Property Label Stability in Wikidata

Thomas Pellissier Tanon, Lucie-Aimée Kaffee

Property Label Stability in Wikidata

Evolution and Convergence of Schemas in Collaborative Knowledge Bases

Thomas Pellissier Tanon*
LTCI, Télécom ParisTech
Paris, France
ttanon@enst.fr

Lucie-Aimée Kaffee*
University of Southampton
Southampton, UK
kaffee@soton.ac.uk

ABSTRACT

Stability in Wikidata's schema is essential for the reuse of its data. In this paper, we analyze the stability of the data based on the changes in labels of properties in six languages. We find that the schema is overall stable, making it a reliable resource for external usage.

ACM Reference Format:
Thomas Pellissier Tanon and Lucie-Aimée Kaffee. 2018. Property Label Stability in Wikidata: Evolution and Convergence of Schemas in Collaborative Knowledge Bases. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3184558.3191643

1 INTRODUCTION

Wikidata [10], originally created as a central storage for information of Wikipedia, developed into a widely used open knowledge base. It has become a source of linked data at large. Not only in but also outside the Wikimedia universe, its linked data is used. All data in Wikidata is assembled by the contributions of a community of users, evolving without a centralized process as in e.g. governmental open data [8] or extraction from a third party (e.g. YAGO [9]). All of the data in Wikidata is contributed by volunteers, including its schema. We define Wikidata's schema as the properties, the structure-giving part of the triples. Wikidata's triples contain usually two items and a property connecting them e.g. *<Ada Lovelace> <occupation> <computer scientist>*, where *<occupation>* is the property. One of the main factors for the reliability of a knowledge base (KB) is its schema. If the schema is stable, the data is easily reusable for third parties and the KB becomes attractive for reuse and further editing. Wikidata's schema is not constrained by e.g. the software, but can be adjusted as of the community's needs. In this paper, we want to investigate if the collaborative modelling of a schema is sustainable. We therefore quantify the way properties are changed. This gives us an insight not only if it works now, but also whether it is reasonable to assume that the schema is still usable in the same way in 5 years from now.

We focus on labels in our investigation. URIs in Wikidata are opaque as their function differs from the one for labels [5]. That means, each entity is addressed by a unique identifier, that does not suggest anything about the concept of the entity describes. For example, the item *Ada Lovelace* is identified with the ID Q7259, the property *capital* with the ID P36.

Labels are used to describe the content of an entity, identified with such an opaque URI. For example the property P31 has the label "instance of" in English and "nature de l'élément" in French. Labels are not only the access point for humans to the data [2], but usually also the way third party application will reuse the data. For example, Question Answering systems [1, 3] or ontology modeling [6] depend on natural language description of entities. Quantifying the stability of labels of the schema gives us an impression of how realistic the reusability of Wikidata's data is on a long perspective time-wise.

Wikidata is inherently multilingual. This means, editors add labels in over 400 languages. However, the coverage of labels differs dramatically between the different languages [4]. This can be attributed to the different sizes of communities. Nevertheless, in all languages stability of properties is an important factor. Therefore, we include five languages our analyzes: English (en), French (fr), German (de), Dutch (nl), Arabic (ar), and Yoruba (yo). As seen in Table 1, those languages are of varying sizes in terms of covered language information and native speaker. Given the close connection between Wikipedia and Wikidata and their editors [7], we include sizes of Wikipedias to gain an understanding of the community sizes.

	WD labels	WP pages	Speakers
en	14,867,057	5,559,376	365
fr	8,104,878	1,951,294	75
de	6,842,263	2,147,568	92
nl	8,710,608	1,922,135	21
ar	829,672	556,464	280
yo	39,389	31,608	28

Table 1: Language Statistics for Wikidata labels (WD), Wikipedia articles (WP), and native speakers as of 2007 in million

We conclude that beside possible concerns given the collaborative editing of a KB, its schema can be stable and reliable. We base our conclusion on our example of Wikidata. Over all six investigated languages of varying coverage, this assumption is supported.

2 METHODS

Wikidata is collaboratively edited. Each edit in Wikidata is recorded in the editing history of the respective entity. We extract the editing

[*] Both authors contributed equally to the paper

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191643

history of Wikidata for each property. For each revision we retrieve the labels and aliases of the property at this revision and the timestamp of the revision. Based on this data we can display timelines of changes in the property label[1]. We analyze the property label changes in four different metrics to get a comprehensive overview on the stability of the schema in Wikidata.

We define *Lifetime* as the time a property has the same label as the current one over the time this property existed.

The metric *Shared Labels* measures how many properties share a label with any other property, making a lookup based on name a challenging task.

Stability describes the probability, that a label of a property picked at a randomly chosen (with uniform distribution) point in time where such label exists will still be this property label or an alias now. In Wikidata, an alias is an alternative label to the main label of an entity, indicated by the property `skos:altLabel`. The intuition is that moving a label to the alias, the property is still discoverable by the same name. There might just happen slight changes to the actual label, while the concept the property refers to stays the same.

Furthermore, we measure *Quick Changes.* As editing of labels is open for any user[2], registered or anonymously, non-usable editing occurs. To count these edits, vandalism or good faith errors, we computed the number of *quick changes.*

We define quick changes as labels that stayed less than a week. Not included in this definition are changes at the beginning of the property life, when often the property semantic is still discussed. A change is only quick change if there has been a label before which stayed more than a week without being changed earlier in the property life.

	en	fr	de	nl	ar	yo
properties	3982	3910	2976	3710	3287	148
existence	1	0.97	0.94	0.96	0.31	0.20
lifetime	0.89	0.88	0.88	0.87	0.29	0.20
stability	0.96	0.94	0.96	0.92	0.97	0.99
changes	2.38	1.50	1.47	1.43	1.17	1.28
major changes	1.77	1.41	1.37	1.38	1.14	1.25
number QC	0.367	0.063	0.067	0.031	0.020	0.027
duration QC (in h)	1.79	0.94	1.12	0.89	1.48	0.0009

Table 2: Results of the analyzes focused each on a language. Properties (number of properties with a label), Existence (avg of ratio of the time span in which there is a label), Lifetime (avg presence of the last label), Stability (ratio of major values still present ponderated by duration), Changes (avg number of changes), Major changes (avg number of major changes), Number QC (avg number of quick changes), and Duration QC (avg duration of a quick change in hours). [3]

3 RESULTS AND DISCUSSION

We analyzed property labels in six different languages towards the stability of the schema.

Lifetime. On average the current English label has been the English label of the property for 87% of the property lifetime. This number is 88% for German and French and 86% for Dutch. There are also very few major changes in property labels. This number decreases in “medium" or “small" languages (on average 1.14 for Arabic) compared to “big” languages (1.77 for English). This suggests that once a property label is set, it is unlikely to change. This trend is even more verified in languages with smaller communities, probably due to the small number of editors in the language. The low lifetime of Arabic and Yoruba labels can be attributed to the relatively recent addition of most of their labels compared to the established languages. When we divide measures by the property life time they are disadvantaged. For example, on average a property has an Arabic label only during 31% of its lifetime, suggesting more recent additions than in languages like German or French (see the *existence* measure in Table 2).

Shared Labels. In English, no property shares a label with an other property (see http://tinyurl.com/yazxc5xq). However, there are 66 properties which English label is an alias of an other property (see http://tinyurl.com/yawtoudt).

Stability. In the case of English, the probability for stability is 96% and is higher than 90% for all the languages analyzed here. That means, looking up a property by a label used for this property at any point in time is highly likely to find the property ID in the current state of Wikidata. This works under the assumption, that it is highly unlikely that labels used before for property PX are now used for property PY. This assumption is supported by the results of the *shared labels* metric.

Quick Changes. Vandalism on labels is quite low. The number of *quick changes* (i.e. changes in labels that have been followed by an other change in less than a week) is of 0.36 per property on average in English. The smaller the language community, the lower this number gets. They are relatively quickly discovered and changed accordingly, as visible in Table 2. This leads us to the conclusion, that vandalism is a minor factor even in a collaborative and openly contributed knowledge base such as Wikidata, independent of the language.

4 CONCLUSION

We analyzed the schema of Wikidata towards its stability in terms of labels in six language. Overall, the results are very promising. The schema is stable, and therefore easily reusable. Labels of properties are rarely changed and not shared between different properties, making a lookup based on name an easy task. This makes Wikidata a multilingual source for a stable schema, that can be reused over various applications.

ACKNOWLEDGMENTS

This research was partially supported by the grant ANR-16-CE23-0007-01 (“DICOS”) and Horizon 2020 Marie Skłodowska-Curie grant agreement No. 642795 (WDAqua ITN).

[1] Like the one available at https://thomas.pellissier-tanon.fr/wikidata/labels-timeline.html

[2] There is only a set of 20 properties of the total 4,262 properties, whose editing is restricted to registered users: https://quarry.wmflabs.org/query/24202 (Query executed January 17th, 2018).

REFERENCES

[1] Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Maret. 2017. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information systems* (2017), 1–41. https://doi.org/10.1007/s10115-017-1100-y

[2] Basil Ell, Denny Vrandecic, and Elena Paslaru Bontas Simperl. 2011. Labels in the Web of Data. In *10th International Semantic Web Conference, ISWC, Part I*. 162–176. https://doi.org/10.1007/978-3-642-25073-6_11

[3] Konrad Höffner, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. 2017. Survey on challenges of Question Answering in the Semantic Web. *Semantic Web* 8, 6 (2017), 895–920. https://doi.org/10.3233/SW-160247

[4] Lucie-Aimée Kaffee, Alessandro Piscopo, Pavlos Vougiouklis, Elena Simperl, Leslie Carr, and Lydia Pintscher. 2017. A Glimpse into Babel: An Analysis of Multilinguality in Wikidata. In *13th International Symposium on Open Collaboration, OpenSym*. 14:1–14:5. https://doi.org/10.1145/3125433.3125465

[5] Elena Montiel-Ponsoda, Daniel Vila-Suero, Boris Villazón-Terrazas, Gordon Dunsire, Elena Escolano Rodriguez, and Asunción Gómez-Pérez. 2011. Style Guidelines for Naming and Labeling Ontologies in the Multilingual Web. In *Proceedings of the 2011 International Conference on Dublin Core and Metadata Applications, DC*. 105–115.

[6] Silvio Peroni, David M. Shotton, and Fabio Vitali. 2013. Tools for the Automatic Generation of Ontology Documentation: A Task-Based Evaluation. *Int. J. Semantic Web Inf. Syst.* 9, 1 (2013), 21–44. https://doi.org/10.4018/jswis.2013010102

[7] Alessandro Piscopo, Chris Phethean, and Elena Simperl. 2017. What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata. In *Social Informatics - 9th International Conference, SocInfo, Part I*. 305–322. https://doi.org/10.1007/978-3-319-67217-5_19

[8] Nigel Shadbolt, Kieron O'Hara, Tim Berners-Lee, Nicholas Gibbins, Hugh Glaser, Wendy Hall, and m. c. schraefel. 2012. Linked Open Government Data: Lessons from Data.gov.uk. *IEEE Intelligent Systems* 27, 3 (2012), 16–24. https://doi.org/10.1109/MIS.2012.23

[9] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *16th International Conference on World Wide Web, WWW*. 697–706. https://doi.org/10.1145/1242572.1242667

[10] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. https://doi.org/10.1145/2629489