Recoin: Relative Completeness in Wikidata

Vevake Balaraman
University of Trento
Trento, Italy
balaraman@fbk.eu

Simon Razniewski
Max Planck Institute for Informatics
Saarbrücken, Germany
srazniew@mpi-inf.mpg.de

Werner Nutt
Free University of Bozen-Bolzano
Bozen-Bolzano, Italy
Werner.Nutt@unibz.it

ABSTRACT

The collaborative knowledge base Wikidata is the central storage of Wikimedia projects, containing over 45 million data items. It acts as the hub for interlinking Wikipedia pages about a specific item in different languages, automates features such as infoboxes in Wikipedia, and is increasingly used for other applications such as data enrichment and question answering. Tracking the quality of Wikidata is an important issue for this project. In this paper we focus particularly on the *completeness* aspect. Several automated techniques have been adopted by Wikis to track and manage completeness, yet these techniques are generally subjective and do not provide a clear quality estimate at the level of entities. In this paper, we present an approach towards measuring *Relative Completeness* in Wikidata by comparison with data present for similar entities. This relative completeness approach is easily scalable with the introduction of new classes in the knowledge base, and has been implemented for all available entities in Wikidata. The results provide an intuition on the completeness of an entity comparing it with other similar entities. Here, we present our implementation approach along with a discussion on strategies and open challenges.

CCS CONCEPTS

• **Information systems** → **Wikis**; **Recommender systems**; *Incomplete data*; *Similarity measures*;

KEYWORDS

Wikidata, Wikipedia, Data Completeness, Data Quality, Knowledge Bases

ACM Reference Format:
Vevake Balaraman, Simon Razniewski, and Werner Nutt. 2018. Recoin: Relative Completeness in Wikidata. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3184558.3191641

1 INTRODUCTION

Wikimedia Foundation projects such as Wikipedia, Wiktionary, and Wikibooks are important sources of information to people across the globe. Wikidata, a younger member of the family of Wikimedia projects, is a collaborative database that acts as a central repository of structured data for other Wikimedia projects, and provides structured data for a variety of other applications, ranging from museum metadata to political transparency and scientific publications.

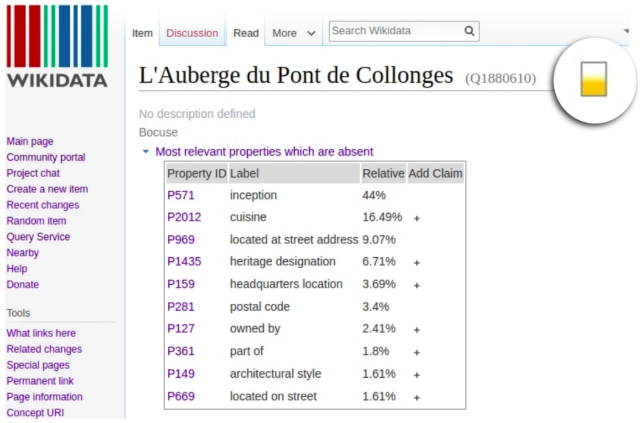

Figure 1: Recoin for *L'Auberge du Pont de Collonges*. The color indicator shows the completeness of the entity, while the list shows relevant absent properties.

Wikidata enables for instance the automatic generation of infoboxes across language version of Wikipedia, and even allows one to import facts directly into article bodies. This way, if for instance a player changes his team, a single change in Wikidata could be sufficient to update information in all (currently) 288 language versions. Several other features of Wikipedia such as list generation on a particular type or related link suggestions, make use of the data from Wikidata.

Given the size of the Wikidata project and its impact, it is important to supplement editors with capabilities to assess the quality, and in particular the completeness of Wikidata. *Completeness is a data quality measure that refers to the degree to which all required information is present in a particular dataset* [14]. Traditional databases are modeled for well-defined domains, with a specific schema defining the contents that can be added to the database. In these cases, completeness of an entity can be easily measured by checking whether all attributes foreseen by the schema have a value that is not null. Wikidata, on the contrary, is an open-domain knowledge base without a fixed schema. While there are a few core properties such as *date of birth* or *place of birth* that are virtually mandatory for the class human, it is possible to express over 4200 properties in Wikidata for various classes, many only applying in specific circumstances (think of *place of detention* or *monastic order*), thus making a strict definition of data that should be present impossible.

Previous work on Wikidata quality has focused on the property level, i.e., assessing whether data for a specific property of an entity

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

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191641

Figure 2: Wikidata page of Antoine Griezmann.

is complete [10]. Given authoritative sources, it is relatively easy to check whether Wikidata contains all clubs that Antoine Griezmann played for, or whether all arrondissements of Lyon are mentioned.

With the Recoin tool (Relative completeness indicator), we look at entity completeness, a different granularity of investigation. Recoin aims to help in answering the question *How complete is data about an entity in Wikidata as a whole?*, based on a relative completeness approach that compares the data present for a given entity with the data present for other, similar entities.

Recoin is available as a gadget to any logged-in Wikimedia user.[1] Figure 1 shows the output of Recoin for a sample entity, *L'Auberge du Pont de Collonges*, the only restaurant in the vicinity of Lyon with 3 Michelin stars.

2 BACKGROUND

Wikidata is a free crowd-sourced knowledge base with more than 19k active users [2] that contains information known to human knowledge, and acts as a central data storage for the structured data to other Wikimedia projects. Any entity known to human knowledge can be represented in Wikidata as an *item* and each item can be described by the five following elements. (In the paper the terms *item* and *entity* are used interchangeably.)

(1) **Labels**: A label is the most common name that an item is known by. Two items can carry the same label.
(2) **Description**: A description is a short phrase to describe an entity.
(3) **Aliases**: Aliases are alternative names, other than the label, that an entity may be known by.
(4) **Sitelinks**: Sitelinks are links to other Wikimedia projects that contain information about that particular item.
(5) **Statements**: Statements represent the information or data about that particular item in Wikidata. Each statement about an item consists of a *property* and a *value*. A single property may contain multiple values.

The labels, descriptions, aliases and sitelinks are multilingual (i.e., the details are inputted for a specific language and may differ between languages) whereas the statements are language independent as they contain the language-independent facts about the item.

[1] https://www.wikidata.org/wiki/Wikidata:Recoin

Figure 2 shows the information present on the Wikidata page of Antoine. Griezmann[2] It captures the relevant information about Antoine Griezmann including both his personal and professional information. Yet the page gives no indication of how complete data about Griezmann is.

Knowing the data quality of an entity in Wikidata gives us insights into the quality of the database and helps the editors/users focus on certain entities. As Wikidata is created by a community of users, providing them information on the entity helps them curate the data better. This helps in improving the quality of the knowledge base.

3 RELATED WORK

Considering the importance of high coverage of data in knowledge bases, research has focused on addressing the quality and enhancement of knowledge bases.

Completeness of knowledge bases: For conceptual and pragmatic reasons, knowledge bases contain only a subset of the information that holds in reality, hence, are incomplete [12]. Paulheim has studied various knowledge graph refinement approaches that aim to identify wrong information and add missing knowledge to the graph [8]. All of them take some data source as a reference for good quality. He distinguishes between gold-standard strategies, which refer to a product of external knowledge, for instance annotations provided by humans, or other knowledge bases, and silver-standard strategies, which take the knowledge base itself as point of reference. He states that "scalability issues are only rarely addressed by current research works" [8]. Aprosio et al. addressed the data coverage issue in DBpedia by proposing a distant supervision approach to extend the coverage of properties in the DBpedia knowledge base using Wikipedia [3]. Färber et al. provided criteria for analyzing the data quality of different knowledge bases [4]. They defined metrics that evaluate the quality of a knowledge base by comparing it with a gold-standard data source. They also conducted an evaluation of selected classes using a manually created gold-standard that defines a small set of core properties that every human should have. Yet such an approach cannot be expected to scale, thus, completeness assessment of knowledge bases needs to utilize other techniques, such as the identification of trends in data [6].

Quality on Wikimedia projects: Quality has long been observed as important. In Wikipedia, various status indicators and templates exist to mark articles e.g. as *excellent*, *unsourced*, *stub-level*. Automated tools are also employed for this purpose, for instance the Objective Revision Evaluation Service (ORES) [1], a web-service that can automatically predict the quality of articles and edits. In 2017, a competition was held about classifying Wikidata edits as vandalism or not, where the best systems could achieve 87% precision [7]. Yet quality manifests itself in many other ways than vandalism. The closest attempt at tracking completeness of Wikidata is the COOL-WD tool [10], a web portal that allows one to record the completeness of values for individual properties of entities, for instance universities in Lyon, or members of the French national soccer team. COOL-WD also aggregates these assertions

[2] https://www.wikidata.org/wiki/Q455462

into a single score. However, this describes only how many properties contain complete information, not whether the entity contains all relevant information. Wulczyn et al. [13] address a similar issue as our paper, but on Wikipedia. They proposed a recommendation system to reduce the article coverage gap across languages, that suggests important missing articles to editors based on his/her interest. Their findings show that such curated suggestions improve the chances of being created by a factor of 3.2 and increase the editor engagement by a factor of 2 [13].

In a recent paper we developed a machine learning approach to predict, given an item, which of two properties people would find more interesting to know about [11]. Such pairwise preferences can be extended in various ways to a ranking of all properties for the item and used to suggest the relevant ones needed to make the item complete. To learn such a model, the approach uses the Wikipedia page corresponding to that Wikidata item. This is a limitation as not all items in Wikidata have a corresponding page in Wikipedia and also most of the existing pages contain very little information to extract.

While there have been studies to improve the quality of Wikimedia projects, unlike in Wikipedia there is so far no tool available that gives real-time quality information on the level of entities.

4 APPROACH

Relative Completeness. The term *completeness* ideally defines if the knowledge base captures all known information about an item in the form of *statements*. Intuitively, a boolean parameter should be sufficient to indicate if the item is complete or not. For a defined domain unlike Wikidata this may be feasible. Also, unlike for properties, for which it is largely possible to indicate whether they contain all values relevant to the item, representing entity completeness by a boolean parameter would convey little information as to the quality of the entity. Labeling an entity that has a single property as incomplete and an entity with over 100 properties as complete does not provide much specific information on their quality. Also identifying the complete entities is a rather hard and infeasible approach because of the fact that certain properties in Wikidata capture information that is bound to change over time, e.g., the property *medical condition* for items of type *human*.

To quantify completeness under these circumstances, we propose the use of a *relative* notion of completeness: capturing recall in comparison with other, similar subjects. For example, to assess the completeness of data about Trump, one should look at the KB contents for other US presidents, and for assessing the completeness of a city, one should look at the data for similar cities.

Formally, relative completeness relies on two components:

(1) a similarity function between subject pairs $sim(S_1, S_2)$ that can be used to compute a (weighted) set of similar subjects $\mathbf{S}$ for a subject S;
(2) a scoring function $score(S, \mathbf{S})$ that computes a score or rank for the completeness of S with regard to a set of comparison subjects $\mathbf{S}$.

We detail next our instantiation of these concepts.

Design Choices. For computing entity similarity, in Recoin, we rely on a simple boolean similarity function that considers two entities as similar if they share at least one type. For entities of type human, we refine this by treating the values of their *occupation* property like types.

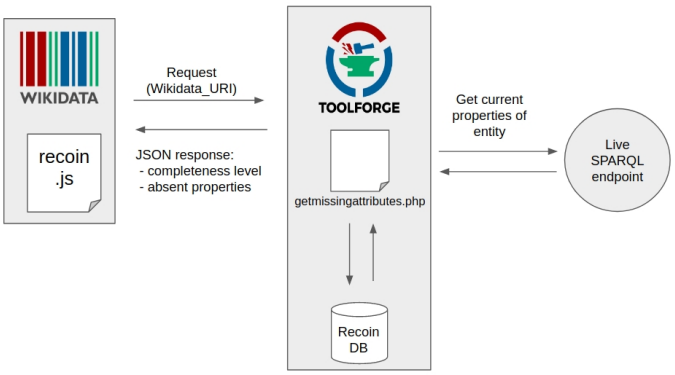

Figure 3: Architecture of Recoin.

We then compute the 5 most common properties in $\mathbf{S}$ that subject S is missing, and show the average frequency of these properties in the comparison set $\mathbf{S}$, subtracted from 1, as $score(S, \mathbf{S})$. One could easily use more than 5 properties, yet, the influence of additional properties is diminishing, so 5 properties was found to sufficiently describe the shape of the distribution of the frequency of missing properties.

In the end, the score is discretized into 5 buckets as follows:

- Level 5 (Very good informativeness) - 100%-95% score
- Level 4 (Good informativeness) - 95%-90% score
- Level 3 (Medium informative) - 90%-75% score
- Level 2 (Basic information) - 75%-50% score
- Level 1 (Very basic information) - 50%-0% score.

As an example, consider Larry Sanger, the co-founder of Wikipedia, who in Wikidata is listed with the professions *philosopher* and *blogger*. Consequently, he would be compared with all other philosophers and bloggers. The most frequent properties among these subjects that Larry Sanger is lacking are *member of*, *award received*, *work location*, *religion* and *described by source*, which occur in 12.09% to 6.35% of these types of subjects, thus leading to an average frequency of missing properties of 8.38%, and a completeness score of 91.62% (Level 4). For comparison, Trump's score is 98.07%, while Tim Berners-Lee's score is 95.63% (each Level 5).

Both similarity and scoring function leave space for refinement. For computing subject similarity, a range of techniques such as similarity of textual descriptions or relatedness measures in knowledge graphs could be used (for a recent survey, see [9]). For scoring subjects, it is desirable to use more informed techniques than simple counts, as frequent properties are not necessarily also important. In our ongoing work, we aim to devise more accurate and subject-specific ranking of properties [11].

5 IMPLEMENTATION

In this section, we describe the implementation of Recoin along with explanations of the adopted strategies. The implementation is split into three modules.

(1) Relevant properties—which computes the most relevant missing properties for an entity based on entities similar to it.

(2) Completeness—which uses the data from the module above to compute the completeness for each entity in real-time.
(3) Integration with Wikidata—which integrates the results from the former modules into Wikidata pages.

In Fig. 3 we illustrate the architecture of Recoin. At a high-level, during run-time (when a user loads a Wikidata page), the *Recoin* plugin queries the script hosted in Toolforge[3] to retrieve the completeness score and the relevant missing properties for the item.

5.1 Relevant properties

As stated in Section 4, the average frequency of missing properties determines the completeness measure of an entity. In the following, we refer to a value of the property *instance of* as a class and to an item having the value as an instance of the class. For each property and each class, we calculate the frequency of the property, that is, we count how many instances have the property. In this way, each class gives rise to a specific ranking of the properties according to their frequency in that class.

For the case of entities that are instances of multiple classes, we compute the *weighted frequency* of each property as the sum of the frequencies over those classes divided by the sum of the cardinalities of those classes. This computation is slightly different from the description in Sec. 4 insofar as this way, entities that share multiple occupations with the entity of interest are counted several times, yet a live computation of property frequencies is not realistic. By weighting property frequencies of types, it is possible to have these precomputed, and the difference is generally minor.

For instance, Emmanuel Macron, the current President of France, has occupations *banker*, *politician* and *statesman*. There are 6,704 bankers, 338,464 politicians, and 792 statesman in Wikidata (as of Dec-2017). Among bankers, 648 (9.7%) have the property work location (P937), while among politicians 105,485 (31.2%) have, and among statesman 30 (3.8%) have. Thus, the final computed frequency this property is the weighted average, equaling $(648 + 105{,}485 + 30)/(6{,}704 + 338{,}464 + 792) = 31\%$.

5.2 Completeness

Our early analysis revealed that computing the completeness based on all the missing properties is not effective. Since an entity is bound to have some less frequent properties missing (maybe because the properties are totally irrelevant) compared to other similar entities, the average frequency would always be very low, giving us no information. Conversely, if we reason about completeness based on the top property alone, an entity could be classified as minimally complete just because one major property would be missing. So, trying out different numbers of properties as the basis of our classification, and inspecting the results for various classes, we found that an optimal balance is achieved by basing the completeness classification on the top-5 missing properties for an entity.

In the special case of properties related to death such as *date of death* or *place of death*, we manually created rules to filter them out whenever a person does not have at least one of them.

[3]https://wikitech.wikimedia.org/wiki/Help:Toolforge

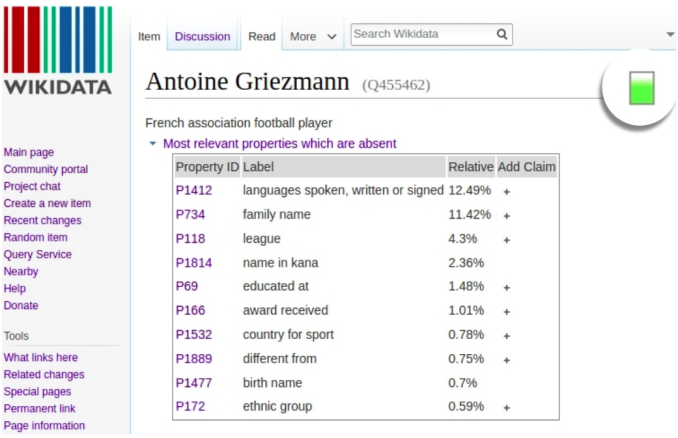

Figure 4: Recoin on the Wikidata page of Antoine Griezmann.

5.3 Integration with Wikidata

The implemented tool is deployed in the hosting environment of toolforge, provided by Wikimedia. An endpoint receives the request to calculat the relative completeness for a given item identifier and returns a *json* output. The data output by this endpoint is then captured by Recoin to integrate in the webpage of Wikidata.

The completeness measure is represented visually in Wikidata by an indicator capturing the extent of information. Fig. 4 shows the output of Recoin on the Wikidata page of Antoine Griezmann. With Recoin, we can address the question established in Section 2 (i.e., *How complete is the data for Antoine Griezmann?*). Although Antoine Griezmann's page contains a large set of information in Wikidata, we notice that there are certain relevant properties that are not listed for him. Recoin highlights this fact, suggesting the most relevant missing properties and summarizing the completeness information in a visual indicator.

Considering community feedback, a special version of the tool was made available that suggests only the properties of datatype *external-id*.

5.4 API

We also provide a standalone API that can be used for programmatic evaluations of Wikidata quality. The API can be accessed at

https://tools.wmflabs.org/recoin/getmissingattributes.php?lang=en&subject=Q1880610

by substituting the Q-code and the language with the desired choices. Unlike the Recoin tool, the API does not require a login to Wikidata, and thus enables the use of Recoin for anyone.

5.5 Analysis using Recoin

Using the above-mentioned API, the completeness score was calculated for all entities having the value *association football player (Q937857)* for the property *occupation* in Wikidata. This allowed us to compute the completeness score for all football players present in Wikidata and better understand the completeness of the whole class of football players. A total of 215,342 entities (as of 24 February,

	Entity count	Average completeness score	Avg. # of properties / entity	Avg. # of statements / entity
Level 1	31 682	30.69	7.10	7.38
Level 2	65 387	67.48	10.86	13.28
Level 3	106 200	85.86	13.60	18.72
Level 4	11 422	91.90	22.54	30.73
Level 5	657	96.15	30.60	40.81

Table 1: Completeness statistics for football players present in Wikidata

	Entity count	Average completeness score	Avg. # of properties / entity	Avg. # of statements / entity
Level 1	334	46.60	1.12	1.15
Level 2	979	64.33	4.65	4.85
Level 3	589	82.04	7.30	7.69
Level 4	81	91.70	8.22	8.71
Level 5	3	96.12	12.00	13.00

Table 2: Completeness statistics for submarines present in Wikidata

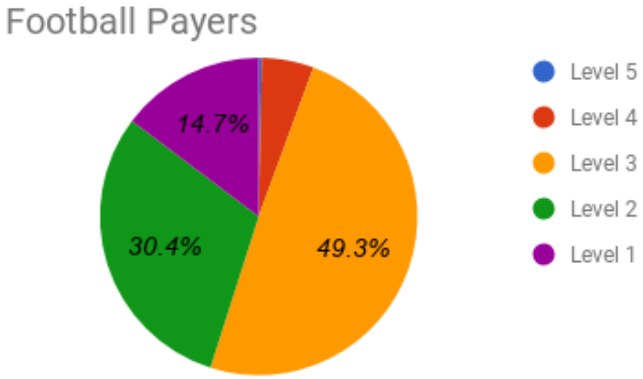

Figure 5: Completeness estimate for all football players in Wikidata.

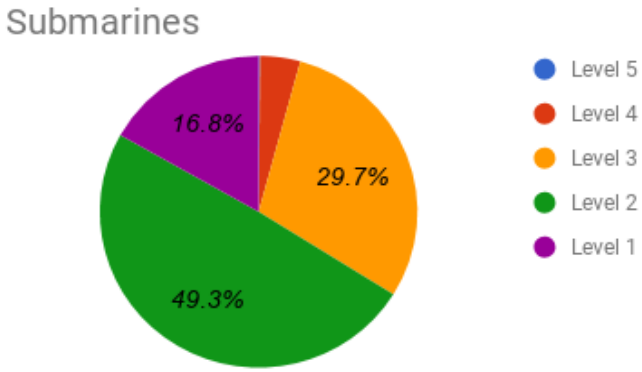

Figure 6: Completeness estimate for all submarine entities in Wikidata.

2018) were found that are instances of class *human* and have this occupation. Table 1 shows the statistics obtained from Recoin and also additional details obtained from Wikidata for this type of entities. Fig. 5 shows the percentage of entities in each completeness level over all football players

The entity count for each completeness level and the corresponding average completeness score are retrieved from Recoin while the average number of properties/entity and average number of statements/entity are obtained from the Wikidata API. A property may have more than one value for an entity, therefore leading to a difference in the number of properties and the number of statements. We notice from the results that only 0.5% of the football players in Wikidata have a highly complete profile while over 40% have profiles that contain only basic information (level 1 and 2).

Similarly, entities that are instances of the class *submarines (Q2811)* were analyzed using Recoin. Table 2 shows the corresponding results while Fig. 6 shows the percentage of entities at each completeness level. A total of 1986 entities (as of 24 February, 2018) were found in Wikdiata for this class. Here, though the number of entities is very low, Recoin can still identify their completeness based on the other entities in that class. We notice that over 60% of the profiles contain only basic information (level 1 and 2).

6 DISCUSSION

In this section we discuss the main challenges towards measuring and quantifying relative completeness.

Potential Non-monotonicity. The first challenge is the right choice of the similarity metric. Our current choice of considering all classes/professions leads to some unexpected results. For instance, Garry Kasparov, a well-known chess player, is evaluated to be less complete than certain other rather unknown chess players, because of the presence of other occupations in his profile. The evaluation expects him to have the relevant properties of other occupation while the properties related to the occupation he is famous for are present. A potential solution is the use of *ranks*: The Wikidata data model allows one to express that certain property values are preferred over others, for instance, in Kasparov's case, the profession *chess player* is preferred over the others. Yet there is no equivalent solution for the case of classes, so here a custom class ranking would be required [5].

Correlation between Homogeneity and Relative Completeness. A second issue concerning relative completeness is its relationship to property homogeneity. If in a class, all entities have nearly the same properties, all entities would have a good relative completeness (missing few of what many others have). In contrast, if a class is very heterogeneous, most entities would miss properties that some

others have, thus achieving on average a worse relative completeness. This is for instance the case for the class *human*. We tackled this, thereby refining the class by profession, yet a correlation between homogeneity and relative completeness remains inherent to our approach.

Cultural Bias. Another challenge is the threat of cultural dominance via relative completeness. Since English content is the most widely used/edited, there is a strong influence of cultural bias from these entities. As the entities with most properties to a class are likely to get higher completeness levels, entities from other countries are almost always bound to get a lower completeness score as their page is expected to have the same level of information as for English entities. This effect may reduce the acceptance of the approach.

Unsuitable Properties. To ensure user experience, we manually filtered a few unsuited property suggestions relating to death, like *date of death* or *place of death*, that would otherwise be often top-ranked. Yet, similar properties also exist for other classes, for companies for instance via the property *dissolved, abolished or demolished*. As a manual treatment is not scalable, an automated approach towards identifying and filtering such properties would be highly desirable.

Ontological Reasoning. Our approach of measuring completeness is purely based on statistical distributions. Yet Wikidata comes with a rich ontology that could help in refining the assessment. Ideally, assertions such as that every human should have a birth date could be exploited directly, or constraints such as that female cyclists do not have a *male cyclist database ID* could be used to filter missing properties.

Though these challenges are somewhat subjective, we believe that they are important aspects to address to provide a better experience for all users and maintain the high quality of Wikidata.

7 CONCLUSION

In this work, we presented an approach to evaluate the *completeness* of entities in Wikidata. We based our approach on data present for other, similar entities, and provide a completeness indicator deployed in Wikidata as Recoin. We believe that objective criteria for assessing quality are important for resource allocation and project management. Recoin provides a useful first step in this direction. Extensions that use a more fine-grained identification of similar entities, and that can assess property relevance beyond pure frequencies are desirable.

Acknowledgment

This work has been partially supported by the project "TaDaQua", funded by the Free University of Bozen-Bolzano.

REFERENCES

[1] ORES objective revision evaluation service. https://www.mediawiki.org/wiki/ORES. Accessed: 21-01-2018.
[2] Wikidata statistics. https://www.wikidata.org/wiki/Special:Statistics. Accessed: 21-01-2018.
[3] A. P. Aprosio, C. Giuliano, and A. Lavelli. Extending the coverage of DBpedia properties using distant supervision over Wikipedia. In *NLP & DBpedia*, pages 20–31, 2013.
[4] M. Färber, F. Bartscherer, C. Menne, and A. Rettinger. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. 9:1–53, 03 2017.
[5] D. Fernández-álvarez, J. Emilio, L. Gayo, and D. Gayo-avello. ClassRank : a method to measure class relevance in knowledge graphs applied to Wikidata. *Semantic Web*, 0:1–14, 2017.
[6] L. Galárraga, S. Razniewski, A. Amarilli, and F. M. Suchanek. Predicting completeness in knowledge bases. *WSDM*, 2017.
[7] S. Heindorf, M. Potthast, H. Bast, B. Buchhold, and E. Haussmann. WSDM cup 2017: Vandalism detection and triple scoring. In *WSDM*, pages 827–828, 2017.
[8] H. Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3):489–508, 2017.
[9] M. Ponza, P. Ferragina, and S. Chakrabarti. A two-stage framework for computing entity relatedness in Wikipedia. *CIKM*, pages 1867–1876, 2017.
[10] R. E. Prasojo, F. Darari, S. Razniewski, and W. Nutt. Managing and consuming completeness information for Wikidata using COOL-WD. *COLD workshop at ISWC*, 2016.
[11] S. Razniewski, V. Balaraman, and W. Nutt. Doctoral advisor or medical condition: Towards entity-specific rankings of knowledge base properties. In *ADMA*, 2017.
[12] S. Razniewski, F. M. Suchanek, and W. Nutt. But what do we actually know. *AKBC*, 2016.
[13] E. Wulczyn, R. West, L. Zia, and J. Leskovec. Growing Wikipedia across languages via recommendation. In *WWW*, pages 975–985, 2016.
[14] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for Linked Data: A survey. *Semantic Web Journal*, 2015.