Reference statistics in wikidata topical subsets

Reference Statistics in Wikidata Topical Subsets

Seyed Amir Hosseini Beghaeiraveri[1][0000−0002−9123−5686],
Alasdair J.G. Gray[1][0000−0002−5711−4872], and
Fiona J. McNeill[2][0000−0001−7873−5187]

[1] School of Mathematical and Computer Sciences,
Heriot-Watt University, Edinburgh, UK
[2] School of Informatics, The University of Edinburgh, Edinburgh, UK

Abstract. Wikidata is the only general-purpose open knowledge graph with the capability of specifying references for every single statement. Currently, about 68% of Wikidata statements have at least one reference but the quality of these references is rarely covered in data quality studies. There is also a lack of a comprehensive framework for evaluating references. In this paper, we investigate the statistics of Wikidata references in 6 topical subsets of Wikidata. We compare these statistics over two Wikidata dumps; one from 2016 and one from 2021.

Keywords: Reference quality · Wikidata · Data quality · Topical subset · WikiProject · Gene Wiki

1 Introduction

Wikidata [31] is a knowledge graph that started in 2012 and is now the most active Wikimedia project. It contains knowledge on a broad range of topics with statements (data asserting a fact) being created and edited through crowdsourcing. A distinguishing characteristic of Wikidata is its ability to capture additional information about statements, such as providing references for each piece of data. According to the Wikidata project,"Wikidata is not a database that stores facts about the world, but a secondary knowledge base that collects and links to references to such knowledge" [7].

Our focus in this paper is on the references of statements. Having good evidence of where the data came from improves the trust and reusability of the data as errors can be traced, and data can be categorized according to where they came from [25,26]. According to Wikidata policy [7], all statements need to be referenced except statements about common human knowledge, statements that refer to an external source, or statements of items that are a source for other statements. Wikidata recommends that references be relevant and authoritative, but these terms are not explicitly defined. Providing appropriate references is the responsibility of the person who adds the statement. Assessment of references

is the responsibility of the Wikidata user community. Currently, about 68% of Wikidata statements already have at least one reference [4]. While there has been some initial work to look at reference quality [27], there no systematic way to assess the quality of a reference.

Wikidata aims to cover a wide range of topics via user collaborations. Users interested in a particular topic form communities called WikiProjects [16]. Besides human users, WikiProjects may use bots to collect and edit a mass of data, including references. Wikidata enforces strict rules for accepting edits by bots [5]. WikiProjects reflect the activity of contributors in covered topics. Investigating WikiProjects provides a topical comparison basis to analyze the functionality of humans and bots in different quality metrics across different Wikidata topics.

In this paper, we perform a statistical analysis on the reference statements of different WikiProjects to provide insight into their quality. Our contributions in this paper are:

1. Creating a topical comparison platform for investigating the quality of references. This is done by extracting 6 topical subsets from Wikidata corresponding to 6 different WikiProjects. We also publish the subsets for further community experiments.
2. Providing a statistical report of references in the 6 subsets.

In Section 2 we discuss related work on reference quality. Section 3 explains reference nodes in the Wikidata RDF model. Section 4 details the process of subsetting Wikidata to build topical subsets for the topical comparison platform. In Section 5, we present the statistics of references in the extracted subsets. Section 6 outlines our position on the importance of studying the quality of references and the initial ideas of a reference quality checking framework. The conclusion of the paper is presented in Section 7.

2 Related Work

Provenance of data in knowledge graphs and its quality is one of the criteria of trustworthiness which is one of the main dimensions of data quality [22]. The analysis by Farber et al. [22] gives Wikidata the full score for the trustworthiness on statement level as Wikidata can provide references for each single statement. They do not give an analysis of how Wikidata uses this feature. Farber et al. reported that the coverage of references over the statements in October 2015 is 1.3% while Wikidata Stats says more than 50% of statements had a reference at that date [4]. The reason for this difference is that Farber et al. counted the number of distinct reference nodes, while a reference node might be shared between more than one statement. We call this shared references.

Accuracy and trustworthiness have not been covered in Wikidata as much as other data quality dimensions [28]. Piscopo et al. [27] proposed an approach to evaluate the authoritativeness and the relevance of Wikidata external sources based on the quality definitions set by the Wikidata community. The approach consists of two main steps. First, a set of sample references is evaluated through

microtask crowdsourcing. Then, this data is fed to a machine-learning algorithm to apply a large-scale evaluation over the whole Wikidata dump (from October 2016). They evaluated only English language sources, mainly because of the limits of performing crowdsourcing for non-English sources. They show that Wikidata external sources are of good quality as 70% are relevant and 80% are authoritative.

Comparing between Wikidata and Wikipedia external references, Piscopo et al. [29] showed that Wikidata has a more diverse pool of sources, in terms of country of provenance, and employs a larger percentage of external databases and reference sources, such as library catalogues, compared to the online encyclopedia. More recently, Shenoy et al. [30] developed a framework to detect and analyze low-quality statements in Wikidata. Their work does not consider the quality of references as a metric. Curotto and Hugan [21] proposed a method of searching and indexing English Wikipedia references to create references for Wikidata facts. This proposal like any other reference-suggesting tool needs to be evaluated in terms of the quality of suggested references which indicates the need for a comprehensive reference quality checking framework.

The few prior work on reference quality [27,29] were applied on the 2016 and 2017 dumps of Wikidata. Given the exponential growth of Wikidata in recent years, there is a need for a comprehensive investigation on the diversity of current Wikidata references, the extent to which bots and humans participate in references, and comparisons between bots and humans regarding to the quality of references. Also, no prior work studied the reference quality across different topics in Wikidata. In this paper, by investigating reference statistics we start a path to a comprehensive review of Wikidata references. We aim to develop a broader framework by precisely defining other data quality criteria for references.

3 Wikidata Data Model

The Wikidata knowledge graph consists of items (entities from the real world) and properties (relationships between items or between items and values). An `(item, property, value)` triple in Wikidata is called a claim. A distinguishing characteristic of Wikidata is its ability to capture the provenance of each claim. This is achieved by enriching the claim with qualifiers (contextual information) and/or references (the source of the claim) to create a statement.

The Wikidata RDF model uses reification [1] for adding references to statements, as shown in Figure 1. Every statement in Wikidata has a statement node (identified by a unique ID in the `wds:` namespace) from which all references, qualifiers, ranks, and values are stored. References are linked through `prov:wasDerivedFrom` edges to reference nodes (identified by a unique ID in the `wdref:` namespace). Reference nodes provide the provenance of the fact by one or more properties like *retrieved date (P813)*, *stated in (P248)*, and *reference URL (P248)*. If a statement has multiple references, there will be a separate reference node for each reference. If two statement nodes share the same provenance, then they link to the same reference node.

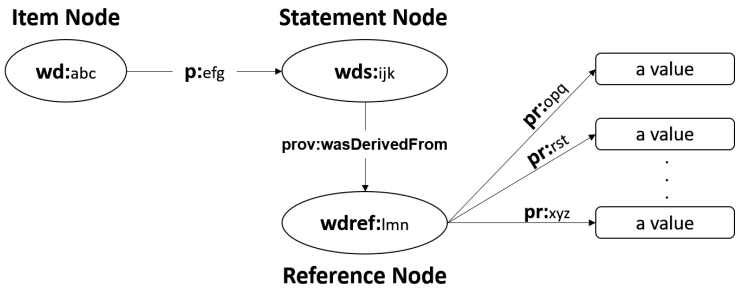

Fig. 1. The reification used in Wikidata for adding references into statements derived from [1]. `abc` is an arbitrary QID. `opq`, `rst`, and `xyz` are arbitrary PIDs.

4 Topical Subsetting of Wikidata

Investigating reference quality requires access to the reference nodes. This can be achieved by querying for the reference nodes by type, e.g. using the basic graph pattern `?item rdf:type wikibase:Reference`. The Wikidata Query Service has blocked these queries for performance reasons [1]. Due to the enormous size of Wikidata dumps, locally indexing a complete Wikidata dump is time consuming, costly, and requires hardware beyond a standard desktop computer. Therefore, we use topical subsets of Wikidata [19] which gives us substantially smaller datasets over which we can research references locally. They also provide a basis for comparing the richness and quality of referencing across Wikidata topics; thus reflecting the work of different communities. These smaller, focused, datasets are more likely to be reused [24].

4.1 Topic Selection Desiderata

A WikiProject [16] is a team of Wikidata contributors who aim to improve Wikidata by working on a specific topic or doing a specific task. A simple query[3] shows that there are 243 WikiProjects in Wikidata, many of which have been created to enrich Wikidata (both A-Boxes and T-Boxes) on a particular topic, such as music, scientific disciplines, or politics. WikiProject contributors typically list classes and properties for their topic so data instances that match these definitions can be added to Wikidata. We can use such definitions to determine the boundaries and the scope of the topic and extract their subset. These subsets are representative of their relevant WikiProject in different experiments (e.g. in reference statistics as we present in this paper).

WikiProjects vary in purpose, scope, activity, and progress. Extracting subsets for each of the projects is not feasible due to their number, nor are all of them suitable. A candidate project must meet the following desiderata:

[3] `https://w.wiki/48Rn` - accessed 27 September 2021

– It should be topical in nature. Task-based projects such as disambiguation pages [9] are not suitable for topical subsetting.
– Contributors should provide information about items, classes, and properties that are added to Wikidata through the project. This information is presented as lists, tables, entity schemas, or UML class diagrams. Using this information, we can determine the boundaries of the covered topic.
– The topic of the project should not be too limited or too broad. For example, in the Scholia project [13], just scholarly articles make up 30% of Wikidata items [6] which is very broad. We would like our candidates to have the same level of independence [12] from the whole Wikidata.
– We would like our experiment to be a good approximation of the whole Wikidata so we need candidates from a wide range of subject areas.

4.2 Selected Projects

Based on our topic selection desiderata we identified the following projects for topical subsetting to enable us to investigate reference quality. We have selected some closely related projects to allow direct comparison, and then some less related ones for contrast. We have selected a combination of scientific and non-scientific topics. The projects are of similar size and scope.

Gene Wiki [20]: Gene Wiki aims to make and maintain Wikidata as a central hub of linked knowledge on Genes, Proteins, Diseases, Drugs, and related items. It is one of the most active WikiProjects. It has five active bots and specified 24 classes of data to be added to Wikidata pictured in a UML class diagram. We include all instances of these 24 main classes and their subclasses into the subset.

Taxonomy [15]: The goal of this project is to populate Wikidata with taxonomic names and their classifications. This project consists of the class of *Taxon (Q16521)* and its hierarchy plus 47 other related classes that are specified in the wiki page of the project. The *Taxon (Q16521)* class and its subclasses are also considered in the Gene Wiki project. Considering it as a separate use case allows investigating the references in this focused part of Gene Wiki as compared to the rest.

Astronomy [8]: The main goal of this project is to define classes and properties for items related to Astronomy. Accurate referencing is one of the main goals of the project. Besides that, an active community, well-structured ontology definition, and usefulness of the project motivate us to consider this project. This subset consists of all instances of *astronomical object (Q6999)* class and its subclasses.

Law [10]: This project aims to cover anything that touches the law, e.g. economic laws, evidences, and legal proceedings. The provided data would be particularly useful for judicial systems. The project intends to be broad in scope, but it has a detailed ontology definition. *Law (Q7748)*, *public order (Q294199)*, and *evidence (Q176763)* are some of the included classes.

Music [11]: This project aims to map and import all music-related data from diverse sources to feed Wikipedia music infoboxes. Referencing is also important in this project. *Musician (Q639669)*, *musical ensemble (Q2088357)*, and *musical work(Q2188189)* are some of the main classes.

Ships [14]: This project aims to establish the most ideal structure for ship data, and create and update claims for all ship items on Wikidata. The project has a well-structured class hierarchy. Based on the mentioned items and classes on the project's web page, all instances of all subclasses of *watercraft (Q1229765)* and *ship class (Q559026)* are in the subset.

Full programmatic definitions of the subsets can be found in the supplementary material for this paper [17].

4.3 Subset Extraction Setup

We use the Wikidata WDumper [23] tool to extract subsets corresponding to each project. For each project, the main classes are identified according to the project's wiki page. Identified classes are then used to write WDumper specification files. The specification files are then enriched with subclasses via a Python script [17]. Finally, the related A-Boxes are extracted via WDumper. Subsets include all statements for A-Boxes along with references, qualifiers, and rank data. T-Boxes have been ignored as referencing does not apply to them. The WDumper specification files for each project are in [17].

For each project, two separate subsets are extracted: one from the 2016 dump (3 October 2016) [3] and one from the 2021 dump (30 June 2021) [2]. The 2021 dump was downloaded from the Wikimedia dump store [4]. We chose the 2016 dump as it is used in prior work on Wikidata reference quality. Thus, it will allow us to compare statistics between the two different snapshots and drawer some conclusions with that earlier work. The extracted subsets in N-Triples are in [18]. For this paper, the subsets were indexed and queried using Blazegraph[5] 2.1.6 triplestore.

5 Reference Statistics

We consider four experiments in which we perform a set of SPARQL queries over each extracted subset to obtain a statistical overview of references in Wikidata. The SPARQL queries for each experiment along with results can be found at the GitHub repository of the paper [17].

5.1 Basic Statistics

Table 1 shows the initial statistics for references in each project. The first two columns are the number of items and the number of statements. The

[4] `https://dumps.wikimedia.org/other/wikibase/wikidatawiki/` - accessed 2 July 2021

[5] `https://github.com/blazegraph/database/releases/tag/BLAZEGRAPH_2_1_6_RC`

Table 1. The basic statistic of references in each subset.

Project	Dump	Items	Statement Nodes	Reference Nodes	Referenced Statements	Shared Reference Nodes
Gene Wiki	2016	2,647,174	17,656,669	169,493	8,789,246(50%)	42,902(25%)
	2021	8,801,623	92,729,475	9,559,517	65,780,005(71%)	4,700,610(49%)
Taxonomy	2016	2,214,088	16,056,914	95,714	8,146,218(51%)	5,971(06%)
	2021	3,225,102	32,536,083	498,535	19,423,938(60%)	204,602(41%)
Astronomy	2016	141,843	888,717	13,260	751,158(85%)	12,198(92%)
	2021	8,416,958	144,637,511	157,558	128,394,763(89%)	112,365(71%)
Law	2016	67,763	174,252	380	48,225(27%)	152(40%)
	2021	433,440	4,236,657	407,409	2,266,462(53%)	317,975(78%)
Music	2016	598,074	3,742,474	80,857	2,298,330(61%)	35,574(44%)
	2021	948,266	11,702,021	1,329,746	6,342,019(54%)	374,440(28%)
Ships	2016	42,873	183,240	857	114,528(62%)	227(26%)
	2021	126,896	1,101,802	59,282	315,381(29%)	16,396 (28%)

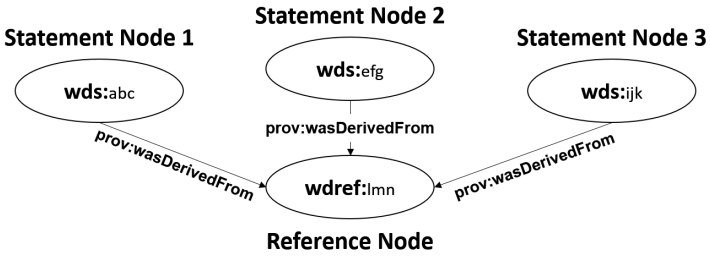

Fig. 2. An example of a reference node that is shared between three statement nodes

third column is the number of reference nodes (i.e. nodes with the type `wikibase:Reference`). The fourth column shows the number and percentage of statements with at least one reference. The difference between the number of referenced statements and the number of reference nodes is significant. This is because a number of reference nodes are common between statements. In other words, a number of statements have exactly the same references. We call these *shared references*, which is shown in Figure 2. The fifth column shows the number and percentage of those reference nodes that are shared between more than two statements.

As we can see from the table, in all projects the number of items, statement nodes, and reference nodes has substantially increased from 2016 to 2021. After the extraction, we recognized that Taxonomy project makes up about 30% of Gene Wiki. The percentage of referenced statements has increased in all cases except Music and Ships. In the case of Ships, the percentage of referenced statements has dramatically decreased. Considering the increase in statements in both, the decrease in referenced statement can show that human users are more active than bots in Ships and Music (if we intuitively accept that bots provide

references more and better than humans). The percentage of shared references for Gene Wiki, Taxonomy, Law, and Ships has increased from 2016 to 2021, while for Astronomy and Music this amount has decreased. Among the 2021 datasets, the highest number of referenced statements belongs to the Astronomy project and the lowest to the Ships project. The increase in shared references in the Gene Wiki and Taxonomy subsets is likely due to the use of bots to populate Wikidata. Considering the 2021 datasets, the highest number of shared references is allocated to the Law project and the lowest to the Taxonomy project.

5.2 Usage of Reference-specific Properties

Wikidata offers a set of properties such as *stated in (P248)* and *reference URL (P854)* to be used in references. In addition, different projects may offer properties for their references, e.g. the Gene Wiki and Taxonomy projects use properties such as *IUCN taxon ID (P627)* even though they are identifier properties. Figure 3 shows the frequency of reference-specific properties used in references in each use case for 2021 subsets. Note that, Figure 3 illustrates only the most used properties; the variety of properties is more but the abundance of the remaining properties is less than 3% overall. For details, see the CSV file at the GitHub repository of the paper[6].

In Gene Wiki, Taxonomy, and Law, the most frequently used properties are *stated in (P248)*, *retrieved (P813)*, and *reference URL (P854)*, while Music makes most use of the first two. This indicates that most of the references in these subsets rely on external sources that were likely populated by bots. For Gene Wiki and Taxonomy, the next most frequently used properties correspond to identifier properties for well known data sources in the life sciences. It is likely that these are used to indicate these data sources as the provider of the claim. The use of external sources accounts for about 60% (Music) to 100% (Taxonomy) of the references. In Astronomy (58%) and Ships (56%), the most frequently used properties are *imported from Wikimedia project (P143)* and *Wikimedia import URL (P4656)*. These properties indicate that the source of the statement is one of the internal Wikimedia projects, e.g. Wikipedia. Mentioning the Wikipedia article as a source for corresponding Wikidata item is not recommended [7], so the extent of these should be carefully considered in future studies.

5.3 Distribution of Triples per Nodes

Via the reference-specific properties, each reference node uses one or more triples to point to the provenance of the claim. Figure 4 shows that the most frequent properties are probably used together. Having more triples in a reference node provides more details about the source which is likely to increase the accuracy. Figure 4 shows the distribution of the number of triples over the total reference nodes in each project in 2016 and 2021 dumps. In all projects except Law

[6] `https://github.com/seyedahbr/Wikidata_Reference_Statistics/blob/main/QueryResults/UsageofReference-specificProperties/PropertyUsage.xlsx`

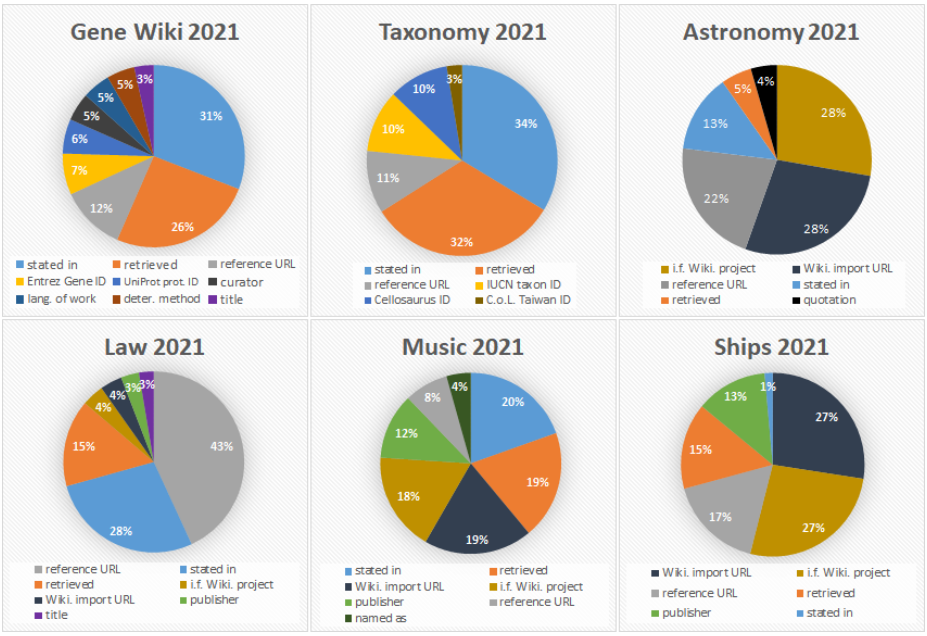

Fig. 3. The frequency of reference-specific properties used in references in each project (2021 subsets).

and Ships, the average number of triples in references has decreased from 2016 to 2021. The best average belongs to Gene Wiki. The similarity of Gene Wiki statistics in both 2016 and 2021 dumps is interesting and is probably related to the steady activities of the project bots. The uniform distribution of triples in taxonomy might be due to the steady activity of the bots in a specific field (as opposed to Gene Wiki, which consists of several fields such as biology, chemistry, and pharmacology). In 2021, Astronomy has the lowest average number of triples in reference nodes, despite having the highest percentage of referenced statements. In the Music project, there are reference nodes with 22 and 35 triples; these outliers are omitted from the figure for presentation purposes. The average number of triples ranges between 1.2 (Ships 2016) and 3.5 (Gene Wiki 2021).

5.4 Distribution of Reference Sharing

Shared reference nodes can affect the quality of references. Having shared references is not necessarily negative as they can reduce redundancy. For example, in Gene Wiki multiple statements about a protein might be taken simultaneously from the UniProt dataset via a bot, so the reference node of all these statements will be the same. Figure 5 shows the distribution of reference sharing of each project in the 2016 and 2021 dumps. In all projects except Astronomy, the reference sharing rate has decreased from 2016 to 2021. Although Figure 5 shows the

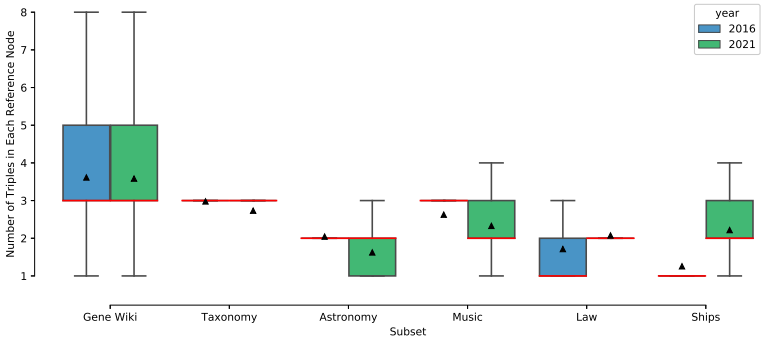

Fig. 4. The distribution of triples per reference nodes. The red lines are the medians. Triangles are mean points. Outliers are omitted for presentation purposes.

Table 2. Rounded mean and maximum of reference sharing in 2021 dump for each project.

	Gene Wiki	Taxonomy	Astronomy	Law	Music	Ships
Mean	13	93	1142	7	14	17
Max	1,281,307	408,522	42,876,186	155,508	1,385,109	96,659

normal distribution rate in shared reference nodes, there are exception reference nodes in each project shared between a very large number of statements. Table 2 shows the mean and maximum of reference sharing in each project in the 2021 dump. In Astronomy, there are about 43 million statements connected to just one reference node, however, there is only one reference node with such situation. In all projects, there are reference nodes that are providing the source of more than 50,000 statements. This amount of sharing might challenge the relevancy condition [7] and should be carefully examined.

6 From Statistics to Quality

To the best of our knowledge, the Piscopo et al. study [29] is the only research on the quality of references in Wikidata, but this work has considerable limitations. They started with the Wikidata edit history in October 2016. They extracted all statements containing external references. Then they excluded statements that do not require reference according to Wikidata policy which leads them to 1,629,102 references. At this step, 89% of references pointed to two specific sources[7] that are excluded from the evaluation as around 98% of these were added by one bot. In the remaining 11%, they evaluated only English sources that are about 46%. In the end, only 83,215 references were reviewed. The number of statements and references is completely changed now. We can see from

[7] uniprot.org and ebi.ac.uk

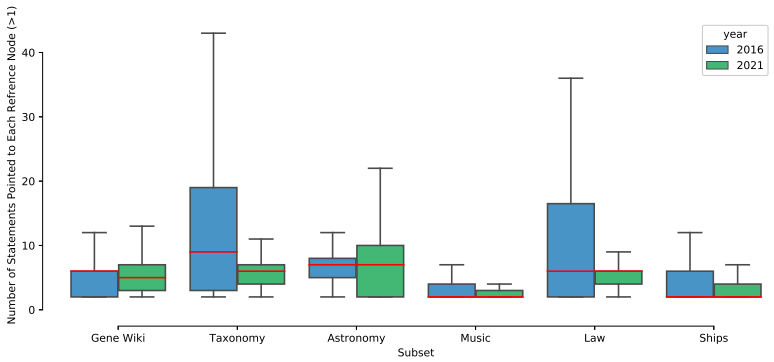

Fig. 5. The distribution of reference sharing (for reference nodes that are connected to ≥ 2 statement). The red lines are the medians. Outliers and means are omitted for presentation purposes.

[4] that the number of statements has increased 10 fold and the percentage of statements referring to external sources has also increased from 25% to 68%. Figure 3 confirms that currently there is a diversity in the most used properties in references. All of these mean that there is a possibility of greater diversity in references and a need for a comprehensive evaluation.

The impact of bots on the quality of references should also be examined. Although Wikidata has strict policies for using bots, the effect of bots on references has not been studied. The challenge here is that tracking bot activities requires processing Wikidata edit history, which is ten times larger than the current Wikidata dump. Shared references can also be a potential factor in reference quality because they can at least challenge the relevancy condition.

Currently, the most important shortcoming is the lack of a framework to examine the quality of referencing in Wikidata and other knowledge graphs. Our idea is a scoring system that can evaluate different criteria on references and quantify the result. For this scoring system, different criteria should be defined according to the references. Relevancy and authoritativeness have been suggested by the Wikidata community for references. There are also data quality criteria such as Accuracy, Accessibility, Consistency, and Completeness that need to be accurately defined according to the context of the references and reference-specific properties. For example, accessibility can be defined as the availability of the links mentioned in the references.

The above criteria apply to single references, but criteria such as shared references should be considered on the whole of Wikidata (or its subsets). Furthermore, some criteria can be measured by the machine, while some others such as relevancy and authoritativeness are subjective, and evaluating them requires machine training with human intervention.

Statistical information can be effective in defining some criteria. For example, using the information of Section 5.3, we can determine a minimum number of

triples that reference nodes should have. Section 5.2 also tells us what properties are most commonly used in references so we will be able to define necessary criteria appropriate to those properties. Our plan for this scoring system is to identify the necessary criteria, provide a precise definition of them, and measure them on the six subsets as well as a random sample from Wikidata. We also plan to separate human references and bot references to compare the quality between them.

7 Conclusion

In this paper, we performed a statistical review of the references in Wikidata. We extracted six independent Wikidata subsets corresponding to 6 different WikiProjects and reviewed reference statistics in them. These statistics can be used by project contributors to improve Wikidata, e.g. correcting the properties used in their project, reviewing shared references, and trying to provide a sufficient number of triples. The subsetting method used can be replicated for other Wikidata projects and other fields of study.

Our statistics show the importance of a more in-depth study of Wikidata references. We stated our position of the need for a reference quality scoring system based on data quality dimensions and provided basic ideas for the system. Such an assessment system can provide precise and detailed suggestions to Wikidatians/WikiProject holders. Our future work is to complete the definition and development of the reference quality scoring system. We aim to perform a comprehensive evaluation on Wikidata references, using WikiProjects along with randomly selected subsets. The challenges for the future work are the large volume of data, tracing bot/human edits, and the subjective nature of the concepts.

Acknowledgements

We would like to acknowledge the useful guidance and fruitful discussions with the ShEx Community Group[8]; Kat Thronton, Andra Waagmeester, Dan Brickley, and Eric Prud'hommaux.

References

1. Wikibase/Indexing/RDF Dump Format - MediaWiki, `https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format`, accessed 2021-06-30
2. Wikidata entity dump (JSON) generated on June 30, 2021 : Free Download, Borrow, and Streaming, `https://archive.org/details/wikidata-20210630.json.gz`, accessed 2021-09-27

[8] `http://shex.io/` - accessed July 2021

3. Wikidata entity dumps (JSON and TTL) of all Wikibase entries for Wikidata generated on October 03, 2016 : Wikidata editors : Free Download, Borrow, and Streaming, https://archive.org/details/wikibase-wikidatawiki-20161003, accessed 2021-06-30
4. Wikidata Stats, https://wikidata-todo.toolforge.org/stats.php, accessed 2021-06-30
5. Wikidata:Bots - Wikidata, https://www.wikidata.org/wiki/Wikidata:Bots, accessed 2021-06-30
6. Wikidata:Statistics/Wikipedia/Type of content - Wikidata, https://www.wikidata.org/wiki/Wikidata:Statistics/Wikipedia/Type_of_content, accessed 2021-06-30
7. Wikidata:Verifiability - Wikidata, https://www.wikidata.org/wiki/Wikidata:Verifiability, accessed 2021-06-30
8. Wikidata:WikiProject Astronomy - Wikidata, https://www.wikidata.org/wiki/Wikidata:WikiProject_Astronomy, accessed 2021-06-30
9. Wikidata:WikiProject Disambiguation pages - Wikidata, https://www.wikidata.org/wiki/Wikidata:WikiProject_Disambiguation_pages, accessed 2021-06-30
10. Wikidata:WikiProject Law - Wikidata, https://www.wikidata.org/wiki/Wikidata:WikiProject_Law#Participants, accessed 2021-06-30
11. Wikidata:WikiProject Music - Wikidata, https://www.wikidata.org/wiki/Wikidata:WikiProject_Music#Overview, accessed 2021-06-30
12. Wikidata:WikiProject Schemas/Subsetting - Wikidata, https://www.wikidata.org/wiki/Wikidata:WikiProject_Schemas/Subsetting, accessed 2021-06-30
13. Wikidata:WikiProject Scholia - Wikidata, https://www.wikidata.org/wiki/Wikidata:WikiProject_Scholia, accessed 2021-06-30
14. Wikidata:WikiProject Ships - Wikidata, https://www.wikidata.org/wiki/Wikidata:WikiProject_Ships, accessed 2021-06-30
15. Wikidata:WikiProject Taxonomy - Wikidata, https://www.wikidata.org/wiki/Wikidata:WikiProject_Taxonomy, accessed 2021-06-30
16. Wikidata:WikiProjects - Wikidata, https://www.wikidata.org/wiki/Wikidata:WikiProjects, accessed 2021-06-30
17. Beghaeiraveri, S.A.H.: Wikidata reference statistics. https://github.com/seyedahbr/Wikidata_Reference_Statistics (2021)
18. Beghaeiraveri, S.A.H.: Wikidata Subsets of 6 Wikiproject (Gene Wiki, Taxonomy, Astronomy, Music, Law, Ships) (Jul 2021). https://doi.org/10.5281/zenodo.5117928, https://doi.org/10.5281/zenodo.5117928
19. Beghaeiraveri, S.A.H., Gray, A.J.G., McNeill, F.J.: Experiences of Using WDumper to Create Topical Subsets from Wikidata. In: CEUR Workshop Proceedings. vol. 2873, p. 13. CEUR-WS (Jun 2021), https://researchportal.hw.ac.uk/en/publications/experiences-of-using-wdumper-to-create-topical-subsets-from-wikid, iSSN: 1613-0073
20. Burgstaller-Muehlbacher, S., Waagmeester, A., Mitraka, E., Turner, J., Putman, T., Leong, J., Naik, C., Pavlidis, P., Schriml, L., Good, B.M., Su, A.I.: Wikidata as a semantic framework for the Gene Wiki initiative. Database (Oxford) **2016** (2016). https://doi.org/10.1093/database/baw015, https://academic.oup.com/database/article-lookup/doi/10.1093/database/baw015
21. Curotto, P., Hogan, A.: Suggesting citations for wikidata claims based on wikipedia's external references. In: Wikidata@ ISWC (2020)

22. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web **9**(1), 77–129 (Nov 2017). https://doi.org/10.3233/SW-170275, https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-170275
23. Fünfstück, B.: Wdumper. https://github.com/bennofs/wdumper (2019)
24. Koesten, L., Vougiouklis, P., Simperl, E., Groth, P.: Dataset Reuse: Toward Translating Principles to Practice. Patterns **1**(8), 100–136 (Nov 2020). https://doi.org/10.1016/j.patter.2020.100136, https://www.sciencedirect.com/science/article/pii/S2666389920301847
25. Lehmann, J., Gerber, D., Morsey, M., Ngonga Ngomo, A.C.: DeFacto - Deep Fact Validation. In: The Semantic Web – ISWC 2012. pp. 312–327. Lecture Notes in Computer Science, Springer (2012). https://doi.org/10.1007/978-3-642-35176-1_20
26. Lucassen, T., Schraagen, J.M.: Trust in wikipedia: how users trust information from an unknown source. In: Proceedings of the 4th workshop on Information credibility. pp. 19–26. WICOW '10, Association for Computing Machinery, Raleigh, North Carolina, USA (Apr 2010). https://doi.org/10.1145/1772938.1772944, https://doi.org/10.1145/1772938.1772944
27. Piscopo, A., Kaffee, L.A., Phethean, C., Simperl, E.: Provenance Information in a Collaborative Knowledge Graph: An Evaluation of Wikidata External References. In: The Semantic Web – ISWC 2017, vol. 10587, pp. 542–558. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-68288-4_32, http://link.springer.com/10.1007/978-3-319-68288-4_32, series Title: Lecture Notes in Computer Science
28. Piscopo, A., Simperl, E.: What we talk about when we talk about wikidata quality: a literature survey. In: Proceedings of the 15th International Symposium on Open Collaboration. pp. 1–11. ACM, Skövde Sweden (Aug 2019). https://doi.org/10.1145/3306446.3340822, https://dl.acm.org/doi/10.1145/3306446.3340822
29. Piscopo, A., Vougiouklis, P., Kaffee, L.A., Phethean, C., Hare, J., Simperl, E.: What do Wikidata and Wikipedia Have in Common?: An Analysis of their Use of External References. In: Proceedings of the 13th International Symposium on Open Collaboration - OpenSym '17. pp. 1–10. ACM Press, Galway, Ireland (2017). https://doi.org/10.1145/3125433.3125445, http://dl.acm.org/citation.cfm?doid=3125433.3125445
30. Shenoy, K., Ilievski, F., Garijo, D., Schwabe, D., Szekely, P.: A Study of the Quality of Wikidata. arXiv:2107.00156 [cs] (Jun 2021), http://arxiv.org/abs/2107.00156, arXiv: 2107.00156
31. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (Sep 2014). https://doi.org/10.1145/2629489, https://dl.acm.org/doi/10.1145/2629489