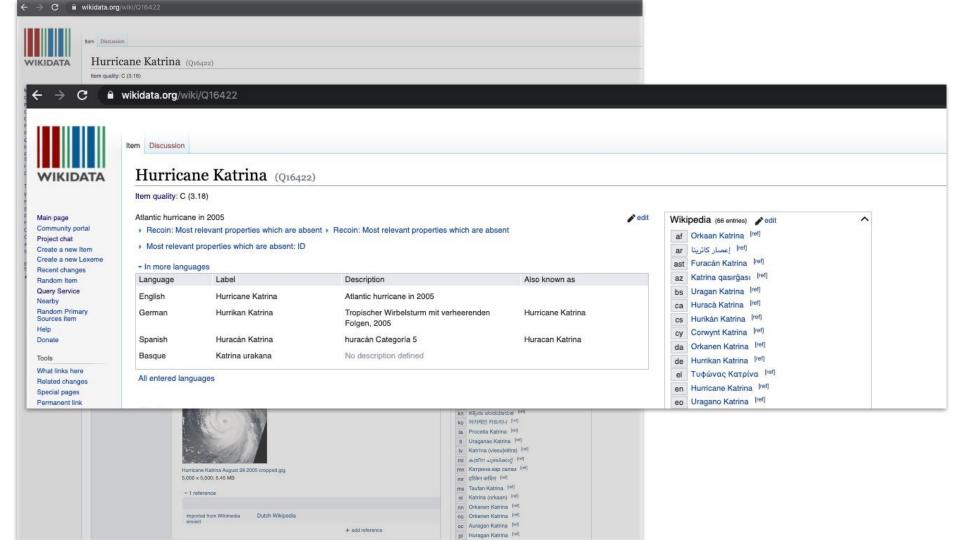
Data

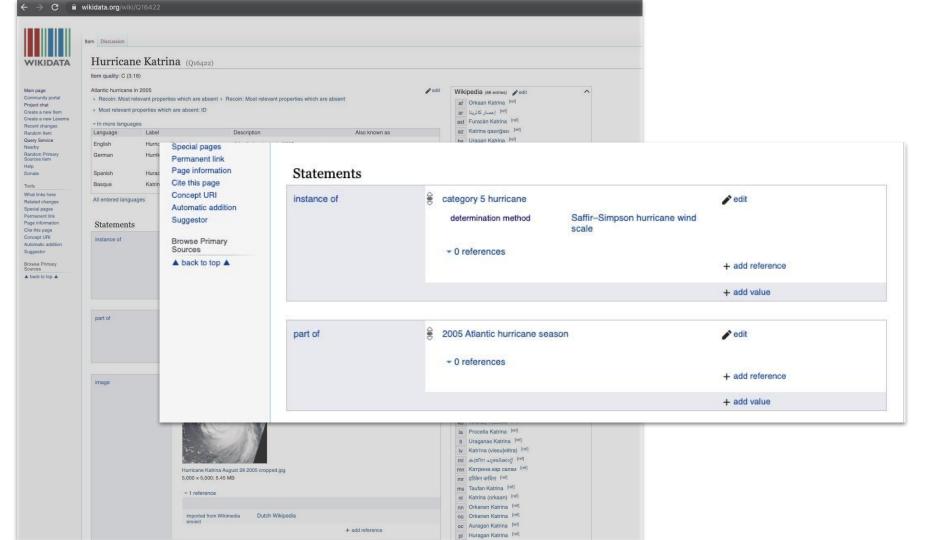


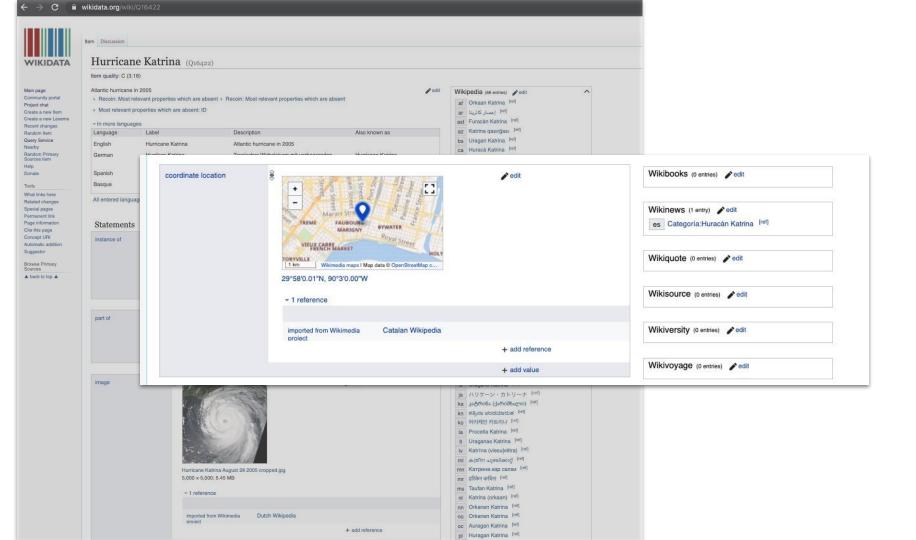


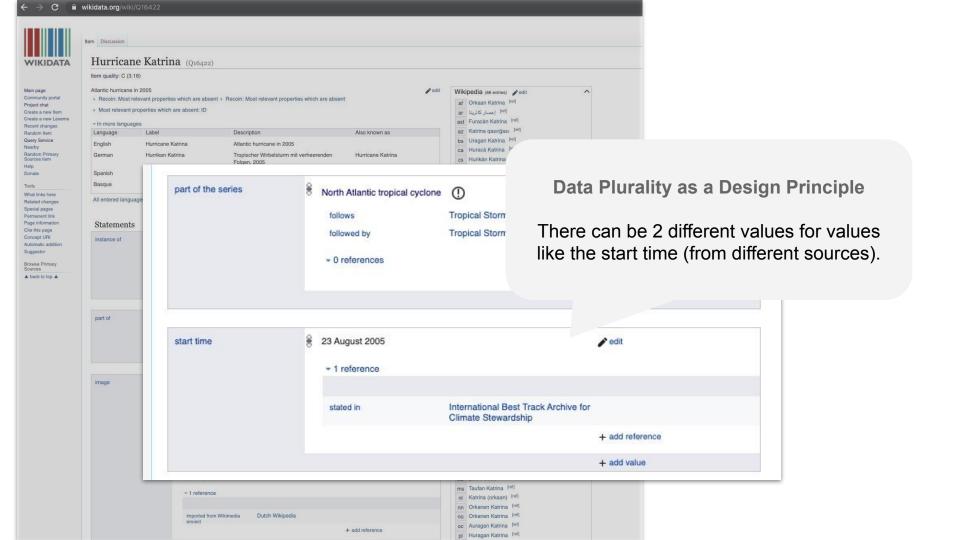
Bot icon by Robot by Creaticca Creative Agency from the Noun Pr User icon by Font Awesome (CC BY 4.0) Tools icon by Verdy p(CC0)

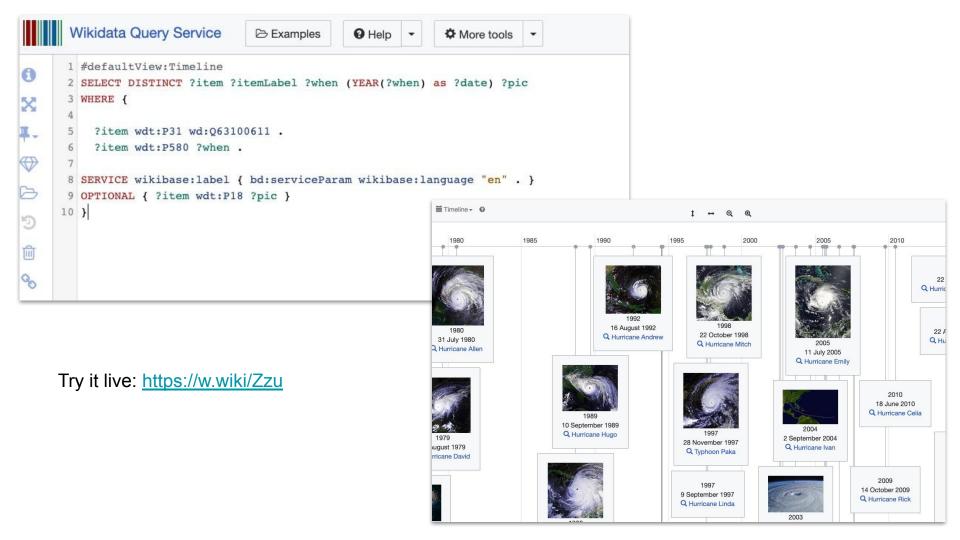




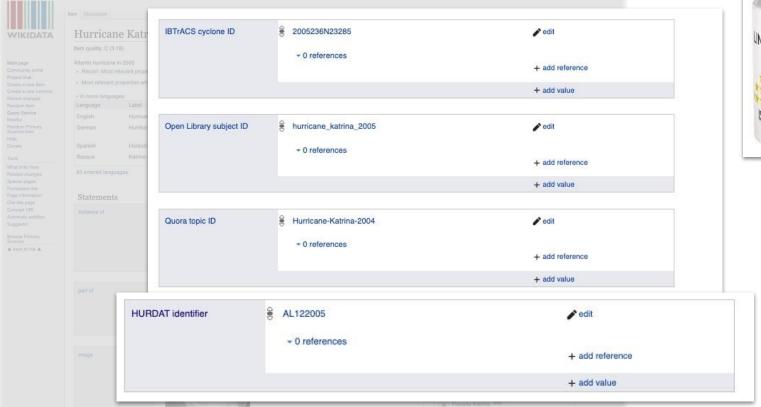








Integrated in the Web Dataspace



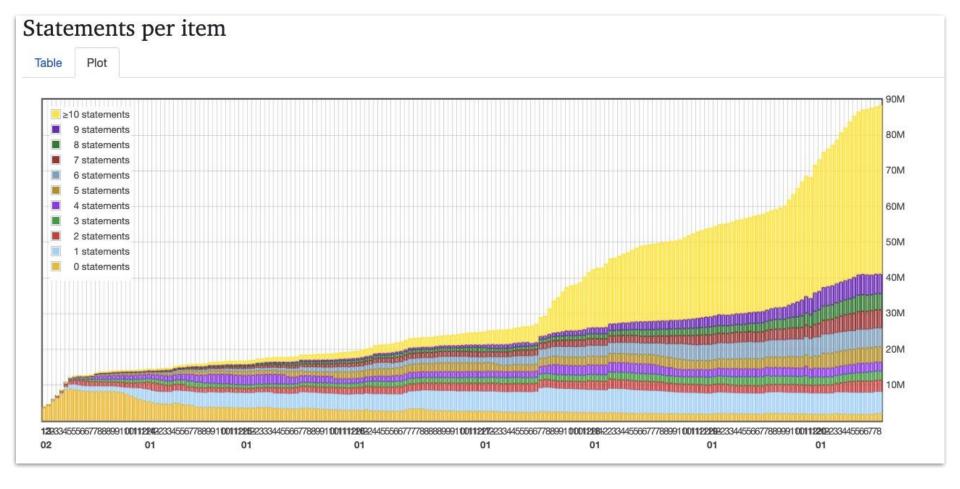


Querying to Compare

```
SELECT (COUNT(*) AS ?count)
where{
  ?item wdt:P31 wd:Q4022 . #river
  ?item wdt:P402 ?openstreetmapID
SELECT (COUNT(*) AS ?count)
where{
  ?item wdt:P31 wd:Q4022 . #river
  FILTER NOT EXISTS {?item wdt:P402 ?openstreetmapID}
SELECT (COUNT(*) AS ?count)
where{
  ?item wdt:P31 wd:O4022 . #river
  FILTER NOT EXISTS {?item wdt:P402 ?openstreetmapID}
  SERVICE <a href="https://sophox.org/sparql">https://sophox.org/sparql>
        { ?openstreetmapE <https://wiki.openstreetmap.org/wiki/Key:wikidata> ?item }
```

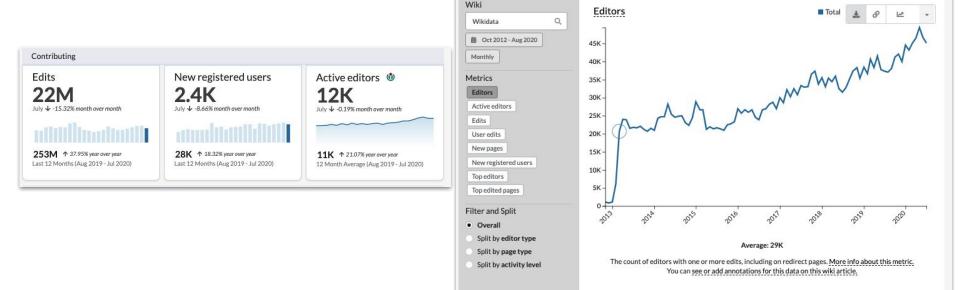
Querying to Compare

```
SELECT (COUNT(*) AS ?count)
where{
  ?item wdt:P31 wd:Q4022 . #river
  ?item wdt:P402 ?openstreetmapID
                                                  service<https://ld.stadt-zuerich.ch/query> {
                                                     ?obser a qb:Observation ;
SELECT (COUNT(*) AS ?count)
                                                      b:dataSet dataset:BEW-RAUM-ZEIT ;
where{
                                                      measure: BEW ?population ;
  ?item wdt:P31 wd:O4022 . #river
                                                      dimension: RAUM ?space ;
  FILTER NOT EXISTS {?item wdt:P402 ?openstree
                                                      dimension: ZETT ?time .
                                                    ?space rdfs:label ?spaceLabel.
                                                    FILTER(?space = code:R30000)
SELECT (COUNT(*) AS ?count)
                                                    FILTER (YEAR(?time) = ?year)
where{
 ?item wdt:P31 wd:O4022 . #river
 FILTER NOT EXISTS {?item wdt:P402 ?openstreetmapID}
 SERVICE <https://sophox.org/sparql>
       { ?openstreetmapE <https://wiki.openstreetmap.org/wiki/Key:wikidata> ?item }
```





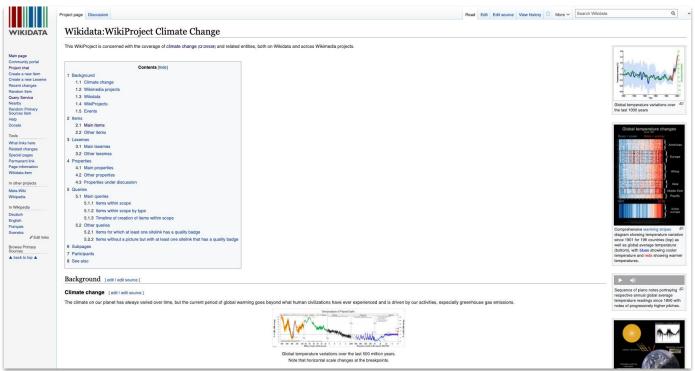
Editors & Edits



SOURCE: https://stats.wikimedia.org/#/wikidata.org/contributing/editors/normal|line|2012-10-21~2020-08-20|~total|monthly

Peer-Production

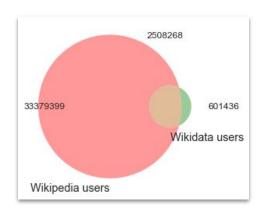
Goal-Oriented Collective Cooperative Action

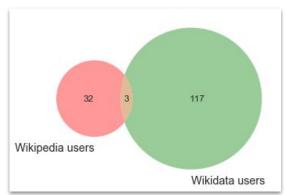


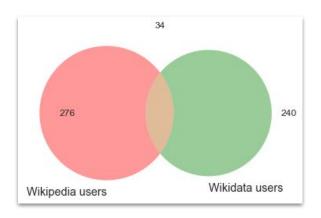
Browse live: https://www.wikidata.org/wiki/Wikidata:WikiProject Climate Change

Wikimedia Communities

As of April 2019







All Users

Users with more than 1M edits

Bots



Top 1K users (by volume of edits)

Types of Editors and User Groups



Unregistered



Human Registered



- New Users
- Autoconfirmed Users
- Administrators
- Bureaucrats
- Translation administrators
- Rollbackers
- Flooders
- Stewards
- Ombudsmen

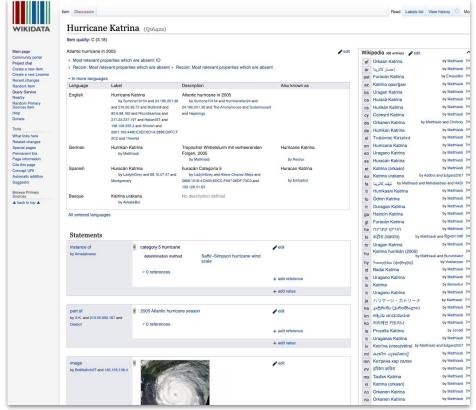


Bots

Transparency, Accountability, Recognition

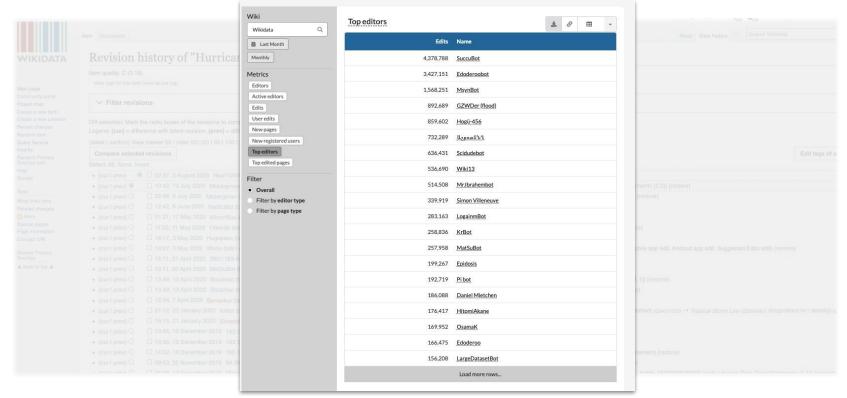


Transparency, Accountability, Recognition



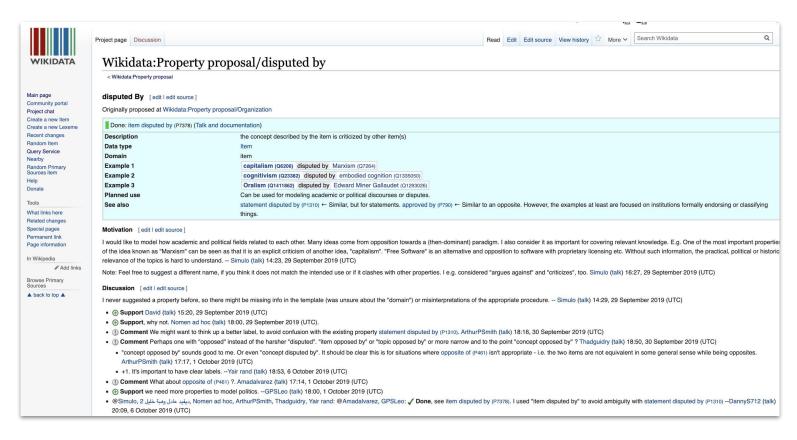
see also: https://www.wikidata.org/wiki/Wikidata:Tools/Enhance_user_interface

Transparency, Accountability, Recognition



SOURCE: https://stats.wikimedia.org/#/wikidata.org/contributing/top-editors/normal|table|last-month|~total|monthly

Deliberation: schema curation



Deliberation: Requests for Comments

The people who did their work on vandalism detection in Wikidata have it easier in academia when they can cite a definition of quality standards.

Nothing in this article prevents you from having a discussion about direct ways to improve Wikidata's usefulness for Wikipedia. But that's not the discussion about the definition of data quality, it's likely to be had elsewhere.

There are many ways to interact with Wikidata. Live and let live.

If you want to improve Wikipedia integration it might make more sense to focus your energies on the practical project of https://www.wikidata.org/wiki/Wikidata:List_generation_input@ ChristianKI (talk) 22:44, 31 August 2016 (UTC)

Academic work does not need to have direct practical implications, but — I believe — it should always take them into account and aim at concrete effects on things. I would not even find it totally correct to ask the Wikidata community to discuss this draft, if I did not think that it could be one day adopted (even partially) by Wikidata or at least be directly bencheficial to it. So yes, it will be good if the outcome of this work will be published, it will be either it is will bring someone to donate to Wikidata, and it will be even better if quality measures will be developed and then used based on this framework. And by the way, no. I do not earn Wikimedia grant; 1—Nessandro Piscoor (alik) 23.39, 2 September 2016 (UTC)

Completeness [edit | edit source]

Hi Alessandro, thanks for the RfC and your valiant participation in the discussion.

Many of the contributions to the discussion point to specific implementations that can help by finding errors easier - such as constraints on data values, or consistency rules. I think those are very important in order to keep control of the data set - but in the end, they are merely a proxy to measuring qualify. They do not really capture your qualify measures, such as 'accuracy' and 'completeness', etc.

While these suggestions and implementations are very important, I would be extremely happy to see you stay focused on the quality dimensions you mention. Your definitions look sane and good. Personally, I would weight these dimensions - e.g. have accuracy very high, consistency rather low, etc. - but your selection looks already quite complete, and I am not sure I would drop anything, besides on the completeness dimension.

For schema completeness, I would argue that since Wikidata does not have a standard class system, maybe we should focus on the properties only.

For item completeness, I would argue that focusing on just the classes and see whether they have all relevant properties is insufficient. My suggestion would be to compare to the Wikipedia article and see whether the most important information about a given item that is mentioned in the article is actually covered.

For population completeness, again I would argue against relying solely on classes.

All of the three subdimensions of completeness I would actually consider maybe complementing or partially replacing with query completeness. Given relevant queries, can we express these queries and do we get all results we are expecting. This would cover your subdimensions, and is also motivated by Wikidata's arguably most important role, to support the Wikimedia projects. Query results are planned to be available for integration and exposure in the other Wikimedia projects at some point (the legendrary Phase 3 of the Wikidata project), and such a query completeness would directly capture how realistic such a goal is.

I am looking forward to see your suggestions on how to operationalize these quality dimensions and the concrete metrics and measurements you will suggest. I hope this helps, and again, thank you! —Denny (talk) 17:47, 1 September 2016 (UTC)

Hi Denny, thanks for your suggestions! The completeness dimension is indeed that one that gave me more doubts (not that the others did not), due to its strong connection with the task at hand, which in my opinion requires to answer the question "What is Wikidata for?"

I agree that query completeness is a good approach; the authors of COOL-WD have already explored it, I would like to have a more careful look at it and think about how it can be improved.

I will start elaborating the metrics in the next days, taking into account the discussions had on this page. As soon as I will devise the most appropriate metrics for the dimensions chosen, I will again update the community again on our findings.

Thanks, "Alessandro Piscopo (talk) 23:26, 2 September 2016 (UTC)

Thanks, that sounds good! -- Denny (talk) 15:02, 6 September 2016 (UTC)

Completeness heuristic by comparing to Wikipédia internal links graphs [edit | edit source]

Thanks to the Wikipédia articles / Items mapping, we could compare the graphs generated by the internal links in wikipedias and the wikidata graph. For example, I expect that for a complete Item, almost every articles linked in a language version of wikipédia (or the union of all versions) be very close - say, one or two statements away - of the items used in the corresponding Wikidata item or that uses it. Maybe a very simple proportion of such items that uses the items we want to know is complete union all the Items used by it compared to the items of the stellaries of the corresponding article could be a good heurstool to evaluate the completeness of an item writ. Wikipedia. author. Tom'T0m / talk page 13:29, 7 September 2016 (UTC)

Hi TomT0m, yours is a good idea, I would be curious to try it, in order to understand how feasibile it is (e.g. I think that using a union of all version may be challenging) and the insights that it would give. My doubt is whether it would shed some light on the degree on interlinking of entities in the two projects, rather than providing information about completeness. Thanks, --Alessandro Piscopo (talk) 13:05, 13 September 2016 (UTC)

Wikidata:Living people [edit | edit | source]

I will not support any proposal that does not include provisions for the protection of information regarding living people. Wikipedias will be less likely to trust us without such a provision.--Jasper Deng (talk) 00:11, 28 October 2016 (UTC)

References [edit | edit source]

1. 1 Lukyanenko, R., Parsons, J., & Wiersma, Y. F. (2014). The IQ of the crowd: understanding and improving information quality in structured user-generated content. Information Systems Research, 25(4), 669-689.

SOURCE: https://www.wikidata.org/wiki/Wikidata:Requests for comment/Data quality framework for Wikidata

Deliberation: Requests for Permissions

82% of bot RfPs were approved (323).

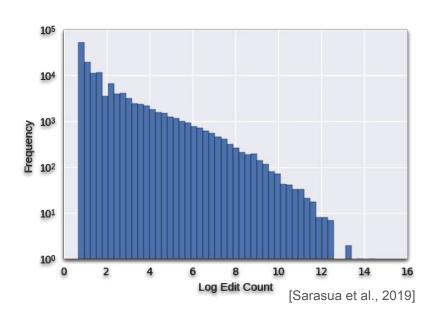
Most frequent reasons to reject RfPs: "non-reactive operators, withdrawn and duplicated requests, no community consensus".

[Farda-Sarbas et al., 2019]

Human-Machine Cooperation

"A balanced contribution of bots and human editors positively influence outcome quality, whereas higher numbers of anonymous edits may hinder performance."

Challenges



"2% of anonymous unregistered human user edits overall are from bots", but also "3% of registered human users edits overall."

[Hall and Halfaker., 2018]

Addressing Participation Inequalities

Governance

Wikibase: Using the Infrastructure

- MediaWiki
- Data Model
- SPARQL endpoint
- QuickStatements to import data

Registry of online Wikibases



Wikibase installation video by Andra Waagmeester



Cristina Sarasua < sarasua@ifi.uzh.ch > Daniel Mietchen < dm7gn@virginia.edu >





Image by Addshore, CC0