Analyzing and Visualizing Translation  
Patterns of Wikidata Properties

John Samuel(B)[©](http://orcid.org/0000-0001-8721-7007)

CPE Lyon, Université de Lyon, Villeurbanne Cedex, France  
[john.samuel@cpe.fr](mailto:john.samuel@cpe.fr)

Abstract. From multi-domain multilingual Wikipedia websites to a single-domain multilingual Wikidata site, online collaboration has taken a major stride. However, achieving a multilingual experience is a rather challenging task for a highly evolving site like Wikidata built with the collaboration of contributors from around the world. It is important to let the contributors analyse and discover how properties are translated and also detect potential problems. This article focuses on developing a tool for understanding and visualizing the translation patterns of Wikidata.

Keywords: Online collaboration • Multilingualism Ontology development • Wikidata

1. Introduction

Wikidata [[11](#bookmark37)] is a multilingual, linked, open and structured knowledge base started by the Wikimedia Foundation in 2012. Since then, it has seen a tremen­dous growth and is widely used and studied by a number of researchers focusing on many aspects including collaboration and multilingual features [[4](#bookmark30),[5](#bookmark31),[7](#bookmark33),[10](#bookmark36)] of the project. Unlike its sister project, Wikipedia that has a dedicated sub-domai[n[[1]](#footnote-1)](#bookmark0)for every supported language (currently around 300 languages), Wikidata is a single domain websit[e[[2]](#footnote-2)](#bookmark1) with the capability to let users change the language set­tings. Such a change enables different users to view the same URL in different languages. That means a user can share these URLs to any other user of a different language and she can see the details in her own native language. For e.g.,<https://www.wikidata.org/wiki/Q9143>refers to the entity ‘programming language’ (en) or ‘langage de programmation’ (fr).

One interesting aspect is the way Wikidata identifies the items. It uses Q- number[s[[3]](#footnote-3)](#bookmark2) for entities called items and P-number[s[[4]](#footnote-4)](#bookmark3) to describe the properties.

A first-time user looking at an entity URL cannot say what it’s about unlike a Wikipedia URL (See footnote 4) or even a DBpedia UR[L[[5]](#footnote-5)](#bookmark4), another structured knowledge store.

Labels form a crucial part of Wikidata [[10](#bookmark36)]. They are short texts to identify a given Wikidata entity in a given language. Additionally, every item two other attributes: description and alias. Both of these are human-understandable texts available in different Wikidata-supported languages to specify a long description and any additional labels respectively.

Wikidata is a collaborative [[5](#bookmark31),[7](#bookmark33)] website, i.e., contributors can create, update or delete new items on the site. However, creation (and deletion) of Wikidata properties is a very long process. Wikidata, with its goal to represent data belong­ing to multiple domains uses both discussion and subsequent voting for the prop­erty creation/deletion. Properties form the key part of Wikidata since they are used to describe the various entities. Take for example, P856, one of the highly used properties is used to specify the official website of an entity. Its English label is ‘official website’. But given the collaborative nature, labels, descriptions and aliases can be changed by any contributor. Hence any changes must be properly monitored to detect any possible vandalism since it may affect the semantics of thousands (or even millions) of Wikidata entries.

Secondly, it is also important to analyze and visualize the way by which properties are translated. As discussed in [[7](#bookmark33),[8](#bookmark34)], though Wikidata provides a number of templates for multilingual discussion, properties are translated in other languages after their creation. Therefore, the role played by bilingual or even multilingual speakers in Wikidata cannot be undermined. Therefore, it is important to visualize the property translation process at a much more granular level not only to detect possible vandalism but also to get useful insights into the translation process.

Instead of doing one-time cumulative analysis of all the properties based on a data dump on a given period, it is important that the contributors can see, analyse and validate the results as well as suggest useful insights at a more gran­ular level. With these goals, the web application WDPro[p[[6]](#footnote-6)](#bookmark5)[[8](#bookmark34)] is built providing Wikidata users and contributors information concerning the properties, though primarily focusing on their multilingual translation. In this article, we will focus on how WDProp is further extended to obtain the details of property translation process.

Section 2 will briefly describe WDProp, its various objectives and its cur­rent capabilities. The implementation of the visualization process along with the results are described in Sect. [3](#bookmark19). Section [4](#bookmark22) presents various related works in multilingual collaborative ontology development and the role of visualization towards this goal. Section [5](#bookmark24) briefly describes the future course of actions and concludes the article.

1. WDProp

WDPro[p[[7]](#footnote-7)](#bookmark6) [[8](#bookmark34)] is an online application developed with the goal of understanding and improving multilingual and collaborative ontology development on Wiki­data. Based on the translation status in different languages, properties can be separated into two categories: translated and untranslated properties. For example, statistics showing already translated propertie[s[[8]](#footnote-8)](#bookmark7) and languages with no translated propertie[s[[9]](#footnote-9)](#bookmark8) are shown in Fig. [1.](#bookmark15) The key advantage of this tool is the ability to navigate the properties in different ways including based on the (missing) translation status, property datatypes, curated property classes, curated WikiProjects etc. Every link is bookmarkable and gives statistics on the fly.

**Count of translated labels Count of translated descriptions**



**Languages with no translated labels Languages with no translated descriptions**



Fig. 1. Statistics on WDProp with clickable and bookmarkable links

Total 177 languages



WDProp also lets the users to search properties by their labels, compare translation statistics among different languages, view already translated labels, descriptions, aliases and the different property discussion templates.

1. Implementation

In addition to showing translation statistics for every property, WDProp is fur­ther extended to show the translation pattern. To implement this feature, it goes through the edit history of a given property. Any edit on Wikidata results in a message is prepended with a phrase given in Table [1.](#bookmark21) There are also otherphrases like ‘wbeditentity-create’, ‘wbsetaliases-add-remove’, ‘wbsetlabeldescrip- tionaliases’ to detect when a property is created, an alias is added or removed or when all the three attributes are set in a single edit. The presence of these patterns are checked to understand the objective of each edit.

Table 1. Edit message for changes related to labels, descriptions and aliases of Wiki­data properties

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Add | Update | Remove |
| Label | wbsetlabel-add | wbsetlabel-set | wbsetlabel-  remove |
| Description | wbsetdescription-  add | wbsetdescription-  set | wbsetdescription-  remove |
| Alias | wbsetaliases-add | wbsetaliases-add-  remove | wbsetaliases-  remove |

We color code the actions and also show the associated time when a given action is made and use the colors blue to show addition, light blue to show any update and the color red to show removal. For each action, we also show the Wikidata language codes (e.g., en for English, fr for French etc.)

Development: WDProp is developed using web technologies like HTML, Javascript and CSS. This has an additional advantage that developers can down­load, setup and use it on their personal desktop or even integrate it with their internal servers. It makes use of Wikidata SPARQL endpoin[t[[10]](#footnote-10)](#bookmark9) to obtain the data on the fly.

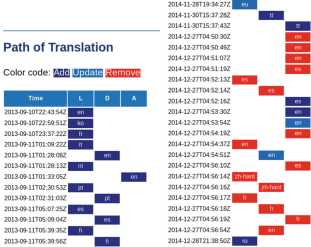
Results: Figure [2](#bookmark16) shows the results of the translation path of property P85[6[[11]](#footnote-11)](#bookmark10)) in chronological order, with four columns: time, label, description and alias. Some more translation path[s[[12]](#footnote-12)](#bookmark11) are given in the Fig. [3](#bookmark17). Contributors can also look at the colors light blue and red to detect any possible problems and if needed, go to Wikidata site to revert them. Some interesting questions can be: is English always the first language used for translation?, is it common to see multiple label or description changes in a given language? Which language immediately follows French? etc.

Limitations: Our approach has certain limits. If multiple changes are made in one single edit by bots, it is currently not possible to detect them. It may require additional information like the actual content changes. For properties that have a long edit history, the SPARQL query may timeout, thereby showing no results.

**P856**

* Property details
* Label missing translation in languages
* Description missing translation in languages
* Alias missing translation in languages
* Alias missing translation in languages
* Label already translated in languages

« Description already translated in languages

**Details**

**C:\Users\ettor\AppData\Local\Temp\FineReader12.00\media\image4.jpeg**

Link <https://www.wikidata.org/entitv/P856>

Translation Path path.html?propertv=P856

Fig. 2. Details of property P856, especially the translation path. Note the red color boxes. Some of these changes were later reverted. (Color figure online)

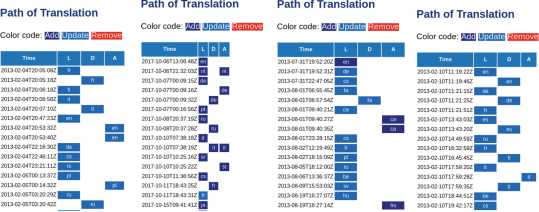
****

Fig. 3. Translation path of P18, P4290, P735, P106 (Color figure online)

1. Related Works

There is a growing interest in building collaborative approaches for knowledge management. Wikipedia, OntoWiki [[2](#bookmark28)], Wikidata [[11](#bookmark37)] are some major examples. All of these follow Wiki-style editing, i.e., users can create, edit, correct, delete, discuss or revert changes to articles. A number of research works are now avail­able especially for Wikipedia that can visualize the above actions. Visualization of deletion discussions is considered by Notabilia [[9](#bookmark35)]. [[1](#bookmark27),[3](#bookmark29)] focuses on analyzing how information flows across different language editions.

Even though Wikidata is a recent entry in this picture, the above questions still remain important. There are several research works that has focused on its multilingual [[4](#bookmark30),[10](#bookmark36)] and collaborative aspects [[5](#bookmark31)]. Recent work on property label stabilit[y[[13]](#footnote-13)](#bookmark12) [[10](#bookmark36)] shows the growing importance on the need of maintaining the stability in property labels and detect any undesired changes.

The goal of WDProp is to take these research works into consideration and provide on-the-fly statistics to Wikidata users and contributors so that they canfind by themselves any linguistic influence [[3](#bookmark29),[6](#bookmark32)] patterns, possible vandalism or any new insight in a transparent, reproducible and shareable manner.

1. Conclusion

Labels and descriptions of Wikidata properties are very important since they form a major role in describing numerous items. In this article, we looked at obtaining the translation path of every property and visualizing them with the online application called WDProp. Visualizing the translation process helps not only in detecting any possible vandalism but also in understanding how more tools and alert systems can be made for bilingual or multilingual speakers. By basing on web technologies and SPARQL endpoint, this work may be extended to other private or independent multilingual Wikibase instances. Our future course of actions include optimizing the performance of results especially for highly used properties with a long edit history.

References

1. Borra, E., et al.: A platform for visually exploring the development of Wikipedia articles. In: Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, 26-29 May 2015, pp. 711-712. AAAI Press (2015). [http://www.aaai.org/ocs/index.php/ICWSM/](http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10565) [ICWSM15/paper/view/10565](http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10565)
2. Frischmuth, P., Martin, M., Tramp, S., Riechert, T., Auer, S.: Ontowiki - an author­ing, publication and visualization interface for the data web. Semant. Web 6(3), 215-240 (2015).<https://doi.org/10.3233/SW-140145>
3. Gottschalk, S., Demidova, E.: MultiWiki: interlingual text passage alignment in Wikipedia. TWEB 11(1), 6:1-6:30 (2017).<https://doi.org/10.1145/3004296>
4. Kaffee, L., Piscopo, A., Vougiouklis, P., Simperl, E., Carr, L., Pintscher, L.: A glimpse into babel: an analysis of multilinguality in Wikidata. In: Proceedings of the 13th International Symposium on Open Collaboration, OpenSym 2017, Gal­way, Ireland, 23-25 August 2017, pp. 14:1-14:5. ACM (2017). [https://doi.org/10.](https://doi.org/10.1145/3125433.3125465) [1145/3125433.3125465](https://doi.org/10.1145/3125433.3125465)
5. Muller-Birn, C., Karran, B., Lehmann, J., Luczak-Rosch, M.: Peer-production sys­tem or collaborative ontology engineering effort: what is Wikidata? In: Proceedings of the 11th International Symposium on Open Collaboration, San Francisco, CA, USA, 19-21 August 2015, pp. 20:1-20:10. ACM (2015). [https://doi.org/10.1145/](https://doi.org/10.1145/2788993.2789836) [2788993.2789836](https://doi.org/10.1145/2788993.2789836)
6. Samoilenko, A., Karimi, F., Kunegis, J., Edler, D., Strohmaier, M.: Linguistic influence patterns within the global network of Wikipedia language editions. In: Proceedings of the ACM Web Science Conference, WebSci 2015, Oxford, United Kingdom, 28 June-1 July 2015, pp. 54:1-54:2. ACM (2015). [https://doi.org/10.](https://doi.org/10.1145/2786451.2786497) [1145/2786451.2786497](https://doi.org/10.1145/2786451.2786497)
7. Samuel, J.: Collaborative approach to developing a multilingual ontology: a case study of Wikidata. In: Garoufallou, E., Virkus, S., Siatri, R., Koutsomiha, D. (eds.) MTSR 2017. CCIS, vol. 755, pp. 167-172. Springer, Cham (2017). [https://doi.org/](https://doi.org/10.1007/978-3-319-70863-8_16) [10.1007/978- 3-319-70863-8\_16](https://doi.org/10.1007/978-3-319-70863-8_16)
8. Samuel, J.: Towards understanding and improving multilingual collaborative ontol­ogy development in Wikidata. In: Wikidata Workshop 2018 (2018). [http://](http://wikiworkshop.org/2018/papers/wikiworkshop2018_paper_12.pdf) [wikiworkshop.org/2018/papers/wikiworkshop2018\_paper\_12.pdf](http://wikiworkshop.org/2018/papers/wikiworkshop2018_paper_12.pdf)
9. Stefaner, M., Taraborelli, D., Ciampaglia, G.L.: Notabilia-visualizing dele­tion discussions on Wikipedia (2011). [http://well-formed-data.net/archives/570/](http://well-formed-data.net/archives/570/notabilia-visualizing-deletion-discussions-on-wikipedia) [notabilia-visualizing-deletion-discussions-on-wikipedia](http://well-formed-data.net/archives/570/notabilia-visualizing-deletion-discussions-on-wikipedia)
10. Tanon, T.P., Kaffee, L.: Property label stability in Wikidata: evolution and con­vergence of schemas in collaborative knowledge bases. In: Companion of the the Web Conference 2018 on the Web Conference 2018, WWW 2018, Lyon, France, 23-27 April 2018, pp. 1801-1803. ACM (2018). [https://doi.org/10.1145/3184558.](https://doi.org/10.1145/3184558.3191643) [3191643](https://doi.org/10.1145/3184558.3191643)
11. Vrandecic, D., Krotzsch, M.: Wikidata: a free collaborative knowledgebase. Com­mun. ACM 57(10), 78-85 (2014).<https://doi.org/10.1145/2629489>

1. e.g., https://[en,fr,ml].wikipedia.org. [↑](#footnote-ref-1)
2. [https://www.wikidata.org.](https://www.wikidata.org) [↑](#footnote-ref-2)
3. e.g., [https://www.wikidata.org/wiki/Q9143.](https://www.wikidata.org/wiki/Q9143) [↑](#footnote-ref-3)
4. [https://www.wikidata.org/wiki/Property:P279.](https://www.wikidata.org/wiki/Property:P279)

   © Springer Nature Switzerland AG 2018

   P. Bellot et al. (Eds.): CLEF 2018, LNCS 11018, pp. 128-134, 2018. <https://doi.org/10.100r/9r8-3-319-98932-r_12> [↑](#footnote-ref-4)
5. e.g., [http://dbpedia.org/ontology/ProgrammingLanguage.](http://dbpedia.org/ontology/ProgrammingLanguage) [↑](#footnote-ref-5)
6. [https://tools.wmflabs.org/wdprop/.](https://tools.wmflabs.org/wdprop/) [↑](#footnote-ref-6)
7. [https://doi.org/10.5281/zenodo.1174371.](https://doi.org/10.5281/zenodo.1174371) [↑](#footnote-ref-7)
8. [https://tools.wmflabs.org/wdprop/translated.html.](https://tools.wmflabs.org/wdprop/translated.html) [↑](#footnote-ref-8)
9. [https://tools.wmflabs.org/wdprop/untranslated.html.](https://tools.wmflabs.org/wdprop/untranslated.html) [↑](#footnote-ref-9)
10. [https://query.wikidata.org/.](https://query.wikidata.org/) [↑](#footnote-ref-10)
11. [https://tools.wmflabs.org/wdprop/path.html?property=P856.](https://tools.wmflabs.org/wdprop/path.html?property=P856) [↑](#footnote-ref-11)
12. [https://tools.wmflabs.org/wdprop/path.html?property=P4290.](https://tools.wmflabs.org/wdprop/path.html?property=P4290) [↑](#footnote-ref-12)
13. [https://thomas.pellissier-tanon.fr/wikidata/labels-timeline.html?entity=P279&](https://thomas.pellissier-tanon.fr/wikidata/labels-timeline.html?entity=P279&lang=en) [lang=en.](https://thomas.pellissier-tanon.fr/wikidata/labels-timeline.html?entity=P279&lang=en) [↑](#footnote-ref-13)