Simple Wikidata analysis for tracking and improving biographies in Catalan Wikipedia

Toni Hermoso Pulido
Centre for Genomic Regulation (CRG), BIST, Barcelona, Spain.
toni.hermoso@crg.eu

ABSTRACT

The advent of Wikidata represented a breakthrough as a collaborative and constantly advancing knowledgebase. As it was originally envisioned, it simplified the linkage and data reuse among different Wikimedia projects. Catalan Wikipedia is one example project where Wikidata has been heavily adopted by its community base: that is the case of integration with article infoboxes or in automatically generated lists. In the following article we highlight the possibilities of taking advantage of structured data from Wikidata for evaluating new biographical articles, so facilitating users to get engaged into diversity challenges or track potential vandalism and errors.

CCS CONCEPTS

• Information systems→Wikis

KEYWORDS

Gender bias, user engagement, curated review

1 Introduction

Biographical articles are a key and common component of Wikipedia projects. For sake of reference, in Catalan Wikipedia, identified thanks to Wikidata [1], they represent around 25% of the total number of articles. They are also important articles to have under scrutiny because they can be a target to abuse for self-promotion by biographed people and, in whole numbers, they can also reflect a prevalent gender bias [2][3]. For addressing this latter question, editor communities maintain wikiprojects (such as Viquidones in Catalan Wikipedia[1]) for monitoring and addressing these issues through editathons, contests or other means. Resources such as the Wikidata Human Gender Indicators (WHGI)[2] already provide a periodical global update of the situation for Catalan and many other Wikipedias. However, in order to engage users to reach a certain milestone of a number of articles or an article percentage of women biographies (for instance, within the time frame of a contest), an automatically generated list of new women articles was regarded more convenient. This list[3] is obtained thanks to a set of Python scripts and updated several times per day.

As a side product of this effort, this rendered useful for monitoring problems related to Wikidata gender property (P21) (e.g., when its value was missing), or for Wikidata entries that got 'human' instance (P31) value dropped, sometimes because of existing vandalism in Wikidata.

By taking advantage of the generated dataset of periodically updated biographical articles in Catalan Wikipedia, authority controls identifiers from their associated Wikidata entries were also retrieved. This was used to generate page reports and lists in order to identify which biographical articles were missing any authority control identifiers and to foster community discussions. Relevant points can be about the role of authority control identifiers in article notability or the suitability of their inclusion for improving reader experience (e.g., when linked in infoboxes or specific ‘Authority control’ templates).

1 https://ca.wikipedia.org/wiki/Viquiprojecte:Viquidones

2 https://whgi.wmflabs.org/

3 https://ca.wikipedia.org/wiki/Viquiprojecte:Viquidones/Progrés

2 Methods

Articles that can be considered as biographies are retrieved using Query Wikidata SPARQL endpoint by selecting those entries which have the instance value of human and have a Catalan Wikipedia sitelink. Article name, Wikidata ID and gender property (P21) are retrieved [4]. These same articles are inspected for their date and user of creation using MediaWiki API available at Catalan Wikipedia. Since this second step is time-consuming and it represents a non-mutating event, data of both processes are stored in a MariaDB relational database so only newly created articles are inspected at every run of the script, with the consequent saving of time.

For authority analysis, a recent Wikidata entity dump[4] is processed every week by retrieving and storing entries with authority control properties included in Catalan Wikipedia Authority Control template[5]. Articles that contain or link certain other pages, such as the forementioned 'Authority Control' template, can also be retrieved by using MediaWiki API.

Whenever possible, reports are shared with the editor community as lists and tables codified in wikitext. For sake of clarity, simple charts are designed and embedded also as wikitext thanks to Graph Extension, which is deployed in most Wikimedia projects.

The involved code, mostly written in Python, its associated documentation, and links to the generated reports can be found at the Git repository: https://github.com/toniher/wikidata-pylisting

3 Future improvements

Apart from biographical information related to gender and authority controls, other relevant and interrelated fields could be included, such as 'occupation', ‘date of birth’, 'employer' or 'position' (e.g., in governments or institutions), which could be used to support thematic challenges (i.e., women scientists for *International Day of Women and Girls in Science*)[6].

In order to facilitate users to take action and get engaged into certain Wikipedia or Wikidata entries, more detailed report subpages can be created or, alternately, provide associated links to resources such as PetScan[7], a tool that generates dynamic lists by cross-matching criteria from both Wikidata and Wikipedia.

4 Discussion

One limiting step for exploring additional properties, and so other aspects related to biographical content, in a close to live fashion is the execution timeout of the involved SPARQL queries. Instead of including several output properties, this could be worked out by performing several queries and combining them instead.

As it is approached by initiatives such as Wikipedia Cultural Diversity Observatory (WCDO) [5], combining outputs of several projects, such as different Wikipedia language versions, could also highlight existing gaps where the editor community can decide to act, for instance related to personalities linked to ongoing news.

Nonetheless, for massive datasets such as authority controls or bibliographical works, this is not realistic at the time of writing and dumps still need to be processed.

Wikidata is not an impassive resource to feed Wikipedia with factual and relational data but, as presented here as well, a dynamic framework that can support the progress and improve the quality of other projects (and also improve itself by supporting them). Favoring the cross-collaboration between Wikidata and other Wikimedia projects is both a technical and community challenges, so that, for instance, current data curator and article editor communities can benefit in the most effective ways. Efforts such as enabling Wikidata property edition from Wikipedia infoboxes would be aligned in this direction[8].

By embedding or porting some features of tools like PetScan or dashboards like WCDO in Wikidata or

4 https://dumps.wikimedia.org/wikidatawiki/entities/

5 https://ca.wikipedia.org/wiki/Plantilla:Autoritat

6 https://ca.wikipedia.org/wiki/Viquipèdia:Viquimaratató_Dia_Internacional_de_les_Dones_i_les_Nenes_en_la_Ciència_2021

7 https://petscan.wmflabs.org/

8 https://www.mediawiki.org/wiki/Wikidata_Bridge

Wikipedia interfaces, editors and curators could be engaged more promptly into addressing gaps or existing problems [6].

ACKNOWLEDGMENTS

My deepest thanks to developer and editor Wikidata and Wikipedia communities for making such thriving projects possible.

REFERENCES

[1] Vrandečić, Denny and Krötzsch, Markus. "Wikidata: A Free Collaborative Knowledgebase." *Communications of the ACM* 57, 10 (October 2014), 78–85. https://doi.org/10.1145/2629489.

[2] Konieczny, Piotr, and Maximilian Klein. “Gender Gap through Time and Space: A Journey through Wikipedia Biographies via the Wikidata Human Gender Indicator:” *New Media & Society*, June 18, 2018. https://doi.org/10.1177/1461444818779080.

[3] Wagner, Claudia, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. “Women through the Glass Ceiling: Gender Asymmetries in Wikipedia.” *EPJ Data Science* 5, no. 1 (December 2016): 1–24. https://doi.org/10.1140/epjds/s13688-016-0066-4.

[4] Malyshev, Stanislav, Markus Krötzsch, Larry González, Julius Gonsior, and Adrian Bielefeldt. “Getting the Most Out of Wikidata: Semantic Technology Usage in Wikipedia’s Knowledge Graph.” In *The Semantic Web – ISWC 2018*, edited by Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, 376–94. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-030-00668-6_23.

[5] Miquel-Ribé, Marc, and David Laniado. “Wikipedia Cultural Diversity Dataset: A Complete Cartography for 300 Language Editions.” *Proceedings of the International AAAI Conference on Web and Social Media* 13 (July 6, 2019): 620–29.

[6] Miquel-Ribé, Marc, and David Laniado. “The Wikipedia Diversity Observatory: A Project to Identify and Bridge Content Gaps in Wikipedia.” In *Proceedings of the 16th International Symposium on Open Collaboration*, 1–4. OpenSym 2020. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3412569.3412866.