WikiMetaData Studio: Dashboards From Data Profiling the Languages, Properties, and Items of Wikidata *

Niel Chah[1] [0000−0002−3377−7823] and Periklis Andritsos[2]

University of Toronto, Faculty of Information
[1]niel.chah@mail.utoronto.ca,[2]periklis.andritsos@utoronto.ca

Abstract. Wikidata is a large collaborative knowledge graph, containing multilingual data for hundreds of languages across millions of entities (items) using thousands of properties. To explore research questions regarding the general state of data on Wikidata (e.g. multilingual research about potential biases or missing data in certain languages), a comprehensive data profile of Wikidata is first needed. This paper presents a data profiling framework that parses the entire Wikidata data dumps to produce a series of granular descriptive statistics that summarize the full extent of data on Wikidata. This method addresses the limitations of using existing SPARQL API querying methods. The output from the data profiling framework is presented through a series of interactive dashboards using Google Data Studio. Future work using output from the data profiling is also proposed.

Keywords: Wikidata · Data profiling · Multilingual data

1 Introduction

What languages are the most widely used for labels, aliases, and descriptions in Wikidata? Which ones are the least? What properties are most frequently used in statements, in qualifiers, and in references? How many statements state that there is "some value" or definitely "no value"? What kinds of subject matter topics are prevalent in Wikidata? These are some of the questions that motivate the data profiling framework and the resulting dashboards that are described in this paper.

Since its launch in 2012, Wikidata is a widely used collaborative knowledge graphs (KG) with data populated in hundreds of languages [5]. Each entity, or *item*, in Wikidata is associated with a unique "Q" identifier. The information for each Wikidata item is shown on a dedicated web page which also lists the labels (i.e. names), descriptions, and aliases that describe it across the languages supported by Wikidata. For instance, the Wikidata *item* for the International Semantic Web Conference is "Q6053150", and its Wikidata page is populated with labels, descriptions, and aliases in various languages in addition to the many properties that link it to other items and literal values.[1]

As KGs like Wikidata continue to grow in volume and usage, it is important to explore Wikidata's data coverage across international languages and communities. This can be seen in initiatives such as the "Whose Knowledge?" Wikimedia campaign[2] and the "Reimagining Wikidata from the margins"[3] initiative ahead of WikidataCon 2021. This paper presents ongoing research that seeks to answer questions on the distribution of multilingual data on Wikidata and eventually pose recommendations for data augmentation and enrichment of the knowledge graph.

2 Related Works

To our knowledge, the most recent comprehensive research study on the multilingual data on Wikidata was done in 2017 by Kaffee et al. [2]. Their paper looked at the language distribution for labels, excluding descriptions, aliases, and other sources of natural language data (see Figure 1). As of 2017, data for Wikidata labels were most populated for the English, Dutch, French, German, Spanish, Italian, Swedish, and Russian languages. This paper will present a view into the distribution of language data and further detailed statistics as of June 30, 2021.

Fig. 1. Distribution of label data in different languages in Wikidata, March 2017 from Kaffee et al. [2]

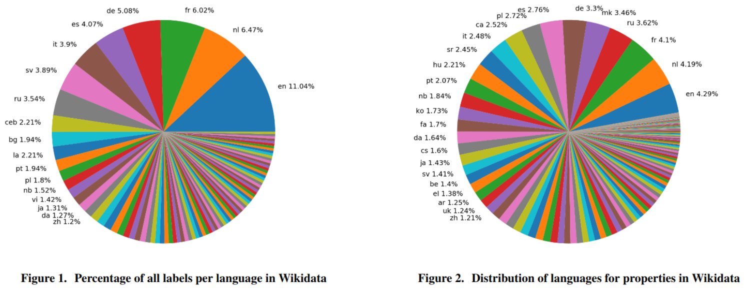


Figure 1. Percentage of all labels per language in Wikidata

Figure 2. Distribution of languages for properties in Wikidata

In addition, many interactive data visualization tools and dashboards have been created to describe certain aspects of the multilingual information on Wikidata, with each tool uniquely tailored for a specific objective.[4] An all-purpose

[1] `https://www.wikidata.org/wiki/Q6053150`

[2] `https://meta.wikimedia.org/wiki/Whose_Knowledge`

[3] `https://www.wikidata.org/wiki/Wikidata:Reimagining_Wikidata_from_the_margins`

[4] `https://www.wikidata.org/wiki/Wikidata:Tools/Visualize_data`

tool that supports all possible features to explore Wikidata exhaustively is not yet extant, nor proposed as the scope of this paper. For instance, the Wikidata Languages Landscape[5] dashboard provides average descriptive statistics on language data in the Wikidata ontology and for a few select items (or entities). ProWD[6] offers interactive dashboards that display the level of "completeness" of Wikidata items at a class-level, but does not look into the multilingual data of those items [6]. Furthermore, WDProp[7] is a dashboard that visualizes the frequency of translated label, description, and alias values in the Wikidata ontology, but not other items [3].

3 Methodology

3.1 SPARQL query limitations

Conventional methods of obtaining descriptive statistics from Wikidata rely on writing SPARQL queries to the Wikidata API endpoint or Wikidata Query Service.[8] However, queries on the scale required for comprehensive data profiling described in this paper are not possible as the API times out due to the computational complexity of even optimized queries.

For reference, the following SPARQL query that gets all scholarly articles (Q13442814) that have a missing French label (`https://w.wiki/4C8H`) fails with a "Query timeout limit reached" timeout error. On the other hand, the same query that gets all doctoral thesis items (Q187685) with a missing French label (`https://w.wiki/4C8K`) successfully completes. This discrepancy may be explained by the significantly larger amount of scholarly article entities compared to doctoral theses. Using the output from the WikiMetaData Profiler as of the June 30, 2021 data dump, it was found that scholarly articles (Q13442814) made up over 34 million items or 38% of all items that have a value using the P31 "instance of" property. While the possible reason for the timeout errors may have been uncovered, this does not address the root issue that SPARQL queries fail for large expansive queries that may very well be the *raison d'être* for their use.

An alternative is to run a local machine, or cloud computing setup to ingest the entire Wikidata data dumps and run queries. However, with these alternatives a consideration that must be kept in mind is the cost of the hardware and software setups. The large size of the data dumps makes it necessary for a large disk and memory size to be used in order for a local solution be viable. The June 30, 2021 gzipped data dump was >100GB in its compressed form, and the latest October 6, 2021 data dump is over 106 GBs [9]. A documented example where the entire Wikidata data dumps were loaded into Apache Jena, an open-source

[5] `https://meta.wikimedia.org/wiki/Wikidata_Languages_Landscape`

[6] `https://prowd.id/#about`

[7] `https://wdprop.toolforge.org/wdprop.html`

[8] `https://query.wikidata.org/`

[9] `https://dumps.wikimedia.org/wikidatawiki/entities/`

triplestore, used Intel Xeon CPUs with 32 cores and 128 GBs of RAM and almost 700 GB of disk space.[10]. As this was done in 2019, more recent data dumps would require commensurately higher technical requirements owing to the larger data dump size.

Plugging the aforementioned requirements into the pricing calculators[11] of popular cloud service providers yields a monthly cost in the range of hundreds of US dollars per week. As experiments and queries are run on these instances over weeks and months, the costs would also scale accordingly. As a a result, this would limit the availability of these tools to researchers and institutions with such resources.

3.2 Data profiling framework and pipeline for generating dashboards.

Regular data dumps of Wikidata are provided by the Wikimedia Foundation online in various formats. The June 30, 2021 Wikidata JSON dump (>100GB in its compressed form) was profiled using the data profiling framework, which is released as open-source code on GitHub.[12] More recent data dumps may also be downloaded and processed through the same pipeline in order to reproduce the results using the code on GitHub.

The code operates in the following manner. Within the nested JSON structure of the data dump, each line contains all of the key: value (or *property: object*) pairs for a single Wikidata item, which correspond to the property and object values in a $(subject, property, object)$ triple. As the entirety of the key: value data is parsed, certain predefined fields and statistics of interest (languages (labels, aliases, descriptions), properties, classes of items, qualifiers, and references) are captured in memory, updated in memory, and eventually written to output files. More precisely, the data that was maintained in memory is first formed into Python `pandas` DataFrames and then written to tab-separated value (TSV) output files [4,**?**]. These output files were uploaded to Google Data Studio (GDS) and presented through a series of data visualizations and interactive dashboards.[13] To monitor the profiling progress, the `tqdm` progress bars were used [1].

The data profiler is able to run on widely available consumer hardware setups, such as commercially sold laptops and desktops. Most computations were run using Jupyter notebooks on two local machines: (1) running Ubuntu 18.04, with 8 CPU cores and 16 GB memory, and (2) running macOS with 6 CPU cores and 16 GB memory. The Google Colaboratory notebooks, each with 2 CPU cores

[10] (non-English:) `https://muncca.com/2019/02/14/wikidata-import-in-apache-jena/`

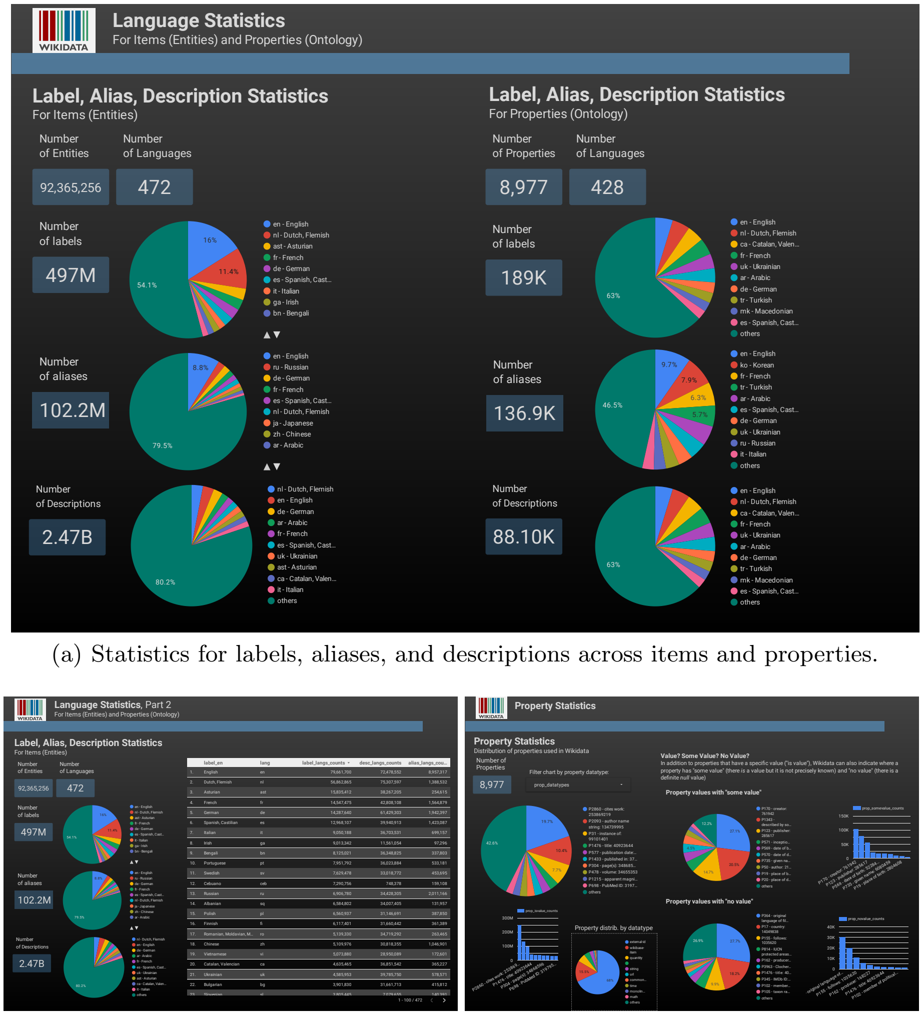
[11] Virtual machines at `https://azure.microsoft.com/en-us/pricing/calculator/` and EC2 instances at `https://calculator.aws/#/createCalculator/EC2`

[12] `https://github.com/nchah/wikidata-profiler`

[13] `https://datastudio.google.com/reporting/7f6f76eb-c24f-4a1d-b7a0-86fc4c2a55c4`

and ~13 GB memory, were also utilized to run further parallel experiments. The approximate time to run this on the macOS machine was ~6 hours.

Fig. 2. Screenshots of the output from the data profiling framework using interactive dashboards on Google Data Studio.



(a) Statistics for labels, aliases, and descriptions across items and properties.

	label_en	lang	label_langs_counts ▾	desc_langs_counts	alias_langs_cou...
1.	English	en	79,661,700	72,478,552	8,957,317
2.	Dutch, Flemish	nl	56,862,865	75,307,597	1,388,532
3.	Asturian	ast	15,835,412	38,267,205	254,615
4.	French	fr	14,547,475	42,808,108	1,564,879
5.	German	de	14,287,640	61,429,303	1,942,397
6.	Spanish, Castilian	es	12,968,107	39,940,913	1,423,087
7.	Italian	it	9,050,188	36,703,531	699,157
8.	Irish	ga	9,013,342	11,561,054	97,296
9.	Bengali	bn	8,125,021	36,348,825	337,803
10.	Portuguese	pt	7,951,792	36,023,884	533,181
11.	Swedish	sv	7,629,478	33,018,772	453,695
12.	Cebuano	ceb	7,290,756	748,378	159,108
13.	Russian	ru	6,906,780	34,428,305	2,011,166
14.	Albanian	sq	6,584,802	34,007,405	131,957
15.	Polish	pl	6,560,937	31,146,691	387,850
16.	Finnish	fi	6,117,401	31,660,442	361,389
17.	Romanian, Moldavian, M...	ro	5,139,330	34,719,292	263,465
18.	Chinese	zh	5,109,976	30,818,355	1,046,901
19.	Vietnamese	vi	5,073,880	28,950,089	172,601
20.	Catalan, Valencian	ca	4,635,465	36,851,542	365,227
21.	Ukrainian	uk	4,585,953	39,785,750	578,571
22.	Bulgarian	bg	3,901,830	31,661,713	415,812
23.	Slovenian	sl	3,805,445	7,079,655	140,391

1 - 100 / 472

Property Statistics
Property Statistics
Distribution of properties used in Wikidata
Number of Properties
8,977
Filter chart by property datatype:
prop_datatypes
19.7%
10.4%
7.7%
42.6%
P2860 - cites work: 253869219
P2093 - author name string: 134739995
P31 - instance of: 99101401
P1476 - title: 40923644
P577 - publication date:...
P1433 - published in: 37...
P304 - page(s): 348685...
P478 - volume: 34655353
P1215 - apparent magni...
P698 - PubMed ID: 3197...
others
prop_isvalue_counts
300M
200M
100M
0
P2860 - cites work: 253869...
P1476 - title: 40923644
P304 - page(s): 34868596
P698 - PubMed ID: 319795...
Property distrib. by datatype
15.5%
68%
external-id
wikibase-item
quantity
string
url
common...
time
monolin...
math
others

Value? Some Value? No Value?
In addition to properties that have a specific value ("is value"), Wikidata can also indicate where a property has "some value" (there *is* a value but it is not precisely known) and "no value" (there is a definite *null* value)

(b) A further detailed page for languages.

(c) A detailed page for properties.

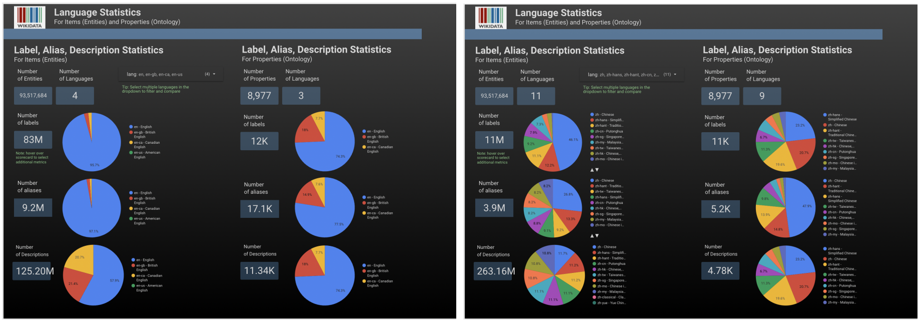
4 Findings

From the interactive Google Data Studio dashboards, the following results that are relevant to language locales are described in further detail. The full set of interactive dashboards allows for additional findings to be made across different aspects of Wikidata.

Distribution of multilingual data, properties, classes of items, qualifiers, and references. A selection of the Google Data Studio dashboards is shown in Figure 2. In the full GDS dashboard, descriptive statistics are shown for the distribution of different language data used in labels, aliases, and descriptions for items and properties (part of the ontology). To discover what kinds of entities exist in Wikidata, a view into the different classes of items using the P31 property ("instance of") is also presented. The distributions of properties that are used for statements, qualifiers, and references are also shown.

Interactive comparison of the distribution of multilingual data. It is also possible to use the dashboards to compare the distribution of values across language variants or dialects. For instance, in Figure 3, the various dialects or variations of a larger language are compared for Chinese (zh) and English (en) dialects. These dialects are shown by a "-" (dash) and additional language code. Across items, the Chinese (zh) dialects are almost evenly distributed for aliases and descriptions, with a slight difference in the even distribution for labels. This is unlike the distribution of significantly fewer localized values in the dialects under English (en): British English (en-gb), Canadian English (en-ca), and American English (en-us).

Fig. 3. Screenshots of the interactive dashboards on Google Data Studio using outputs from the data profiling framework.



Language Statistics
For Items (Entities) and Properties (Ontology)

Label, Alias, Description Statistics
For Items (Entities)

Number of Entities: 93,517,684
Number of Languages: 4
lang: en, en-gb, en-ca, en-us (4)
Tip: Select multiple languages in the dropdown to filter and compare
Number of labels: 83M
Note: hover over scorecard to select additional metrics
Number of aliases: 9.2M
Number of Descriptions: 125.20M

Label, Alias, Description Statistics
For Properties (Ontology)

Number of Properties: 8,977
Number of Languages: 3
Number of labels: 12K
Number of aliases: 17.1K
Number of Descriptions: 11.34K

Language Statistics
For Items (Entities) and Properties (Ontology)

Label, Alias, Description Statistics
For Items (Entities)

Number of Entities: 93,517,684
Number of Languages: 11
lang: zh, zh-hans, zh-hant, zh-cn, z... (11)
Tip: Select multiple languages in the dropdown to filter and compare
Number of labels: 11M
Note: hover over scorecard to select additional metrics
Number of aliases: 3.9M
Number of Descriptions: 263.16M

Label, Alias, Description Statistics
For Properties (Ontology)

Number of Properties: 8,977
Number of Languages: 9
Number of labels: 11K
Number of aliases: 5.2K
Number of Descriptions: 4.78K

(a) A comparison of English (en) dialects. (b) A comparison of Chinese (zh) dialects.

Dominant languages continue to stay largely dominant, with a few upstart changes. Among the various dashboards, a number of general findings on the languages is described. According to the data profiling work done as of 2017, prominent languages for label data included English (en) at 11.04%, Dutch (nl) at 6.47%, French (fr) at 6.02%, German (de) at 5.08%, and Spanish (es) at 4.07% (see Figure 1, from earlier in this paper) [2]. This approximate order is found once more in a relatively recent snapshot of Wikidata, with the addition of the Asturian (ast) language occupying the third most frequent language for labels and Irish (ga) and Bengali (bn) in the eighth and ninth positions respectively. The relative proportion of all labels values have also increased for many of the top languages, such as English (en) reaching 16.4% compared to 11.04%. As interesting as the rise in prominence of certain languages may be, the main reasons for their rise (e.g. active bots or a rising community of users) cannot be uncovered from the current data dumps.

5 Limitations

While the Wikidata data dumps are comprehensive, they do not capture other important aspects of the Wikidata system. The myriad of users and bots that contribute information to Wikidata are not tracked in the data dumps. As such, data such as the provenance of Wikidata contributions, the speed and frequency of edits (and reversions), and the general degree of activity of users are not easily obtained from a dedicated data dump. Furthermore, the active discussions and coordination that take place on the “Talk” pages of Wikidata pages are not captured.

This kind of provenance data would have a potential application in determining, for example, the upstart rise of Wikidata labels, descriptions, and aliases in Asturian (ast) and other previously underrepresented languages (Irish (ga) and Bengali (bn)). Were bots or automatic processes responsible for the massive influx of data? Or, were there grassroots community movements, hackathons, or other local initiatives that brought together people to add data to the languages?

6 Conclusion and Future Work

This paper presented early work on a data profiling framework and pipeline to parse the entire Wikidata data dumps and visualize the output in interactive Google Data Studio reports. In a forthcoming paper, the output data from this data profiling framework was used to generate heatmaps that depict the distribution of multilingual data (labels, aliases, and descriptions) across machine learned and human annotated topical domains in Wikidata. These heatmaps are then useful for determining where multilingual data is concentrated and sparsely populated.

Future work will use the output from the data profiling framework for research questions on the current state of Wikidata (e.g. with questions on the prevalence of language data for various classes of items, and questions on the frequency of “high quality” properties that use qualifiers and references) and machine learning approaches to programmatically recommend data enrichment regimes (e.g. using knowledge graph embeddings for targeted link prediction in certain under-served languages).

Acknowledgements This research was supported by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Costa-Luis, C.d., Larroque, S.K., Altendorf, K., Mary, H., richardsheridan, Korobov, M., Yorav-Raphael, N., Ivanov, I., Bargull, M., Rodrigues, N., CHEN, G., Lee, A., Newey, C., James, Coales, J., Zugnoni, M., Pagel, M.D., mjstevens777, Dektyarev, M., Rothberg, A., Alexander, Panteleit, D., Dill, F., FichteFoll, Sturm, G., HeoHeo, Kemenade, H.v., McCracken, J., MapleCCC, Nordlund, M.: tqdm: A fast, Extensible Progress Bar for Python and CLI (Sep 2021). https://doi.org/10.5281/zenodo.5517697, `https://doi.org/10.5281/zenodo.5517697`
2. Kaffee, L.A., Piscopo, A., Vougiouklis, P., Simperl, E., Carr, L., Pintscher, L.: In: Proceedings of the 13th International Symposium on Open Collaboration - OpenSym '17. pp. 1–5. ACM Press (2017). https://doi.org/10.1145/3125433.3125465, `http://dl.acm.org/citation.cfm?doid=3125433.3125465`
3. Samuel, J.: Towards understanding and improving multilingual collaborative ontology development in wikidata. In: Companion of the The Web Conference 2018 on The Web Conference. pp. 23–27 (2018)
4. team, T.p.d.: pandas-dev/pandas: Pandas (Feb 2020). https://doi.org/10.5281/zenodo.3509134, `https://doi.org/10.5281/zenodo.3509134`
5. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (Sep 2014). https://doi.org/10.1145/2629489, `http://dl.acm.org/citation.cfm?doid=2661061.2629489`
6. Wisesa, A., Darari, F., Krisnadhi, A., Nutt, W., Razniewski, S.: Wikidata Completeness Profiling Using ProWD. In: Proceedings of the 10th International Conference on Knowledge Capture - K-CAP '19. pp. 123–130. ACM Press, Marina Del Rey, CA, USA (2019). https://doi.org/10.1145/3360901.3364425, `http://dl.acm.org/citation.cfm?doid=3360901.3364425`