Analysis of Consistency between Wikidata and Wikipedia Categories

Leila Feddoul[1,2], Frank Löffler[1,3] and Sirko Schindler[2]

[1]*Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena, Jena, Germany*

[2]*Institute of Data Science, German Aerospace Center DLR, Jena, Germany*

[3]*Competence Center for Digital Research, Michael Stifel Center, Jena, Germany*

Abstract

Wikipedia categories play a significant role in organizing articles by topic. They form a hierarchy, which groups related articles into larger collections. Wikidata provides a corresponding item for each category and allows to define membership of other items to the specific category by a SPARQL query or by specifying classes and properties. This provides us with multiple, redundant sources of category membership which may deviate quite substantially. In this paper, we investigate inconsistencies between Wikipedia and Wikidata category members and analyze possible reasons. We propose a candidate category generation and evaluation workflow that traverses the category hierarchy of Wikipedia in all available languages and compares the results with information obtained from Wikidata. This workflow can be executed either online using the publicly available endpoints or offline based on the provided dumps. Furthermore, we formulate concrete suggestions to harmonize category membership definitions between Wikipedia and Wikidata.

Keywords

Wikidata, Wikipedia, Wikipedia Category

1. Introduction

Wikipedia has grown to be a valuable source of semi-structured information, written and maintained by a large community and provided for everyone to use. As of 2022, it contains over 6.5 million articles in its English section[1], but is also available in 329 other languages[2]. It has a community of about 280,000 active editors and more than 100 million registered users. The basic building blocks of Wikipedia are *articles* that are interlinked among each other.

Wikidata [1] is, like Wikipedia, free and open, but instead of a collection of articles that are intended primarily to be read by humans, it is a knowledge base that is intended to be read and edited by humans and machines. Wikidata is a source of open data that other projects, including Wikipedia, can use to enrich their services. The basic building block of Wikidata is an *item*, which represents any kind of real-world topic, concept, or entity that is uniquely identified.

[1]https://en.wikipedia.org/wiki/Wikipedia:Statistics

[2]https://en.wikipedia.org/wiki/Wikipedia

Figure 1: Wikidata category: *Association football video games* (*Q13199045*).

Wikipedia established ways to structure its building blocks (articles): Categories, i.e. sets of articles or subcategories, are among them. They play an important role since they support finding sets of articles having the same characteristics without knowing individual articles beforehand. The Wikipedia category structure has also been exploited for other tasks like entity retrieval [2] or document classification [3]. Wikipedia's categories[3] group articles with similar topics. E.g., *Category: Former countries* groups a set of articles related to the concept of a former country. This does not only include articles about the respective former countries like *Inca Empire*, but also subcategory pages, e.g., *Category: Former countries in fiction*. Categories can contain subcategories, but the resulting data structure is not a tree, but a more general graph because articles and subcategories can be members in multiple parent categories and, while discouraged, even loops can exist[4].

There is a quite close connection between categories in Wikipedia and Wikidata. In general, for each page (article, category, or otherwise) in Wikipedia, there exists a corresponding Wikidata item that is unique for all languages [4]. Wikidata category items are instances of *Wikimedia category* (*Q4167836*). This type of Wikidata item has specific properties, some of which describe a criterion for membership of a given Wikidata item to the considered category: (i) *Category contains* (*P4224*) is described as *category contains elements that are instances of this item* and consists of a value together with qualifiers[5] if available. The property value refers to the type of items contained and is referred to as *target* in this paper. (ii) *Wikidata SPARQL query equivalent* (*P3921*) is described as *SPARQL code that returns a set of entities that correspond with this category or list*. Figure 1 shows the Wikidata item for the category *Association football video games* (*Q13199045*), with *video game* (*Q7889*) as a target, *genre* (*P136*) together with *sport* (*P641*) values as qualifiers, and a corresponding SPARQL query. In addition, a list of corresponding Wikipedia articles is linked. This exemplifies the multiple sources for category membership used

---

[3]https://en.wikipedia.org/wiki/Wikipedia:Categorization

[4]https://en.wikipedia.org/wiki/Wikipedia:FAQ/Categorization

[5]Qualifiers provide additional information about a specific statement that may not be represented in a single triple statement. For more details, kindly refer to https://www.wikidata.org/wiki/Help:Qualifiers.

in Wikipedia and Wikidata. If more than one source is given, the resulting category members could in theory differ. As we will show later, this is often the case in practice and poses a possible consistency problem.

To the best of our knowledge, no previous work has analyzed the (in)consistencies between Wikipedia category members and items retrieved using the SPARQL queries or targets attached to the respective Wikidata categories and proposed a solution on how to reduce inconsistencies. In this paper, we analyze them by comparing their content, elaborating on possible reasons for and against making all sources consistent, and suggesting some potential future research directions. The key contributions of this paper are: (i) A workflow for the automatic generation of candidate categories together with their SPARQL Wikidata and Wikipedia members (mapped to Wikidata) derived from traversing the Wikipedia category hierarchy in all available languages. (ii) An analysis of inconsistencies within Wikidata categories and between Wikipedia and Wikidata. (iii) An automatic investigation of possible reasons for inconsistency.

The source code for the dataset generation is publicly available [5, 6] under an MIT License and works on both online Wikipedia/Wikidata public endpoints and offline SQL/JSON dumps. All generated data [7], cache files [8] containing data retrieved from dumps as well as experiment results [9] are published on Zenodo. This makes the whole analysis fully reproducible (on dumps of historic versions of the sources) as well as reusable (assuming the underlying items and articles in Wikipedia and Wikidata do not change too much).

2. Related Work

Various works have investigated different aspects of leveraging Wikidata and Wikipedia content. H. Turki et al. [10] focus on explaining how Wikipedia and Wikidata can be processed using existing techniques for data parsing and querying. Furthermore, they raise awareness about the usefulness of the integration of Wikipedia and Wikidata categories for different semantic applications and provide some ideas to enhance the quality of both sources (e.g., removing non-transitive relations from the Wikipedia category graph through the analysis of Wikidata statements). Driven by the observation that a large number of Wikidata entities lack corresponding Wikipedia articles in some languages (orphans), N. Ostapuk et al. [4] propose a pipeline to map Wikidata orphan entities to Wikipedia articles' sections. Their goal is to enrich orphans with additional facts and properties that are derived from their corresponding textual description in Wikipedia. As a result they provide a dataset consisting of a collection of Wikidata entities together with their potential links to related Wikipedia pages in different languages. I. Johnson [11] analyzed how Wikidata content is referenced within the English Wikipedia and proposed a taxonomy that categorizes Wikidata transclusions based on the reader impact. In the context of Wikidata enrichment from external sources, A. Boschin et al. [12] proposed a method based on knowledge graph embeddings to predict new facts (e.g., triple completion) using the hyperlinks between Wikipedia articles. P. Curotto et al. [13] proposed a Wikipedia-based approach for automatic suggestion of authoritative references for Wikidata statements. The goal is to support editors while referencing Wikidata claims. To evaluate the accuracy of the automatic recommendations, they also provide a gold standard dataset of sample claims and their corresponding external references in the English Wikipedia.
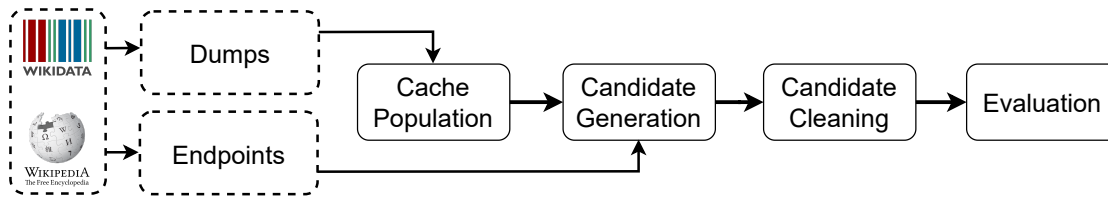
Figure 2: Workflow for candidate generation and evaluation.

3. Approach

As the base of our analysis, we need to retrieve categories and their members from Wikipedia and Wikidata respectively. The corresponding pipeline is outlined in Figure 2. It can be executed either using the public APIs offered by Wikipedia[6] and Wikidata[7] or the regularly provided SQL/JSON dumps from both sites[8]. We employ a cache to hold all information needed. When using the public APIs, this cache is filled successively after each request. In case the data dumps are used, a preprocessing step extracts all relevant data from the provided files and populates the cache accordingly. In both cases, the cache allows to prevent redundant requests and speeds up the processing considerably.

We start the candidate generation by retrieving all items from Wikidata that correspond to a Wikipedia category, i.e. instances of *Wikimedia category* (*Q4167836*). For each of those, we further store: (i) The Wikidata identifier, (ii) the target given by the value of *category contains* (*P4224*) including a list of subclasses, (iii) further qualifiers attached to the target, (iv) the corresponding SPARQL query via *Wikidata SPARQL query equivalent* (*P3921*), if existing, including the results after running the query, (v) and the corresponding Wikipedia category pages in all languages. For each Wikipedia member, we also retrieve the direct types and all their properties and their corresponding values[9]. Some categories are removed from further consideration as the corresponding SPARQL queries do not adhere to the structure of a single target and associated qualifiers. Among the deviations, there are: the use of multiple targets, lack of an *instance of* (*P31*) relation, and queries involving property paths.

Next, we turn to Wikipedia. For each of the previously identified categories, we fetch the members and traverse the hierarchy of subcategories if necessary. Wikipedia versions in different languages are maintained independently from each other[10]. Membership in categories is maintained manually and, hence, also differs across languages. Hence, we have to traverse categories for each language independently. For each member, we store the corresponding Wikidata identifier. While traversing the hierarchy of subcategories using a Breadth-First Search, we apply type checks using the target of the initial category: If fewer than 50% of member

---

articles (excluding any subcategories) are instances of the target or any of its subclasses[11], the traversal in this branch will end. After this step, we have acquired not only member items from Wikidata through the provided SPARQL query but also the manually curated list of members from Wikipedia.

Finally, we apply a cleaning step that removes some categories from consideration. Categories will be omitted if one of the following criteria applies: (i) The category has more than one target. (ii) The category has no corresponding Wikipedia members. This may be due to, e.g., the type check already failing for the members of the initial Wikipedia category. (iii) The corresponding SPARQL query yielded no results. (iv) Multiple SPARQL queries were supplied.

4. Evaluation

The pipeline was executed using the Wikidata JSON dump of 2022-05-02 and the Wikipedia SQL dumps of 2022-05-01. At that time, Wikidata contained roughly 4.99 million categories[12]. Out of these, only $2,280$ have a corresponding SPARQL query (*P3921*), $749,385$ have a target (*category contains P4224*), and only 516 have both of them. Using the restrictions outlined in Section 3, this leaves us with 206 categories used for evaluation.

Our goal is to perform an analysis of the consistency between Wikipedia and Wikidata with respect to the categories' content and an automatic investigation of possible reasons. For this purpose, we compare the two member sets by calculating the *precision* and *recall* of items corresponding to Wikipedia articles $\{WIKI\}$ with respect to SPARQL query results $\{SPARQL\}$:

$$Precision = \frac{|\{SPARQL\} \cap \{WIKI\}|}{|\{WIKI\}|} \qquad (1)$$

$$Recall = \frac{|\{SPARQL\} \cap \{WIKI\}|}{|\{SPARQL\}|} \qquad (2)$$

Results reveal an *average precision* of $\sim$ 0.65 and an *average recall* of ~ 0.75. Figure 3a and Figure 3b show the distribution of both metrics for the 206 candidate categories. Based on Figure 3a, we observe that for 136 out of 206 categories, at least 80% of items retrieved using SPARQL also appear as Wikipedia members, 88 categories share more than 90% of the items, and 19 of the categories have a low *recall* of below or equal to 30%. Figure 3b shows a rather uniform distribution of the *precision*, except for categories having more than 90% precision which applies to 61 out of 206. Overall, a rather high *recall* can be observed. Items retrieved by SPARQL but not found via Wikipedia (causing lower *recall*) can be attributed to one of the two reasons: Either the entity was not added to the category by any Wikipedia editor or the traversal has been stopped too early and the respective subcategory was not visited.

Since the overall *precision* provides a rather mixed picture, we conducted a more detailed investigation into possible reasons. *Precision* gives insights about items that were found in Wikipedia but not by SPARQL. We define the following possible reasons for an item not being found via SPARQL queries:

---

[11]SPARQL: ?entity wdt:P31/wdt:P279* ?target

[12]Retrieved via the following query: *SELECT (COUNT(DISTINCT ?cat) AS ?count) WHERE { ?cat wdt:P31 wd:Q4167836 .}*
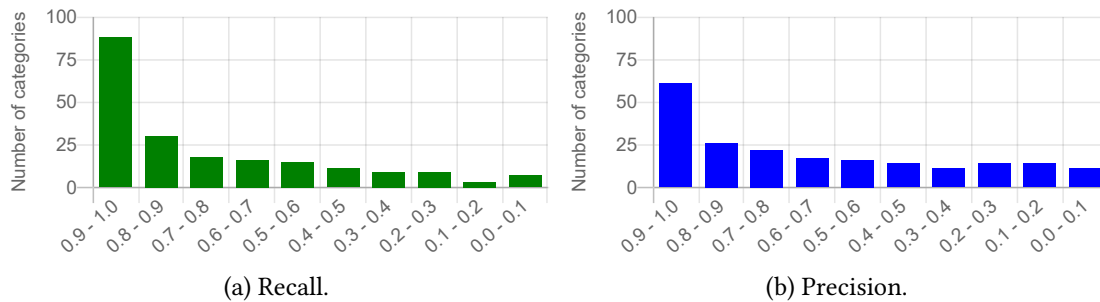
(a) Recall.
(b) Precision.

Figure 3: Precision and recall per category for Wikipedia member articles with respect to the Wikidata items retrieved using corresponding SPARQL queries.

- **Missing property (missingProp)**: The Wikidata item does not own a given property.

- **Different property value (diffPropValue)**: The Wikidata item contains the property but it has a different value.

- **Usage of another property (otherPropUsage)**: The Wikidata item points to a specific value but uses a different property.

We then analyzed the distribution of issues over all items that were not found by SPARQL (items appearing only as Wikipedia members) for all categories. Based on Figure 4a, we notice that the most spread issue type is the *diffPropValue* with $\sim 87\%$ of not found items, followed by *missingProp* with $\sim 15\%$, and *otherPropUsage* with $\sim 0.41\%$. Note that the fraction of items for each issue does not sum up to 100% because the same item may be counted multiple times if the SPARQL query contains multiple properties. E.g., it is possible that one property is missing while the other one has a different value.

We also analyzed the consistency of a SPARQL query with the target and qualifier information available in the *category contains* (*P4224*) property of the Wikidata category. Here, the same issue classes as before apply as well. Figure 4b shows the distribution of issues over all categories in this case. We notice that the most wide-spread issue type is *missingProp* with 52 categories, followed by *otherPropUsage* with 19 categories, and *diffPropValue* with 4 categories – the remaining categories (131) show no issues. An example category with the *otherPropUsage* issue is *Category:Uruguayan beach volleyball players (Q22136982)* with *category contains* (*P4224*) consisting of the following qualifiers: *<occupation (P106): beach volleyball player (Q17361156), country for sport (P1532): Uruguay (Q77)>*, and with a SPARQL query: *?item wdt:P31 wd:Q5; wdt:P27 wd:Q77; wdt:P106 wd:Q17361156*. In this case, the property used within the target's qualifiers, *country for sport (P1532)*, has been replaced by a similar albeit not equal property, *country of citizenship (P27)*.

For all categories, we further considered the correlation between the fraction of items with a specific issue type and the number of Wikipedia items not found by SPARQL. Based on Figure 5a and Figure 5b, we notice that categories with no issues mostly have a small number of Wikipedia items not found by SPARQL, observing some outliers with a very low fraction of items with the issue but with more than $1,000$ not found entities. Furthermore, categories with all items having the issue are observed for categories with a rather small number of items not found.
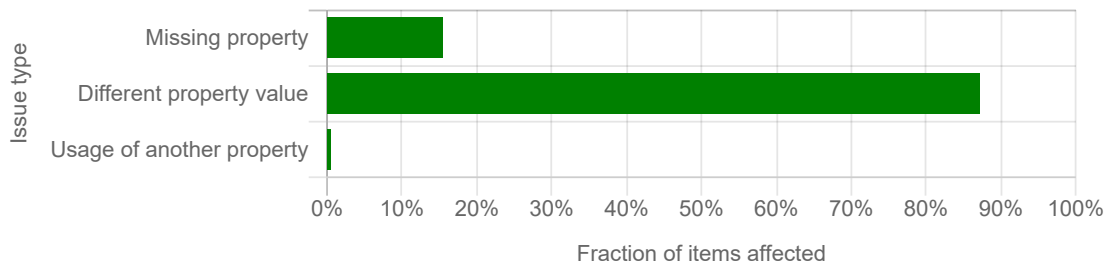
(a) Comparing Wikipedia categories and Wikidata SPARQL queries.
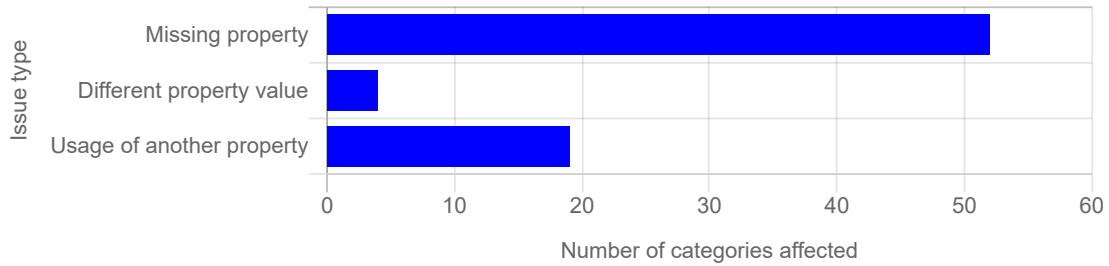
(b) Comparing target and associated qualifiers with SPARQL queries in Wikidata.

Figure 4: Relevancy of issue classes for different sources of category membership.

The remaining categories do not follow any specific trend since for categories with a similar number of items not found, we observe a great variety for the fraction of items affected by the issues. Figure 5c shows that most of the categories affected by the respective issue have a number of Wikipedia items not found by SPARQL ranging from 100 to 10, 000 with a rather low percentage of items affected since it is the less wide-spread issue. In general, we do not see a clear trend for a correlation between the size of the Wikipedia items not found and the items affected by the issue.

5. Conclusion and Future Work

We analyzed the consistency between Wikidata and Wikipedia categories and investigated possible reasons. We also compared the available information within a Wikidata category (SPARQL query with the target and qualifiers). For this, we proposed a workflow for automatic generation of candidate categories. It traverses Wikipedia's category hierarchy in all available languages and retrieves corresponding members as long as certain conditions hold. Results reveal differences of various degrees between all sources and show three possible causes.

The underlying reason for the discovered inconsistencies are rooted in the manual curation of three separate sources answering in essence the same question: Which items/articles should be members of a given category? To increase consistency, we suggest to treat Wikidata's *category contains* (*P4224*) as the main source of truth. From an automation standpoint, this provides the most structured information. From this, SPARQL queries could be automatically generated. Finally using these queries, the members of Wikipedia's categories can be derived. As Wikidata
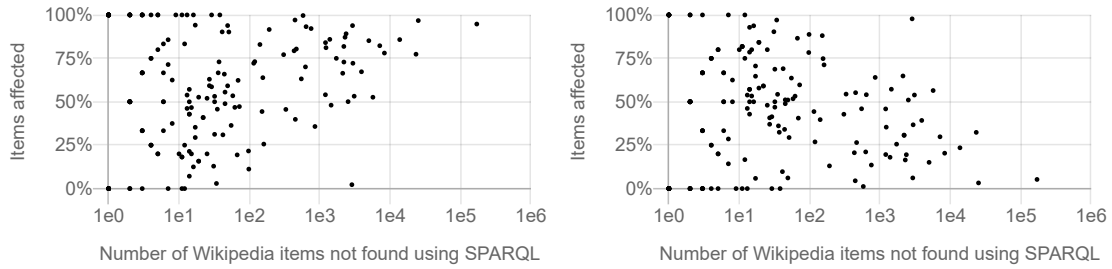
(a) Different property value.

(b) Missing property.
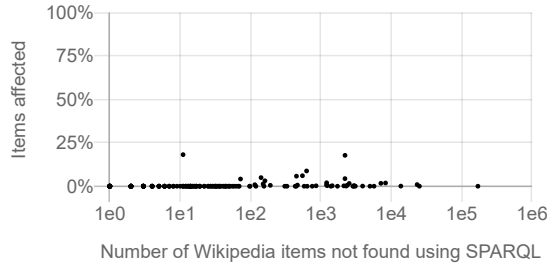
(c) Usage of another property.

Figure 5: Affected items over Wikipedia items not found using SPARQL. One point per category with precision below 100% (173 categories). X-axis using a log scale.

albeit growing remains incomplete, we may further use the current category membership in Wikipedia together with Wikidata's *category contains* (*P4224*) to complete the information of items of Wikidata.

Two approaches are possible to improve this situation: First, new changes to any source are verified against the information contained in the other two. Editors may get a warning if they seemingly violate these constraints. The cause might not be their current action but a mismatch with another source. So, editors may still overrule the warning and commit their change. Second, we can create an interactive interface to review the changes proposed previously. As we can not be certain which information is wrong or incomplete, here human editors may verify the assumptions of an automated system and only verified changes will be propagated to Wikipedia and Wikidata respectively. Both approaches will over time increase the consistency and quality of both Wikipedia and Wikidata and as a consequence improve their usefulness in other applications.

We base our work on the assumption that the definitions of categories, i.e. their semantics, are consistent across Wikidata and all languages within Wikipedia. This might fail, though. Categories like *Category:American singers* (*Q7063228*) can be seen in at least two ways, both of which are legitimate interpretations: based on the *country of citizenship* (*P27*) as Wikidata currently does or based on the *place of birth* (*P19*). Although we do not have evidence of similar divergences in existence, they would require a community process to converge on a common interpretation before applying our suggestions.

Acknowledgments

References

[1] D. Vrandečić, M. Krötzsch, Wikidata: A Free Collaborative Knowledgebase, Commun. ACM 57 (2014) 78–85. doi:10.1145/2629489.

[2] R. Kaptein, J. Kamps, Exploiting the category structure of Wikipedia for entity ranking, Artificial Intelligence 194 (2013) 111–129. doi:10.1016/j.artint.2012.06.003, artificial Intelligence, Wikipedia and Semi-Structured Resources.

[3] J. F. Medeiros, B. P. Nunes, S. W. M. Siqueira, L. A. P. P. Leme, TagTheWeb: Using Wikipedia Categories to Automatically Categorize Resources on the Web, in: Lecture Notes in Computer Science, Springer International Publishing, 2018, pp. 153–157. doi:10.1007/978-3-319-98192-5_29.

[4] N. Ostapuk, D. E. Difallah, P. Cudré-Mauroux, SectionLinks: Mapping Orphan Wikidata Entities onto Wikipedia Sections, in: Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference, Virtual Conference, volume 2773 of *CEUR Workshop Proceedings*, 2020.

[5] L. Feddoul, S. Schindler, fusion-jena/wiki-category-consistency, 2022. URL: https://github.com/fusion-jena/wiki-category-consistency.

[6] L. Feddoul, S. Schindler, fusion-jena/wiki-category-consistency v1.0.2, 2022. doi:10.5281/zenodo.6963599.

[7] L. Feddoul, F. Löffler, S. Schindler, wiki-category-consistency-dataset, 2022. doi:10.5281/zenodo.6913282.

[8] L. Feddoul, F. Löffler, S. Schindler, wiki-category-consistency-cache, 2022. doi:10.5281/zenodo.6913134.

[9] L. Feddoul, F. Löffler, S. Schindler, wiki-category-consistency-eval, 2022. doi:10.5281/zenodo.6913332.

[10] H. Turki, M. A. H. Taieb, M. B. Aouicha, Coupling Wikipedia Categories with Wikidata Statements for Better Semantics, in: Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, volume 2982 of *CEUR Workshop Proceedings*, 2021.

[11] I. Johnson, Analyzing Wikidata Transclusion on English Wikipedia, in: Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference, Virtual Conference, volume 2773 of *CEUR Workshop Proceedings*, 2020.

[12] A. Boschin, T. Bonald, Enriching Wikidata with Semantified Wikipedia Hyperlinks, in: Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, volume 2982 of *CEUR Workshop Proceedings*, 2021.

[13] P. Curotto, A. Hogan, Suggesting Citations for Wikidata Claims based on Wikipedia's External References, in: Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference (OPub 2020), Virtual Conference, volume 2773 of *CEUR Workshop Proceedings*, 2020.