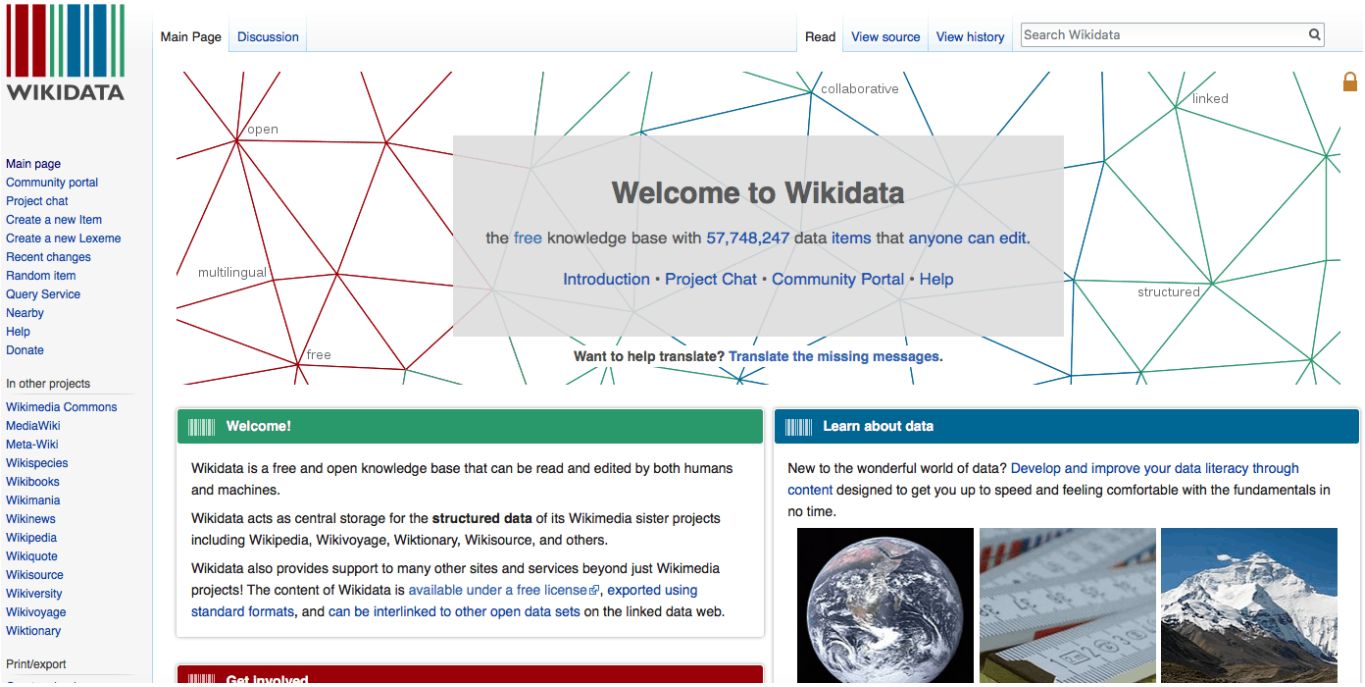
SocietyByte
Wissenschaftsmagazin des BFH-Zentrums Digital Society
https://www.societybyte.swiss

How to estimate completeness of classes in Wikidata

Autoren : Michael Luggen, Djellel Difallah, Cristina Sarasua, Ginaluca Demartini, Philippe Cudré-Mauroux

Datum : 5. Juli 2019



The way in which general encyclopaedic knowledge is collected, curated and used has changed dramatically. Anyone in the world can edit Wikipedia at any time, as a registered user or anonymously. This happens in parallel in different languages. To include all even minority languages and avoid a misalignment of information, the Wikimedia community has created Wikidata, write our authors in this article.

With the introduction of Wikipedia, the way common encyclopedic knowledge is collected, curated, and accessed by the majority of people vastly changed. Anyone in the world can edit Wikipedia any time, as a registered user or anonymously. The paradigm switch from an authority-based process to a democratic and self-organized collaborative process opened a wide range of new opportunities, including, for instance, the potential for a less biased narrative thanks to the crowdsourcing nature of the project, more up-to-date information and broader information coverage. However, Wikipedias in different languages evolved independently, and often topics being edited in, e.g., the English Wikipedia were not updated simultaneously to other Wikipedias, leading to a misalignment of information, leaving especially smaller Wikipedias from minority languages behind.

To address this problem, the Wikimedia community created Wikidata, a central and multilingual knowledge base, containing structured data that Wikipedia and other Wikimedia projects may consume to gain general information. Wikidata operates with similar community principles and

wiki infrastructure as Wikipedia; it is also collaboratively curated and maintained by a large community of thousands of volunteers. With currently more than 55M data items and over 5.4K properties that help describe these data items, Wikidata is linked to many Wikimedia projects (e.g. Wikipedia, Wikimedia Commons, and Wiktionary), and has become the new interlinking hub and center of the Linked Open Data Cloud connected to sources such as Europeana, VIAF, and OpenStreetMap. Wikidata's data is also consumed by end-user applications such as Google Search, Siri, and applications to browse scholarly information. Given the attention that Wikidata has received from data consumers and the impact that it has on Wikipedia, it is important to know if its data is complete and hence useful.

The Challenge: Is it Complete?

Being a collaborative, crowdsourced effort, Wikidata's data is highly dynamic. Editors can create items individually (e.g., a new instance representing a natural disaster that just happened), or in bulk (e.g. importing data about all the pieces of art in a city) about any topic that satisfies the notability criteria defined by the community. The open curation process leads to a knowledge graph evolving dynamically and at various speeds. Moreover, the large community is composed of people with diverse backgrounds, knowledge and interests, and people edit what they know best. While such a process is beneficial for data diversity and freshness, it does not guarantee the total (or even partial) *completeness* of the data. Previous research (Wang et al., 1996) has shown that data consumers identify completeness as one of the key data quality dimensions, together with accuracy and freshness.

With such a decentralized approach of independently-run data entry and import efforts, it has become very difficult to understand and measure what is still missing in Wikidata. The Wikidata community has already endorsed a series of initiatives and tools that encourage efforts towards population completeness. For instance, there are Wikiprojects that aim at populating Wikidata with bibliographic references, genes, or notable women. However, it is still challenging to know to what extent Wikidata has the complete list of mountains, municipalities of Switzerland, or volcanoes. This uncertainty may hinder data consumers from trusting and using the data. Hence, it is of utmost importance to provide mechanisms to measure and foster data completeness in such a collaborative knowledge graph.

Leveraging the Edit History of Wikidata

We conducted a research project (see Luggen et al., 2019) with the focus on the specific problem of *estimating class completeness* in a collaborative knowledge graph and experimentally evaluated our methods in the context of Wikidata. Given a finite class of instances (e.g. *Observation Towers*), our goal was to estimate the number of distinct entities of this class.

The field of ecology and bio-statistics has defined several so-called *capture-recapture* methods to estimate the number of existing species (Bunge and Fitzpatrick, 1993). These methods draw a sample at random from a population and estimate the number of unobserved items based on the frequency of the observed items. Inspired by this field, we took a data-driven approach to

solving the problem of estimating class completeness by leveraging capture-recapture statistical models in the context of Wikidata. We used the data in the knowledge graph, as well as the complete edit history of Wikidata, that states for each edit (i.e., an addition, change or deletion) the time at which the edit occurred, the item / property / page affected, and the description of the action.

On this basis we calculated the cardinality of classes and built estimates for the class convergence to the true value. By calculating the expected class cardinality, we were able to measure class completeness given the number of instances currently present in the knowledge graph for that class. We specifically looked at the problem of estimating the size of a given single domain class (e.g., Volcanos) or composite classes (e.g., Paintings created by Vincent van Gogh) in Wikidata. We thereby limited our methods to the family of finite classes, i.e. to classes with a fixed number of instances. A class is, by definition, complete once the count of distinct instances is equal to the class size.

The capture-recapture data collection protocol that we followed is based on observations recorded in a series of n periods of time that we refer to as *sample periods*. We considered the edit history of Wikidata (containing all edits that all users have done over time), sliced it in sample periods and automatically identified the frequency classes appear in the edits – e.g., if a user makes an edit to add the information that *Paris* is the owner of the *Eiffel Tower*, that edit leads to one mention of the class monument and one mention of the class city (Figure 1). Each mention is composed of an instance (i.e. the entity belonging to a class), the class the instance belongs to, and a timestamp.

The collected mentions served as input for our class size estimators.

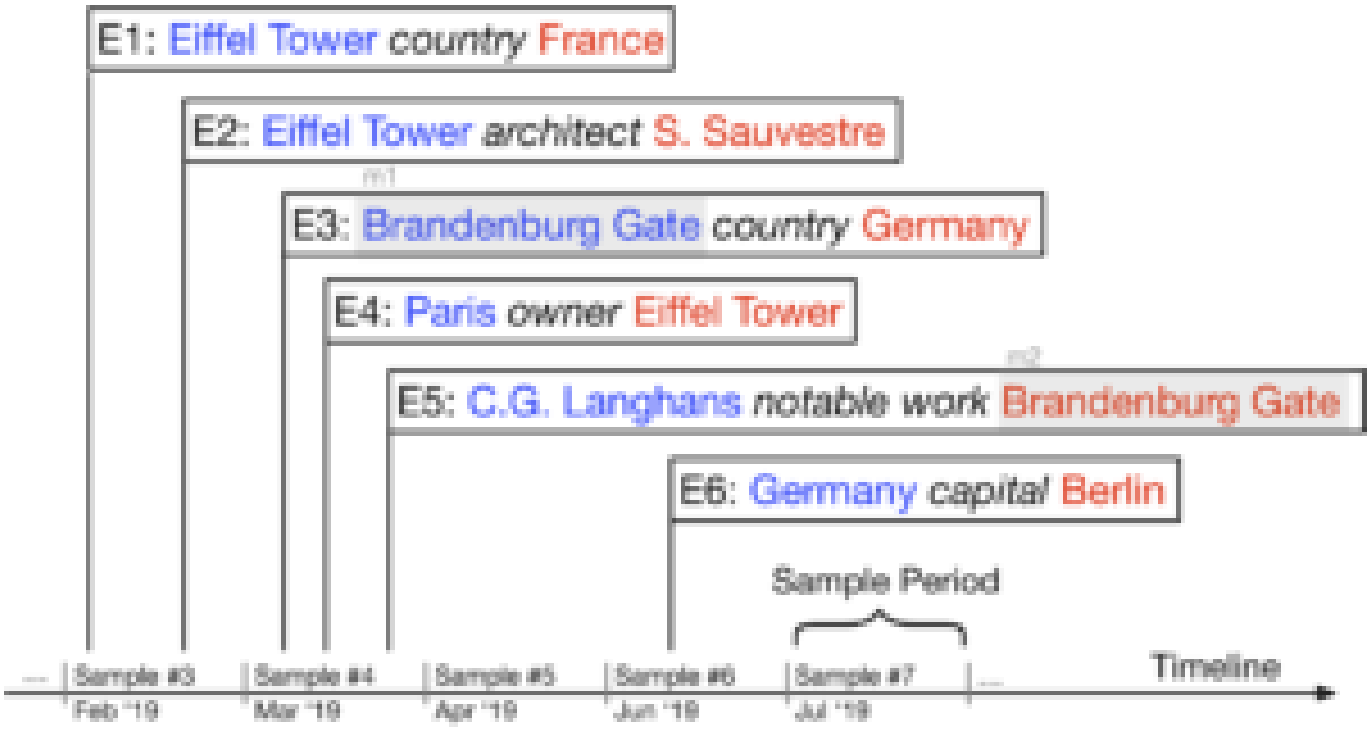


Figure 1: Every time an edit (E1-E6) occurs, we collect mentions for every referenced instance which belongs to a class. The mentions are used as signals for the estimators.

Class Size Estimators

To estimate class size and, thus, its completeness, we considered non-parametric methods that primarily use the frequencies of observations among instances.

Wikidata provides information about how instances are related to classes (e.g., the item *Eiffel Tower* belongs to the class *Observation Tower*). Every time editors change information about an item description, the edit history traces these edits. Looking at the edit history, we can interpret observations about the classes, analysing the edits that users implement on instances of classes. A recently added class which is not yet complete will have instances with only one or few observations. In contrast, in a class that is close to completion, most instances already have multiple observations. To estimate class size, we tested several estimators which take into account how many times (i.e., frequency) an instance of a class was observed over the available samples:

- **Jackknife Estimator** [Jack1] removes a sample period from the data, computes a pseudo estimate for the remaining sample periods and averages across all of them.
- **Sample Coverage and the Good-Turing Estimator** [N1-UNIF] focuses on the ratio of the number of instances that have been observed so far.
- **Singleton Outliers Reduction** [SOR] aims at balancing between low and high dispersion of frequencies of classes of size 1 with respect to other frequency counts.
- **Abundance-based Coverage Estimator** [Chao92] assumes that the capture probabilities can be summarized by their mean and their coefficient of variation.

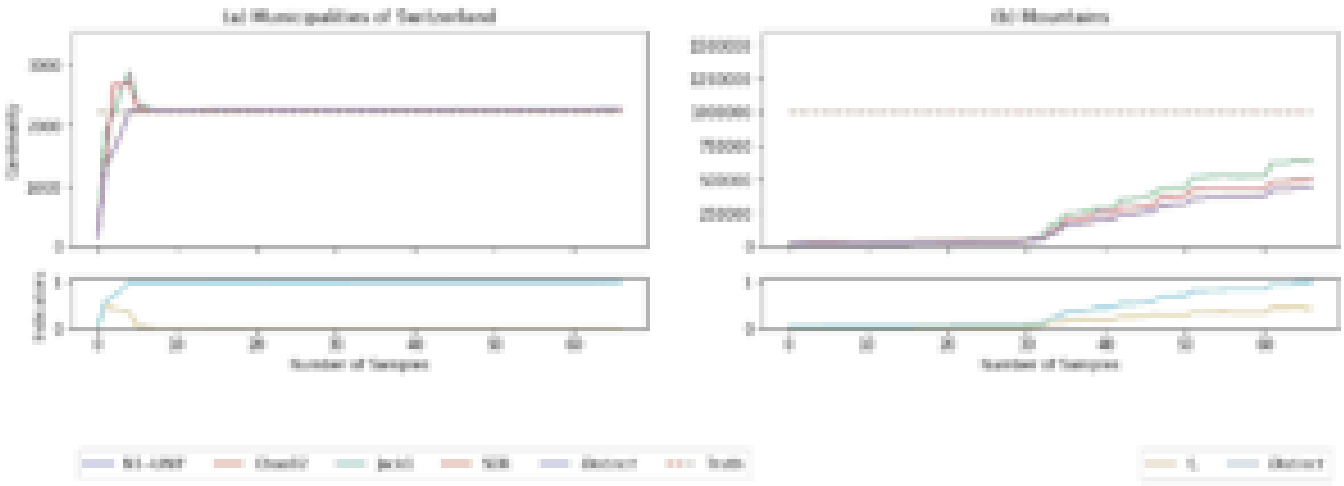


Figure 2 : The performance of the estimators is illustrated on the two classes – "Municipalities of Switzerland", which was completed early on, and "Mountains", which is still growing (as of August 2018). The upper graph shows the estimated number of instances and compares this to

our ground truth (Truth). The lower graph shows the development of the number of instances with only one observation (f1). Both graphs show the current number of observed instances (Distinct).

To evaluate the different class size estimators, we used two measures:

- **Error Metric** (?) that aims at capturing the bias of the estimates as the absolute distance from the ground truth when known.
- **Convergence Metric** (?) that aims at evaluating the convergence of a given estimate, acting as the main measurement tool in a real scenario where we do not have access to the ground truth. The closer the metric is to zero, the more confident we are that the class has converged to its complete set.

Our experimental results (Figure 2) unveiled key properties in terms of the sensitivity and conditions under which some estimators perform better than others. Generally speaking, all estimators beat the lower bound of distinct numbers in the error metric ? almost for all classes investigated. We observe that more conservative estimators like N1-UNIF or Chao92 perform worse than Jack1 and SOR for incomplete classes, which is why we recommend the last two as estimators of class size.

Our evaluation showed that convergence metrics ? are low (0.1). Hence, our convergence metric can be leveraged to identify gaps in the knowledge graph (see Table 1 for examples from Wikidata).

	SOR $\rho < 0.001$	Distinct		SOR $\rho > 0.1$	Distinct
municipality of Japan	0.0000	739	urban beach	0.1759	683
Philippine TV series	0.0009	822	hydroelectric power station	0.2975	2,936
Landgemeinde of Austria	0.0000	1,116	aircraft model	0.1800	3,919
district of China	0.0009	975	motorcycle manufacturer	0.1758	690
nuclear isomer	0.0002	1,322	local museum	0.1760	1,150
international border	0.0000	529	waterfall	0.1942	5,322
commune of France	0.0001	34,937	race track	0.2783	946
village of Burkina Faso	0.0005	2,723	film production company	0.2107	2,179
supernova	0.0005	5,906	red telephone box	0.3469	2,716
township of Indiana	0.0002	999	mountain range	0.2390	21,390

Table 1: Randomly selected examples of classes which seem to be complete (?0.1) inside Wikidata as of August 2018.

Conclusions

Our experimental results show that many estimators yield accurate estimates when provided

with enough observations that reflect the actual underlying distribution of the instances of a class. Our work has direct implications for both Wikidata editors and data consumers. We can provide convergence statistics on the estimated class completeness by domains to point to knowledge gaps. Such statistics could aid newcomers, who often feel insecure about what to edit, to decide what to contribute or what to focus on.

Acknowledgment

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 683253/GraphInt).

References

1. Bunge, J., Fitzpatrick, M. (1993): Estimating the number of species: a review. Journal of the American Statistical Association 88(421), 364–373.
2. Luggen, M., Difallah, D., Sarasua, C., Demartini, G., Cudré-Mauroux, P. (2019): Non-Parametric Class Completeness Estimators for Collaborative Knowledge Graphs - The Case of Wikidata. To Appear In: International Semantic Web Conference 2019.
3. Wang, R.Y., Strong, D.M. (1996): Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems 12(4), 5–33.