Analysing the Evolution of Community-Driven (Sub-)Schemas within Wikidata

Sofia Baroncini[1], Margherita Martorana[2], Mario Scrocca[3], Zuzanna Śmiech[4] and Axel Polleres[5]

[1]*University of Bologna, Bologna, Italy*
[2]*Vrije Universiteit, Amsterdam, Netherlands*
[3]*Cefriel – Politecnico di Milano, Milan, Italy*
[4]*AGH University of Science and Technology, Cracow, Poland*
[5]*Vienna University of Economics and Business & Complexity Science Hub Vienna, Wien, Austria*

Abstract

Wikidata is a collaborative knowledge graph not structured according to predefined ontologies. Its schema evolves in a bottom-up approach defined by its users. In this paper, we propose a methodology to investigate how semantics develop in sub-schemas used by particular, domain-specific communities within the Wikidata knowledge graph: (i) we provide an approach to identify the domain sub-schema from a set of given classes and its related community, considered domain-specific; (ii) we propose an approach for analysing the such identified sub-schemas and communities, including their evolution over time. Finally, we suggest further possible analyses that would give better insights in (i) the communities themselves, (ii) the KG vocabulary accuracy, quality and its evolution over time according to domain areas, raising the potential of Wikidata improvement and its re-use by domain experts.

Keywords

Ontology Evolution, Empirical Semantics, Domain-specific Communities, Wikidata

1. Introduction

Since its initiation by the Wikimedia Foundation in 2012 [1], the Wikidata collaborative knowledge graph has now a collection of almost 100 million items. Wikidata users, editors and contributors, can describe and navigate through real-world concepts by querying the knowledge graph based on its entities, properties and attributes [2]. In contrast with the usual approach to knowledge graph (KG) engineering, Wikidata does not comply with a specific and predefined ontology for the creation and editing of items. Wikidata relies on its community of volunteers to support and expand its knowledge graph across different domains and expertise, as well as provide a connection to the Linked Data Web [3]. The fact that the community is responsible for the maintenance and expansion of the knowledge graph, makes a very interesting use case in the context of collaborative ontology engineering [4].

Wikidata'22: Wikidata workshop at ISWC 2022
sofia.baroncini4@unibo.it (S. Baroncini); m.martorana@vu.nl (M. Martorana); mario.scrocca@cefriel.com (M. Scrocca); zsmiech@student.agh.edu.pl (Z. Śmiech); axel.polleres@wu.ac.at (A. Polleres)
0000-0002-5636-8328 (S. Baroncini); 0000-0001-8004-0464 (M. Martorana); 0000-0002-8235-7331 (M. Scrocca); 0000-0002-1690-4093 (Z. Śmiech); 0000-0001-5670-1146 (A. Polleres)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

We would like to investigate the *"empirical semantics"* resulting from the creation, usage, modification and adaptation of a (sub-)schema. The collaborative nature and availability of the edit history of Wikidata provide a perfect proving ground for the study on empirical semantics and its evolution over time. While the quality of the Wikidata ontology and its evolution over time has already been investigated [4], a comparison considering the role of (sub-)communities is – to the best of our knowledge - still missing. To this end, we lay the foundations to investigate whether and how different domain-specific communities utilize distinct, different schemas, and how such sub-schemas evolve over time, which in turn can be compared with each other as well as with the evolution of the Wikidata ontology as a whole.

Along these lines, we define the following concrete research questions:

1. *How can domain-specific community schemas be defined and identified within the Wikidata KG?*
2. *Which patterns and metrics can be used to describe the empirical semantics adopted in a community-driven schema?*
3. *What is the evolution of a community-driven schema over time?*
4. *How do different communities within Wikidata compare with respect to the metrics and evolution of the schemas they use?*

In this position paper, we address these questions by proposing a structured approach based on the analysis of the literature and a preliminary assessment of the existing challenges. An overview of the approach is presented in Figure 1.

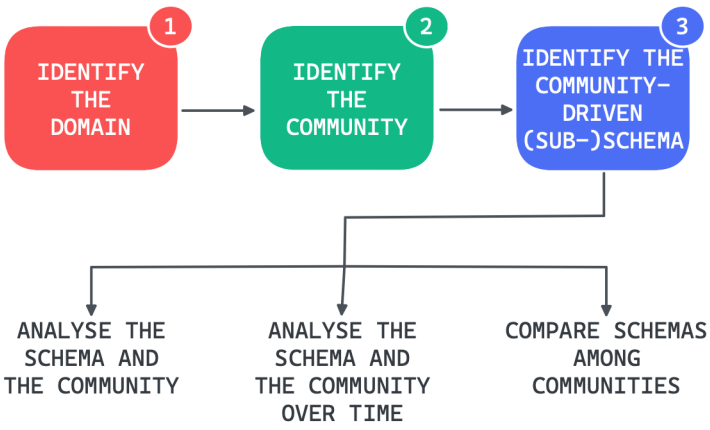

Figure 1: Overview of the proposed approach.

2. Preliminaries and Related Work

Wikidata is a collaborative knowledge graph based on a schema defined by the relations between the items in the graph (i.e. classes, properties and entities). To illustrate the taxonomy of such relationships, the Wikidata graph has properties such as *instance of* (P31) and *subclass*

of (P279), allowing us to understand the structure that builds the graph and distinguishes between which items are classes and which are entities. The Wikidata schema - made of classes and properties, their definitions and restrictions - is collaboratively built by its users with a bottom-up approach and does not follow a predefined ontology formalised through the Web Ontology Language (OWL) [5].

The current version of the Wikidata graph is publicly accessible through the Wikidata SPARQL query service[1]. Historical data about the edits made over time can be retrieved through the dumps made available by the Wikimedia foundation[2]. Pellissier Tanon and Suchanek in [6] analyse the problem of querying the edit history of Wikidata, proposing the implementation of a SPARQL endpoint to make data more accessible. A public endpoint, the *History Query Service*[3], is available and supports the analysis of the edit history of Wikidata until 2019. The endpoint allows executing queries considering the global state of the graph after each revision, the triples added and/or deleted at each revision, and the metadata about each revision (e.g., the user performing the edit).

The peculiar characteristics of the community-driven Wikidata knowledge graph are studied considering many different perspectives. Piscopo and Simperl in [7], reviewed the existing literature identifying different quality dimensions that could be considered in the analysis of Wikidata and its overall schema. The same authors in [4] have also investigated user roles in Wikidata, along with an analysis of the ontology engineering practices primarily followed by these roles. The paper also proposes a definition for the identification of the *Wikidata ontology* and a relevant set of metrics for analysing its quality and its evolution over time. We build upon both these contributions in the definition of our methodology. Additionally, we propose to consider other metrics to describe and analyze the schema of a community, such as the concept of *characteristic sets* introduced by Neumann and Moerkotte [8] originally to support cardinality estimation methods in RDF triplestore.

Kartik et al. in [9] analysed the quality of the statements available in Wikidata as a way to investigate current practices applied by the community. The analysis was based on removed statements, deprecated statements and constraint violations. As opposed to our work, the authors propose an analysis of the general quality of statements in Wikidata, while we focus on an approach to investigate *schemas* considering specifically a domain- or community-specific perspective. Indeed, as mentioned also in [10] as future work, the analysis of users and edits is influenced by the set of topics and/or categories considered, i.e., by their domain of expertise.

Different aspects are highlighted in the literature as relevant problems to be taken into consideration in the definition of community-driven schema within the Wikidata knowledge graph. In particular, various works have discussed the challenges of multi-linguality [11, 12] as well as the role of bots within the Wikidata community, especially with respect to the edits they make [4, 13]. Particularly, given that specific languages are a potential defining attribute of specific (sub-)communities, multi-linguality is, therefore, also potentially relevant to us. Likewise, the presence of bots could significantly skew the activity traces of real human sub-communities of editors.

[1] https://query.wikidata.org/

[2] https://dumps.wikimedia.org/

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

3. Proposed approach

This section describes the proposed approach for the identification and analysis of community-driven (sub-)schemas within the Wikidata knowledge graph. A more extensive and complete visual representation of the approach, in addition to what is already shown in Figure 1, is available on Zenodo[4].

3.1. The Wikidata Surface Schema

As a first step, we define the *Wikidata Surface Schema*, i.e., the schema within the Wikidata KG considered as input for the identification of community-driven (sub-)schemas. Following the approach reported in [4], the Wikidata schema can be defined as the set of properties and the set of items that are used as classes, i.e., those that are the object of *instance of* (`P31`), or subject/object of *subclass of* (`P279`). The Wikidata direct claim graph contains only the *truthy statements*[5], defined as statements that have the best non-deprecated rank for a given property. These statements can be identified through the Wikidata `wikibase:directClaim` special predicate, or the `wikibase:directClaimNormalized` one if also properties from external vocabularies should be considered. We define the *Wikidata Surface Schema* as the schema extracted by applying the definition from [4] to the surface claim graph.

3.2. (Sub-)Communities within Wikidata

The broad definition of the Wikidata community encompasses all users contributing to the development of the Wikidata knowledge graph. This definition can be narrowed to identify different groups of users:

- **based on the user role** - this definition distinguishes communities depending on users' role in the Wikidata community [4],
- **based on domains of knowledge** - the community is defined as the set of users editing entities belonging to a specific domain,
- **based on languages** - in this sense, the definition of community is a set of users contributing to Wikidata in a particular language [11, 12].

In this study, we focus on the definition of communities based on a specific domain of knowledge, since it is a relevant aspect still not addressed by the literature.

3.3. (Sub-)schema for a Wikidata (Sub-)Community

We propose a set of definitions and a methodology to identify a certain community contributing to a domain-specific part of the Wikidata knowledge graph, and the schema defined and adopted by users belonging to the community.

[4] https://doi.org/10.5281/zenodo.6961940
[5] https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format

1. **Identify the domain**. An *initial vocabulary* of Wikidata items associated with the considered domain should be identified, e.g., a set of relevant classes selected by domain experts. The **core-community schema** is a sub-schema of the Wikidata schema that leverages the *initial vocabulary* by selecting all classes and properties related to the items identified respectively through the *instance of* (P31), *subclass of* (P279) and *subproperty of* (P1647) properties.
2. **Identify the community**. The *community* is identified as the set of users, excluding bots [6], that created at least K1-number classes and/or properties of the core-community schema. A broader definition may be adopted, including also users that created at least K2-number instances of the classes defined in the core-community schema. The value of K1 and K2 can be parameterized, the higher the parameter value, the greater the probability that the group of filtered users has knowledge in the specific domain.
3. **Identify the community-driven (sub-)schema**. The *community-driven schema* can be obtained from the *Wikidata Surface Schema* by extracting: (i) all the classes created by users in the identified community, (ii) all the properties associated with at least one instance of the obtained classes.

The implementation of the approach requires the analysis of the Wikidata edit history accessible programmatically through the described dumps and/or the dedicated endpoint described in Section 2. The parametric nature of the approach is made necessary by the different roles that users may assume in editing the Wikidata knowledge graph [4], indeed, it is necessary to filter out users making edits across different domains. The proposed approach assumes that the usage of classes and instances in the Wikidata knowledge graph is proper, erroneous and/or inconsistent modelling decisions adopted by users may require more advanced methods for identifying a community and its (sub-)schema.

3.4. Analysis

Both the identified (sub-)schema and the community should be analysed. The metrics proposed by Piscopo and Simperl in [4] offer a structured approach to evaluate both the ontology quality and the role of users considering the evolution over time of the schema. The same set of metrics can be assessed on Wikidata snapshots over time to detect patterns in its evolution. In our approach, we propose the adoption of the same set of metrics but reducing the scope of the schema and the community, from the whole Wikidata ontology and community to the ones identified through the described approach. Furthermore, we describe a set of additional analyses that can be performed to extract relevant insights in the context of community-driven (sub-)schemas within Wikidata.

1. **Analyse the schema and the community**. The features discussed in [4] about ontology quality (e.g., number of classes, number of instances, chains of sub-classes relations) and users (e.g., number of edits on different types of items, the proportion between the number

[6] Bots can be identified by a flag and by the word "bot" in the user's name, as described by https://www.wikidata.org/wiki/Wikidata:Bots. As not all the bots are identified, more accurate methods for their detection can be based on other parameters, such as the users' behaviour [14]

of edits and number of items edited) provide relevant metrics to analyse both the schema and the community. Additional factors that we propose to investigate are related to the analysis of the schema concerning its actual usage within the Wikidata knowledge graph. A simpler analysis can consider the top-k *used* classes and properties by assessing which elements of the extracted schema are more widely adopted by the community. A more detailed analysis, can consider the concept of *Characteristic sets* [8] to detect patterns of usage. *Characteristic sets* identifies the set of subjects characterized using the same set of predicates within RDF datasets and allows for the identification of semantically similar subjects. We argue that this approach can also be used to characterise and compare patterns in the analysis of empirical semantics adopted by a community. In this way, it is possible to analyse if there is consistency in the usage of the same schema structure to describe a given class, or if there are significant variations. If a recurrent characteristic set emerges, it can be then supposed that this empirical application of Wikidata schema is preferred by the domain-specific community of interest.

2. **Analyse the schema and the community over time**. The same set of metrics defined in the first step should be computed on different snapshots of the Wikidata KG to analyse the evolution of the schema over time. Moreover, we also would like to investigate how, considering a fixed community of users (i.e. the group of users as defined in point 2 of Section 3.3, considering the overall revisions made in the snapshots selected), the identified (sub-)schema evolves over time. A first analysis should consider the quantitative evolution of the schema by comparing numerical metrics on classes and properties. The additional analysis proposed are: (i) top-k new classes and properties introduced in the schema and their usage[7], (ii) diachronic analysis of statements, e.g., considering what were the first created properties and/or sub-classes for a specific class.
3. **Analyse the schemas among communities**. A comparison of the results obtained considering different communities and their (sub-)schemas within the Wikidata knowledge graph is needed to complement the approach. This type of analysis opens to a comparison between structural ways used by different communities to express a certain type of knowledge.

4. Discussion and Conclusions

In this position paper, we sketched an approach to identify a community-driven schema starting from a given (set of) domain (core classes and properties), as well as vice versa to identify a community from a schema in a knowledge graph and perform analyses over such community-driven schemata. We have proposed to re-use existing metrics defined by [4] on the one hand and add new additional metrics that shall allow us to compare communities and their schema usage and to detect the evolution of a community-driven schema over time. As a result, we contribute to the current state-of-the-art by providing a methodology for the analysis of the Wikidata schema considering the role of communities of users.

[7] Can be extracted considering the maximum Q and P identifiers from the previous snapshot and selecting items and properties with higher identifiers in the current snapshot

Future works include the evaluation of the approach on a set of case studies, as well as a more extensive definition of community-driven schemas. Such schemas, for example, could be defined by using not only the *instance of* (P31) and *subclass of* (P279) properties, but also the additional meta-modelling properties introduced in the Wikidata schema and analysed by Haller et al. in [5]. Moreover, we suggest a further thorough analysis of the domain community by subdividing it into different roles/sub-communities such as (1) the group of editors and (2) the group of contributors. This approach applying the study conducted by [4] to a domain-specific community would potentially give interesting insights into the use of sub-vocabularies, evolution and users dynamics of sub-communities, with the final goal to compare the different communities coexisting in a knowledge graph. Furthermore, this approach would allow us to have a better understanding of the domains represented in Wikidata. Firstly, it would be possible to understand the accuracy of the domain vocabulary usage by comparing the initial vocabulary considered relevant by a domain expert with the actual schema implemented. Secondly, it allows an evaluation of the Wikidata quality according to domain areas, identifying, on one hand, those domain schemas that need some improvements or, on the other hand, to rate the domains that have the most excellent quality to enhance the reuse of their data by domain experts.

References

[1] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.

[2] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods, Semantic web 8 (2017) 489–508.

[3] A. Haller, A. Polleres, D. Dobriy, N. Ferranti, S. J. Rodríguez Méndez, An analysis of links in wikidata, in: The Semantic Web, Springer International Publishing, Cham, 2022, pp. 21–38.

[4] A. Piscopo, E. Simperl, Who Models the World?: Collaborative Ontology Creation and User Roles in Wikidata, Proceedings of the ACM on Human-Computer Interaction 2 (2018) 1–18. doi:10.1145/3274410.

[5] A. Haller, A. Polleres, D. Dobriy, N. Ferranti, S. J. Rodriguez Mendez, An Analysis of Links in Wikidata, in: The Semantic Web, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2022, pp. 21–38. doi:10.1007/978-3-031-06981-9_2.

[6] T. Pellissier Tanon, F. Suchanek, Querying the Edit History of Wikidata, in: The Semantic Web: ESWC 2019 Satellite Events, volume 11762, Springer International Publishing, Cham, 2019, pp. 161–166. doi:10.1007/978-3-030-32327-1_32, series Title: Lecture Notes in Computer Science.

[7] A. Piscopo, E. Simperl, What we talk about when we talk about wikidata quality: a literature survey, in: Proceedings of the 15th International Symposium on Open Collaboration, OpenSym '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–11. doi:10.1145/3306446.3340822.

[8] T. Neumann, G. Moerkotte, Characteristic sets: Accurate cardinality estimation for

RDF queries with multiple joins, in: 2011 IEEE 27th International Conference on Data Engineering, 2011, pp. 984–994. doi:10.1109/ICDE.2011.5767868, iSSN: 2375-026X.

[9] K. Shenoy, et al., A study of the quality of Wikidata, Journal of Web Semantics 72 (2022) 100679. doi:10.1016/j.websem.2021.100679.

[10] C. Sarasua, et al., The Evolution of Power and Standard Wikidata s: Comparing Editing Behavior over Time to Predict Lifespan and Volume of Edits, Computer Supported Cooperative Work (CSCW) 28 (2019) 843–882. doi:10.1007/s10606-018-9344-y.

[11] G. Amaral, et al., Assessing the Quality of Sources in Wikidata Across Languages: A Hybrid Approach, Journal of Data and Information Quality 13 (2021) 23:1–23:35. doi:10.1145/3484828.

[12] L.-A. Kaffee, Multilinguality in knowledge graphs, phd, University of Southampton, 2021. URL: https://eprints.soton.ac.uk/456783/.

[13] L. N. Zheng, C. M. Albano, N. M. Vora, F. Mai, J. V. Nickerson, The Roles Bots Play in Wikipedia, Proceedings of the ACM on Human-Computer Interaction 3 (2019) 215:1–215:20. doi:10.1145/3359317.

[14] A. Hall, L. Terveen, A. Halfaker, Bot detection in wikidata using behavioral and other informal cues, Proc. ACM Hum. Comput. Interact. 2 (2018) 1–18.