What do Wikidata and Wikipedia Have in Common? An Analysis of their Use of External References

Alessandro Piscopo
University of Southampton
United Kingdom
A.Piscopo@soton.ac.uk

Pavlos Vougiouklis
University of Southampton
United Kingdom
pv1e13@ecs.soton.ac.uk

Lucie-Aimée Kaffee
University of Southampton
United Kingdom
kaffee@soton.ac.uk

Christopher Phethean
University of Southampton
United Kingdom
C.J.Phethean@soton.ac.uk

Jonathon Hare
University of Southampton
United Kingdom
jsh2@ecs.soton.ac.uk

Elena Simperl
University of Southampton
United Kingdom
E.Simperl@soton.ac.uk

ABSTRACT

Wikidata is a community-driven knowledge graph, strongly linked to Wikipedia. However, the connection between the two projects has been sporadically explored. We investigated the relationship between the two projects in terms of the information they contain by looking at their external references. Our findings show that while only a small number of sources is directly reused across Wikidata and Wikipedia, references often point to the same domain. Furthermore, Wikidata appears to use less Anglo-American-centred sources. These results deserve further in-depth investigation.

ACM Classification Keywords

H.5.3. Group and Organization Interfaces : Collaborative computing; H.1.1. User/Machine Systems: Human information processing; K.6.4. System Management: Quality assurance

Author Keywords

Wikidata; Wikipedia; citations; provenance.

INTRODUCTION

Wikidata is a community-driven knowledge graph run by the Wikimedia Foundation since 2012 [17]. Knowledge graphs are large collections of terms describing real world entities and their relations [13]. They are key to providing structured data that can be easily processed and reasoned upon by a range of applications on the Web. Several knowledge graphs already exist. Nonetheless, in spite of its relatively young age, Wikidata has already raised considerable interest among practitioners and researchers, due to some of its prominent features.

One strength of Wikidata is to rely on a community to edit and maintain its content. Since the inception of the project, the number of registered editors has grown up to more than one hundred thousand. These users have gathered facts about around 24 million entities and are able, at least theoretically, to further expand the coverage of the knowledge graph and continuously keep it updated and correct. This is an advantage compared to a project like DBpedia, where data is periodically extracted from Wikipedia and must first be modified on the online encyclopedia in order to be corrected.

Another strength is that all the data in Wikidata can be openly reused and shared without requiring any attribution, as it is released under a CC0 licence[1]. Because of this, Wikidata may act as a source for a virtually unlimited number of information-based systems and applications. Projects relying on expert-curated knowledge bases, as in the case of the question answering system Wolfram Alpha, are able to provide high quality information, but due to their high maintenance costs they cannot be freely reused by third parties.

Wikidata's knowledge is encoded in a property-value pairs model. Whereas this format may not be appealing to read for humans, it allows machines to process data and easily extract tailored information. In contrast, Wikipedia articles may be more pleasant to read for humans, but the natural language they are written in is difficult to process automatically. Wikidata was initially conceived as a structured backbone for Wikipedia, in order to improve consistency between different versions of the online encyclopedia and reduce the workload for editors [18]. These two projects are thus interconnected at several levels. Data is actively exchanged and reused across the two platforms, e.g. Wikidata was used to populate Wikipedia infoboxes with structured data. Their communities partially overlap, as noted in [14]. Although research has already addressed some aspects of these links, no study has investigated yet the relationship between Wikidata and Wikipedia in terms of their information. An increased understanding of this relationship would be beneficial for these projects, in order to facilitate the discovery of problematic aspects concerning the production of knowledge and of points where synergies between them may be exploited. Furthermore, Wikidata was designed with knowledge diversity in mind [19]. Examin-

OpenSym '17 August 23–25, 2017, Galway, Ireland

ACM ISBN 978-1-4503-5187-4/17/08.

DOI: https://doi.org/10.1145/3125433.3125445

[1] https://creativecommons.org/publicdomain/zero/1.0/, consulted on 15 April 2017.

ing its connection with different Wikipedias can provide an indication of the extent to which this has been achieved.

This paper seeks to gain insights about how Wikidata's and Wikipedia's knowledge are interconnected by analysing their external sources. Both projects require to provide supporting references for each piece of information. We performed a descriptive analysis of external references in Wikidata and Wikipedia, overall and across different language versions, under the assumption that a higher number of common external sources would indicate a closer relationship between the two. The difficulty to track sources from which a knowledge base was created may cause a lack of transparency in the information extracted from it [5]. Therefore, looking at Wikidata and Wikipedia sources may shed light on their trustworthiness and possible biases. Our work aims to be a starting point for further research delving into how information is generated and circulates between Wikidata and Wikipedia. In particular, our contributions are: (***i***) a comparison between Wikidata and Wikipedia information sources, concerning their number and types; (***ii***) an examination of sources reuse across the two platforms; (***iii***) an analysis of the relationship between Wikidata and different language versions of Wikipedia.

Our findings show that while only a low percentage of pages is directly shared between the two projects, these have several web domains in common. Moreover, sources on Wikidata are less US– and UK–centred than those used over all Wikipedias.

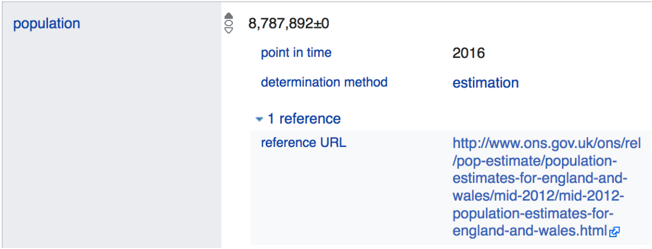
BACKGROUND INFORMATION AND RELATED WORK

The first part of this section describes Wikidata and its connection with Wikipedia. Following, we compare their policies and guidelines about reference use and present a selection of work of relevance about the topic.

Wikidata: a Collaborative Knowledge Graph

Like Wikipedia, Wikidata adopts a wiki model to enable user collaboration. Anyone can edit, content is arranged in pages organised in namespaces, and spaces are created for community discourse. Users are responsible to edit and maintain the whole knowledge graph. This includes instances, i.e. entities in the real world such as "London", and the conceptual structure of knowledge, which defines classes like "capital city" and their relationships with instances.

Prior to Wikidata, the same wiki model has been used in semantic wikis to incorporate semantics and machine-readable data into text, thus intertwining structured and unstructured data [9]. Conversely, Wikidata is purely a collection of structured data. Furthermore, most semantic wikis have a specialist focus, in contrast with Wikidata's aim to have general coverage. Another relevant project was Freebase, a knowledge graph which combined expert and layman contributions to edit and store structured data [1]. However, it was shut down in 2015 and its community was likely smaller than Wikidata's[2].



population
8,787,892±0
point in time 2016
determination method estimation
1 reference
reference URL http://www.ons.gov.uk/ons/rel/pop-estimate/population-estimates-for-england-and-wales/mid-2012/mid-2012-population-estimates-for-england-and-wales.html

Figure 1. An example of Wikidata statement, representing the population of London— which is the subject item and is not shown in the image. Qualifiers specify the point in time to which the information is referred and how the figure was determined. An external reference is attached.

The building blocks of Wikidata are *items* and *properties*. Items represent any concrete entity, e.g. the Victory of Samothrace[3], or abstract concepts, such as the class of all Hellenistic sculptures. Properties are used to state relationships between items or between items and literal values. Both items and properties use alphanumeric language-independent identifiers, which consist respectively in a Q or a P followed by a number. Items are the subject of property-value pairs, called *claims*, which can be enriched by qualifiers and/or references, forming a *statement*. Wikidata items consist in a series of item-property-value statements, rather than in discursive articles like in Wikipedia. Qualifiers provide further contextual information to claims and can set a limitation to its validity, such as the date that it refers to (see Figure 1). References link to sources which provide evidence for a claim. The ensuing section provides further details about references.

Wikidata and Wikipedia

Wikidata has been design having in mind some of the issues already affecting Wikipedia [18]. One of the first functions of Wikidata has been to act as a centralised hub for Wikipedia's inter-language links. Wikidata's language-independent identifiers remove the need to have localised versions of items and properties, which can then store all the links connecting different language versions of Wikipedia articles [18]. Due to this, a Wikidata item exists for each Wikipedia article, but not the other way round, i.e. there may be Wikidata items with no corresponding Wikipedia article. Another effect of Wikidata multilinguality is to avoid divergent, if not contradicting, information on the same topic, an issue known to affect Wikipedia, due to its different language versions [15]. As in Wikipedia, Wikidata editors can communicate through talk pages. However, these are seldom used, given the language diversity of the system [18]. Wikidata allows different versions of the same claim, provided they are backed by references and defined by qualifiers. This approach reduces the possibility of the outbreak of edit wars, moving away from Wikipedia's consensus-based mechanism to solve conflicts between contrasting viewpoints. Edit wars have long affected the free encyclopedia [7]. Moreover, Wikidata's liberal approach facilitates the expression of diverse points of view, addressing another problematic point of its sister project [3].

[2] The literature consulted did not provide any precise information about the size of the Freebase community. However, all the information pointed to a smaller user pool than Wikidata.

[3] https://www.wikidata.org/wiki/Q216402, consulted on 11 July 2017.

The body of literature covering both Wikidata and Wikipedia mainly focuses on editing activity patterns or community features. Steiner *et al.* analyse the share of work carried out by different types of users, i.e. bots, anonymous, and registered human editors, in different language versions of Wikipedia and in Wikidata [16]. Wikidata has higher levels of bot activity, but lower percentages of anonymous edits, compared to the majority of Wikipedias. In [14] we observed a partial overlap between the communities of Wikidata and Wikipedia. The overwhelming majority of highly active users on Wikidata has been previously engaged in the free encyclopedia. A common pattern among these users is a gradual shift of their activity focus from Wikipedia to Wikidata, increasingly assuming responsibilities within the community of the latter.

Wikidata References

Wikidata has been defined a secondary database, collecting claims from primary sources, rather than stating facts about the real world [19]. Policies and guidelines define types of references, when they are needed, and what criteria they should comply with [20, 22]. Every Wikidata statement, at least theoretically, must be supported by a verifiable source. All statements not providing a reference are deemed unverified and should be removed. Some statements are exempt from including a reference, though. These are undisputed claims, representing common knowledge, e.g. *Earth*, *instance of*, *planet*; when a statement connects an item to an external source of information, e.g. an Id in a database; and when the source for a statement is an item itself, e.g. for a book and its author.

Various types of references exist in Wikidata. Internal references connect to other items within the knowledge graph and are mainly connected through property P248 (*stated in*). External ones point instead to web pages through property P854 (*reference URL*). Other properties can be used to specify sources, such as P143 (*imported from*), P887 (*based on heuristic*), and P1343 (*described by source*), or to add further details to the reference. This paper focuses on external references, i.e. those using P854, since they provide direct information about the sources used to build the knowledge graph and can be compared across Wikidata and Wikipedia. Sources must be authoritative, i.e. they should be trustworthy, free of bias, and up-to-date [24]. Authoritative sources include books, academic publications, news and media, and policy and legislation sources. These types of sources are generally represented by items. Web pages must instead meet the requirements set in Wikidata verifiability policy to be considered authoritative. This policy explicitly refers to its homologue in Wikipedia. Pages published by a company, organisation, or government agency, or having an identifiable author are generally authoritative. Self-published sources can be used as sources of information only with regard to their author. Finally, references pointing to user-generated content, such as blogs or the IMDb[4], are deemed not authoritative and deprecated.

Citations in Wikipedia

All information in Wikipedia articles must be verifiable and supported by a reliable source [24]. Potentially contentious material should include a citation providing evidence for it, otherwise it may be removed by. the community. Wikipedia defines its verifiability policy in more detail than Wikidata[5]. Three factors determine the reliability of a source: the source itself, its publisher, and its author. As a general rule, all sources that underwent some sort of publishing process and editorial oversight should be regarded as authoritative. Up-to-date sources are preferable to older ones. Since Wikipedia must not contain original research, the use of secondary sources is advised, rather than primary ones, which may open to different interpretations. Scholarly and news sources are generally considered authoritative. Biased sources should be avoided. These may include vendor, or political or religious association sites, or even news sources in some cases. Similarly to Wikidata, self-published sources should only support statements about their authors and user-generated content is not acceptable.

Citation Analysis

Information source use is part of a more general information-seeking behaviour [25]. Sources can be analysed and their use compared between different cohorts, in order to understand their knowledge-building processes. An example of this type of study is in [12]. Citations in faculty publications and student theses are compared, in order to understand patterns in information source use, as an indicator of research habits and purposes and a background for managing journal collections. The analysis is carried out in a quantitative fashion, by comparing several statistics related to number, type, and currency of citations. Type and age of materials used are similar between the two groups, whereas there are differences concerning the number of citations of journals and monographies.

While no research has addressed yet the use of external sources in Wikidata, several studies have covered Wikipedia citations. Ford *et al.* assess Wikipedia sources by comparing them to the guidelines established by the community [4]. They manually classify a sample of 500 sources by type, publisher, and author, according to the verifiability policy discussed above. Their findings show a mismatch between the requirements set for high-quality sources and those actually used. The majority of citations are from primary sources, which should be avoided. Furthermore, the most common authors of citations are by far identifiable individuals (50%) or organisations (45%), compared to only 5% of collectively authored sources. Only 16% of sources are published by academic institutions, which should be preferred, according to the Wikipedia policy.

Other studies compare citations in Wikipedia against other sources, in order to gain an understanding of the quality and reliability of the online encyclopedia. The correlation between number of outbound links to scientific journals and journal statistics from Journal Citation Reports (JCR) is examined in [11]. The number of citations of journal papers in Wikipedia appear to be highly correlated to the journals' number of of citations in JCR and to a lower extent to their impact factor. [10] draw attention to the Wikipedia verifiability policy as a means not only to prevent inaccuracies, but also to increase the credibility of the encyclopedia. Moving from this assumption, the authors of the research evaluate references from a sample

[4] http://www.imdb.com/, consulted on 16 April 2017.

[5] We refer to the English version of this policy.

of Wikipedia's history articles, by comparing them with issues of the Journal of World History (JWH). The comparison takes into consideration quantitative features, such as the total number of citations and the number of citations per statement. The results are not favourable to Wikipedia, as a large number of claims are not supported by any citations and the reference pool lacks diversity, as it includes mainly sources from the US government and online media news [10].

Huvila follows a qualitative approach to understand how Wikipedia contributors retrieve their information [6]. The most used sources are web sites, books, and news, similarly to what reported by other works, e.g. in [4]. Users can be classified into a number of profiles, according to the prevalent type of source they rely upon when adding new content or modifying existing one. As an example, *surfers* tend to primarily use search engines to find their sources, whereas *scholars* work on their area of expertise and rely on a large number of academic and scholarly sources.

The current study is the first comparative analysis of information sources use in Wikidata and Wikipedia. A qualitative approach such as that in [6] would have permitted us to gain a somewhat deeper understanding of the information-seeking behaviour of editors in these two platforms. However, this research aims to provide a descriptive overview, which could be a starting point for future studies. Thus, we followed an approach similar to [12], by comparing the types of external sources found in Wikidata and Wikipedia and examining to extent to which they are used for the same topic. Next section presents the research questions addressed, while the following illustrates the measures collected and the methods used for that purpose.

RESEARCH QUESTIONS

In the previous sections we have looked at the relationship between Wikidata and Wikipedia are connected and described their approaches to external source use. The aim of our first research question is to understand how their different requirements translate quantitatively into actual use, specifically with regard to coverage and type of sources.

RQ1 How does the use of external sources differ in Wikidata and Wikipedia, with regard to their number and type?

Moreover, in order to understand how Wikidata and Wikipedia are connected with respect to the information they contain, we decided to look at the use of sources on the same topic across the two platforms. We therefore posed the question:

RQ2 To what extent are sources reused across Wikidata and Wikipedia?

Wikipedia is not a monolithic project. The same topic can have different coverage in various language versions or be presented under different viewpoints [15]. Therefore, stronger connections with one or another version may signify that Wikidata is affected by a determined cultural bias, in contrast with the diversity sought by its creators [19]. Besides, the influence of Wikipedia may affect various knowledge domains to a different extent. Therefore, the last two research questions were:

RQ2.A How does the number of common sources vary between different language versions of Wikipedia?

RQ2.B How does the number of common sources vary between different domains of Wikidata?

METHODS

We adopted a descriptive quantitative approach to answer our research questions. We present this approach in this section.

In order to address RQ1 we wanted to provide a comprehensive comparison of the use of external references in Wikidata against Wikipedia. Specifically, we looked at a number of features:

- *number of references per Wikidata item vs. per Wikipedia articles*, across all versions and per language. We counted each reference in each item/article, regardless of whether multiple references used the same source. This feature was chosen to describe the extent to which items and articles are supported by external sources in the two projects;
- *number of uses per single reference (URL) in Wikidata vs. Wikipedia*, across all versions and per language. This feature and the next two aim to provide a measure of the diversity of sources used in each project;
- *number of domains*[6] *used in Wikidata vs. Wikipedia*, across all versions and per language;
- *number of uses per single domain in Wikidata vs. Wikipedia*, across all versions and per language;
- *distribution of top-level domains (TLDs) in Wikidata vs. Wikipedia*, across all versions and per language; this feature was chosen to reflect the geographic diversity of the sources employed [4].

Using TLDs to describe geographic diversity may sound simplistic. An example of a more elaborate approach is the Wikiwhere algorithm developed by Körner *et al.* in [8], which applies machine learning to classify the provenance country of Wikipedia citations. Although its accuracy is higher than simpler approaches, such as using IP location or TLDs, Wikiwhere has been trained to generate labels only for a limited number of countries. Therefore, we relied on TLDs to determine the geographic provenance of sources. This approach has shown an accuracy of around 60% [8].

In order to understand how sources are reused across Wikidata and Wikipedia (RQ2) we computed the number of matching references between Wikidata items and their corresponding Wikipedia articles. With this term, we refer to unique sources that were common to items and articles. Furthermore, we computed the number of matching domains. This is the number of Wikidata item reference domains that were also found in the corresponding Wikipedia articles. The rationale of this choice was that Wikidata and Wikipedia follow slightly different principles when it comes to citing external sources. Hence, they may add citations to different types of statements, even in articles/items regarding the same topic. Matching domains would indicate a common source about the same topic, although possibly with respect to different aspects. For example,

[6] A domain is the highest level in the administrative hierarchy of a web site, e.g. `gov.uk/` is the domain of `https://data.gov.uk/data/search?theme-primary=Environment`.

the 'London' Wikidata item may use a reference from the BBC to support its statements about population figures, whereas the corresponding English Wikipedia article may rely on the same website to provide a citation about the new mayor. Therefore, we can assume that in both cases the BBC was considered as a reliable source of information. Matching references and matching domains aimed to estimate the share of Wikidata items that have sources in common with their analogous Wikipedia articles We gauged the following features to answer RQ2:

- *percentage of Wikidata items with matching references* with the corresponding Wikipedia articles, over the total number of items in our dataset;
- *percentage of Wikidata items with matching domains* with the corresponding Wikipedia articles, over the total number of items in our dataset;
- *percentage of matching references* between Wikidata items and corresponding Wikipedia articles, over the total number of Wikidata references;
- *percentage of matching domains* between Wikidata items and all corresponding Wikipedia articles, over the total number of domains in Wikidata references;
- *number of matches per reference* between Wikidata items and all corresponding Wikipedia articles;
- *number of matches per domain* between Wikidata items and all corresponding Wikipedia articles;
- *correlation between the number of common domains and common authors* between Wikidata items and all corresponding Wikipedia articles. This feature is explained below.

A higher percentage of items with matching references and domains would suggest a more direct inter-dependence between the two projects. Regarding the correlation between the number of common domains and common authors, Wikidata and Wikipedia belong to the same Wikimedia ecosystem and their communities partially overlap (see section 2.2). Hence, sources from the same domain in corresponding items and articles across the two platforms might be added by the same user. High positive correlation between common domains and common authors would confirm this hypothesis and suggest the existence of users specialised in determined topics across Wikidata and Wikipedia. Matching domains were chosen instead of matching references, as different pages within a domain can cover various aspects of the same topic. To check for common editors, we relied on the 'Unified login' features, which allows Wikimedia editors to have the same username across different projects[7]. We verified whether users who edited a Wikidata item were active also on the corresponding Wikipedia articles. We then counted the number of matching users and tested for correlation between this figure and the number of matching domains.

RQ2.A is intended to shed light on the connections between Wikidata and different language versions of Wikipedia. In order to do that, we analysed the number of matches per language version of Wikipedia:

- *percentage of matching references* between Wikidata items and corresponding articles by Wikipedia language version, over the total number of matching references;
- *percentage of matching domains* between Wikidata items and corresponding articles by Wikipedia language version, over the total number of matching domains;
- *percentage of single matching references* between Wikidata items and corresponding articles by Wikipedia language version, over the total number of matching references. Compared to the previous features, which take into account several matches for each reference or domain, this and the following feature count only unique matching references and domains;
- *percentage of single matching domains* between Wikidata items and corresponding articles by Wikipedia language version, over the total number of matching domains.

Wikipedia language versions differ greatly in size. The English version is by far the largest, with several million articles; others instead contain only a small amount of material. To counterweight this skew, we disregarded URLs that appeared in more than one Wikipedia version. The rationale for this approach was to increase the visibility of smaller Wikipedias and highlight possible common provenance patterns between them and Wikidata.

Finally, in order to tackle RQ2.B we classified all Wikidata items in our dataset by using their taxonomic information, i.e. their *instance of* (P31) and *subclass of* (P279) relationships. The classification used follows the one generated by the Wikidata community in [21]. Although this is a very broad and high-level classification, it has the advantage to be already used in Wikidata, thus being commensurable with an already existing classification of the whole knowledge graph. The variables used to address RQ2.B were chosen with the aim to show the type of items which are the closest to their Wikipedia counterparts in terms of information sources:

- *percentage of items with matching references per class*;
- *percentage of items with matching domains per class.*

DATA

We extracted all Wikidata items with at least one reference by issuing a query to the Wikidata SPARQL query endpoint[8]. References pointing to pages in Wikipedia were left out, as these are not authoritative sources according to the verifiability policy. The query was executed by one of the developers of the Wikidata development team at Wikimedia Deutschland, as it required longer than the maximum time allowed to external users. The results of the query included the IDs of the items matching the query, the property used in the statement containing the reference, and the reference URL. Another query was used to obtain all the sitelinks, i.e. links to corresponding pages on other Wikimedia projects, for each of the items from

[7] See https://meta.wikimedia.org/wiki/Help:Unified_login, consulted on 18 April 2017.

[8] https://query.wikidata.org/, consulted on 16 April 2017.

the results of the first query. Both queries are updated to the day of their execution, i.e. 11 March 2017. The total number of items obtained was $1,480,744$, which were linked to a total of $6,239,219$ external sources. Since we relied on item sitelinks to find corresponding Wikipedia pages, we excluded from our dataset all the items which had none, leaving us with $616,528$ items. For each item in this sample we downloaded the text of the corresponding Wikipedia articles in all the language versions available. Due to the time required to download all the pages, items and Wikipedia pages may be updated to different points in time, with a maximum difference of four days. However, from official Wikimedia statistics[9] it can be estimated that the Wikipedia daily edit rate is 0.02 edits per page: only 2-8% of pages are potentially affected by changes between the end date of our Wikidata dataset and the date in which they were downloaded. Afterwards, we parsed the text of all the extracted Wikipedia article links to obtain the external references, by matching a regular expression to the tags that are used to mark Wikipedia citations (`<ref>...</refs>`). $409,790$ Wikidata Items had corresponding Wikipedia articles which had any reference. All the data extracted, the code used for the analysis, and links to interactive versions of the graphs have been made freely available on GitHub[10].

RESULTS

This section presents a comparison between Wikidata and Wikipedia use of external sources (RQ1). Moreover, it addresses the reuse of references across the two projects (RQ2) and covers the relationship between Wikidata and different language versions of Wikipedia (RQ2.A). Finally, it shows how items with matching references are distributed across Wikidata classes (RQ2.B).

Wikidata vs. Wikipedia References

The Wikidata items included in our analysis had on average a lower number of external references per item than Wikipedia articles, both across all languages and for each of the largest five Wikipedias per number of editors (see Table 1). Sources in Wikidata serve as references for multiple statements, to a larger extent than in Wikipedia, as they are used on 2.8 times on average. The difference increases looking at each of the single language versions in Table 1. Additionally, Wikidata presents a less diverse range of sources, as the high ratio between unique references and unique domains shows.

Concerning source types, we looked at the most commonly found domains in each of the two projects. The distribution of Wikidata domains is highly skewed, as the top ten account for around 70% of total uses (Figure 2(a)). The most common domains are reference resources or databases (Figure 3), with a smaller presence of news outlets. On the other hand, a large number of the most cited sources in Wikipedia are from news and media outlets. Four out of the first ten most common domains (Figure 2(b)) are either from newspapers or broadcasters. Nevertheless, none of the domains has a much higher percentage of uses than the rest, as the most common one accounts for only 1.6% of all sources. With regard to the geographic distribution of sources, a small number of TLDs accounts for the vast majority of references in both platforms. Wikidata's top ten TLDs cover around 90% of all references. This percentage is still very high in Wikipedia, where the ten most common TLDs are used for 79% of all citations. Spearman correlation between the proportion of each TLD in the two datasets was very high ($\rho = .87$), showing a similar geographical distribution of sources overall. However, looking at the most common TLDs in both projects might suggest a different picture. Wikidata includes several non-English speaking TLDs among its top ten ones (Figure 4). On the contrary, Wikipedia is more US– and UK-centric. Considering only the top ten TLDs in the encyclopedia, the domains `com`, `org`, `uk`, `gov`, and `edu` are 67.8% of the total, compared to 48.4% of all Wikidata sources.

Sources Reuse across Wikidata and Wikipedia

Regarding the number of matching pages and domains (**RQ2**), items and articles seldom share the same sources ($4,189$; 0.85%). Around one fourth of the items in our dataset had one or more domains in common with their equivalent in Wikipedia and about half of the domains in Wikidata items were also in their Wikipedia counterparts (see Table 2). The most common matching domains were somewhat surprising, as only one (IMDb) was among the top ten most used sites in both projects and four were in either of them, but not in both (Figure 3). Finally, we found a low positive correlation between the number of matching domains and the number of common users between each Wikidata item-Wikipedia article couple ($r = .157$ and $p = 0$).

Variations across Wikipedia Language Versions

RQ2.A tackles the issue of the relationship between Wikidata and different language versions of Wikipedia. Considering all the times that a source was used in corresponding items/articles, the Southern Min Wikipedia, a variety of Chinese, is by far the version with the largest percentage of matching references and domains. Nevertheless, an analysis of the single matching pages, i.e. counting each page only once, this percentage lowers to less than 1%. This is explained by the activity of a bot, which automatically added a link from the US census site as a citation for several US cities. The same link is broadly used also in Wikidata. The rank of the Wikipedias with the highest percentages of common references and domains to Wikidata shows that the English one, unsurprisingly, is the one with the most matching. However, not all the Wikipedias in the top ten (Figure 5) for matching references are among the first ten language versions with most editors or edits[23].

Variations across Wikidata Classes

The class distribution of items which share sources with their Wikipedia counterpart differs from the overall distribution in Wikidata (**RQ2.B**, Figure 7). In particular, the categories *human*, *administrative territorial entity*, and *film* are overrepresented among the items with matching domains.

DISCUSSION

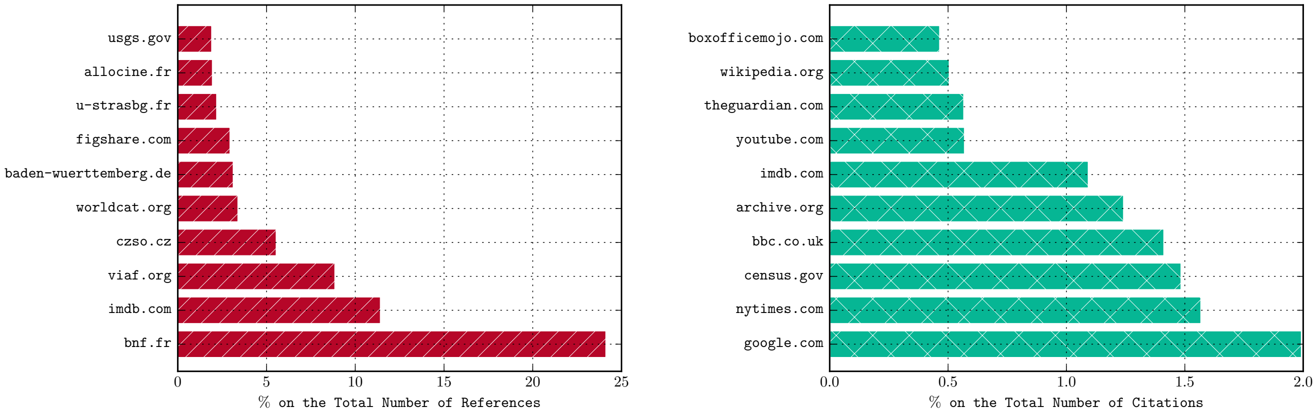
In the current study we explored the relationship between Wikidata and Wikipedia under the point of view of the information

[9] `https://stats.wikimedia.org/EN/ReportCardTopWikis.htm`, consulted on 10 July 2017.

[10] `https://github.com/pvougiou/Wikidata-Referencing`.

Feature	WD	WP all	WP en	WP de	WP fr	WP es	WP ja
Number of Items/Articles	409,790	1,324,413	247,423	107,620	91,662	56,636	38,181
Avg. No. of References per Item/Article (σ)	3.4 (±6.2)	7.5 (±19.9)	11.2 (±26.2)	6.8 (±14.1)	8.1 (±20.9)	8.8 (±23.7)	10 (±24.3)
Median No. of References per Item/Article	2	3	4	4	3	4	4
Avg. No. of Uses per Reference (σ)	2.8 (±64)	2.3 (±19.3)	1.8 (±6.6)	1.8 (±6.1)	1.7 (±9.7)	1.7 (±1.6)	1.8 (±2.1)
Median No. of Uses per Reference	1	2	2	2	2	2	2
No. of Unique References	492,793	1,324,413	297,021	107,620	96,606	67,497	51,108
No. of Unique Domains	22,745	464,701	219,965	84,221	72,748	52,246	40,197
Unique References to Unique Domains ratio	21.7	2.8	1.1	1.3	1.3	1.1	0.9

Table 1. Statistics about use of references in the analysed samples of Wikidata (WD) and Wikipedia (WP). Concerning the latter, we included figures about the whole encyclopedia and by language version, for the five languages with the largest number of editors, i.e. English (en), German (de), French (fr), Spanish (es), Japanese (ja).



(a) Ten most common domains in Wikidata references.

(b) Ten most common domains in all Wikipedias citations.

Figure 2. Top ten domains in Wikidata (left) cover a much higher share of the total, compared to Wikipedia.

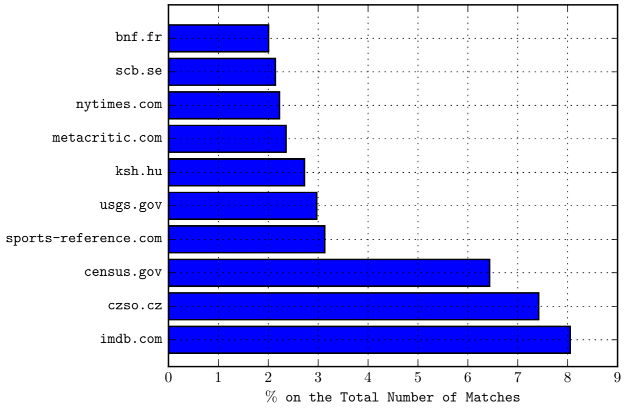


Figure 3. Ten most common matching domains between Wikidata items and Wikipedia articles. Only one domain is in the top ten of both projects' most used domains.

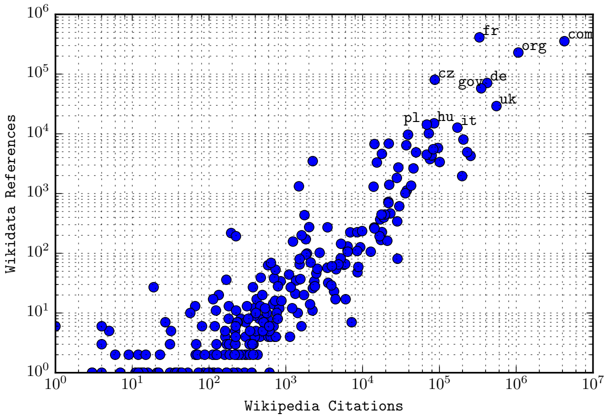


Figure 4. Top-level domains (log) in Wikidata and Wikipedia references.

sources they rely upon. The lower number of references per item in Wikidata was partially surprising. We analysed only Wikidata external references, leaving out internal ones. This might be one of the reasons for the difference. Furthermore, Wikidata is still in an early stage of development, compared to Wikipedia. Nevertheless, Wikidata's verifiability policy prescribes that virtually every claim must be supported by a source [22], whereas the free encyclopedia requires citations only for challenging statements [24]. Further research should focus on the extent to which references are used in Wikidata, by addressing internal references and carrying out an examination at statement level.

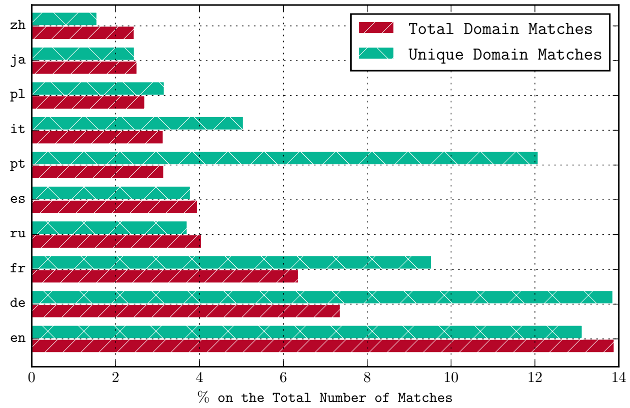


Figure 5. Domain matches between Wikidata items and corresponding Wikipedia articles, by Wikipedia language version. Unique domain matches mean that each domain in common between the two projects is counted only once.

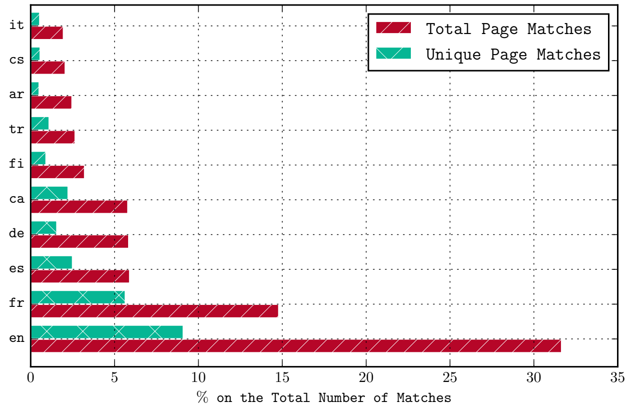


Figure 6. Single web page matches between Wikidata items and corresponding Wikipedia articles, by Wikipedia language version. Unique page matches mean that each web page in common between the two projects is counted only once.

The most frequently used domains differ between Wikidata and Wikipedia, both with regard to their type and their distribution. In Wikidata, databases (e.g. https://simbad.u-strasbg.fr/simbad/) and catalogues (e.g. http://data.bnf.fr/) form the bulk of the external sources. On the other hand, news domains play a stronger role among the sources over the whole of Wikipedia, in line with what was noted in [2]. This can be explained by the inherent characteristics of the two projects. Wikidata's *subject-property-object* triples convey information in a punctual, "atomic" fashion, such as *Charlie Chaplin – place of birth – London*. In contrast, the articles of an encyclopedia express information in a more elaborated, nuanced way, which may need different types of references. Moreover, Wikidata is already connected to several resources in the Linked Open Data cloud. Therefore, links to identifiers in external catalogues are used to disambiguate entities across knowledge bases. The same reasons may be behind the divergences between the class distribution of the items with matching domains and Wikidata as a whole. Classes that are over-represented among matching items may be those for which the type of sources required are more similar. An example is the category of films, which has around twice the percentage of items among the ones with common sources, compared to all Wikidata. This connects well with IMDb being the domain with the highest number of matches between the two projects. Moreover, according to Wikidata and Wikipedia verifiability policies, IMDb is not an authoritative source as it is authored by a community of users.

Metric	Value
Items with Matching References	1.71%
Items with Matching Domains	24.49%
Matching References	0.85%
Matching Domains	50.76%
Avg. Matches per Reference (σ)	9.7 (±377.8)
Avg. Matches per Domain (σ)	150.8 (±4321.1)

Table 2. Percentages of items with references or reference domains matching the ones used by their Wikipedia article counterparts and percentages of matching references and reference domains. Whereas the number of exact link matches is low, a large number of domains appears to be reused in the two projects.

With regard to the distribution of domains, the top ten account for around 70% of all sources. We may related this to the high share of bot activity [16], which adds large numbers of references to the same resource automatically. In terms of practical implications of this finding, Wikidata may be required to diversify its sources with respect to the domains they belong to. This could be achieved by a stronger involvement of various Wikipedia language communities. These may in turn benefit from such involvement, as Wikidata may function as a hub where information spreads by osmosis from one language community to another. Furthermore, an understanding of which item classes have more sources in common may help different communities become aware of the areas in which their contribution to Wikidata is more needed.

Another interesting aspect is the low correlation between the number of item-article sources' matches and their number of common editors. This finding may be a hint that matching domains could be due to the popularity of certain domains, rather than to an active engagement of users in adding the sources about the same topic across Wikimedia projects. This is an issue that is worth investigating, in order to comprehend whether popular sources can polarise some topics and reduce the diversity of points of view about them. Moreover, it would be interesting to understand whether the Wikidata and Wikipedia communities are aware of this partial identity of information sources.

While our analysis showed a high positive correlation between percentage of TLD uses in Wikidata and Wikipedia, an observation of the most commonly used domains showed that US and UK sources represent a much larger portion in the latter. This issue has already been known to affect some parts of the free encyclopedia according to [10]. Being a multilingual project, able to represent different cultural points of view is one of the outstanding goals of Wikidata [19]. Do our findings mean that Wikidata has achieved this goal? The mea-

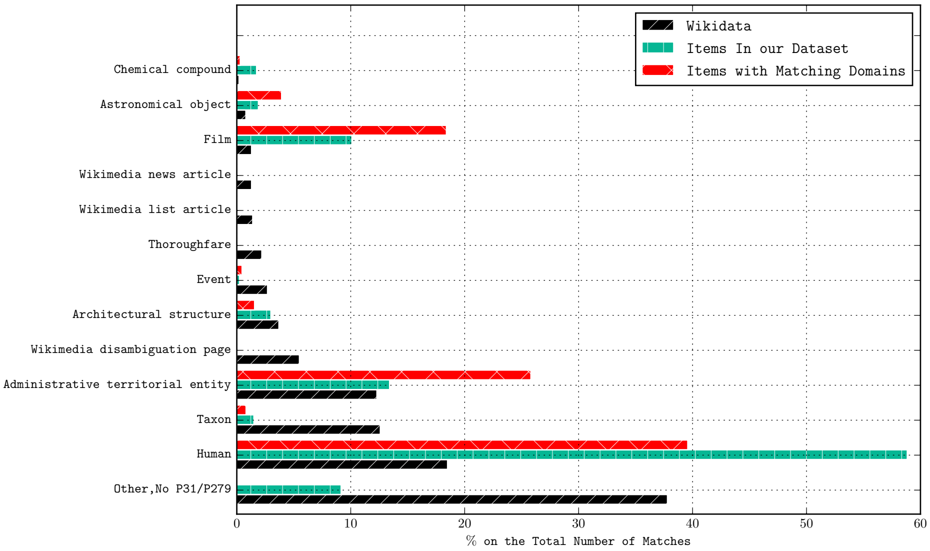


Figure 7. Distribution of item classes over the whole of Wikidata, in our dataset, and in items with matching domains.

sures produced in this work allows us only to get a glimpse into this issue. A definitive answer about that should be the aim of future research investigating Wikidata cultural diversity and multilinguality. Regarding the relationship between Wikidata and different language versions of Wikipedia, the patterns revealed by our analysis do not differ significantly from what could be expected by looking at the relative size of the various Wikipedias. Nonetheless, some languages perform better, compared to their relative size. The community dynamics behind this should be investigated in the future. Future work around the topic should also focus on the possibility to track user activity across different Wikimedia projects, to identify common editing patterns in Wikidata and in different Wikipedias.

LIMITATIONS AND CONCLUSION

Due to the breadth of the subject and its novelty, we faced some limitations with our approach. Notwithstanding the affinities between Wikidata and Wikipedia, the two projects have profound differences. As we have seen, this entails different needs in terms of external sources, but also divergent strategies with regard to the type and structure of the information contained. A comparison of these aspects would be key to a contrasting study of the two project, but it was out of the scope of this work. A second limitation is related to the dataset employed. We conducted our analysis by considering only the external sources of Wikidata, but these are just one of the types of references allowed in the knowledge graph. Including Wikidata internal sources would have introduced to this work a further level of complexity, which should be covered in future work.

This work aimed to shed light on the reuse of web sources between Wikidata and Wikipedia. We analysed what type of citations are employed in each of the two projects and how. The differences observed between the types of sources found in the two projects could be explained by their respective peculiarities. External databases and reference sources, such as library catalogues, are dominant among Wikidata sources. On the contrary, Wikipedia uses a large number of news citations, which may be unfit to provide the punctual, 'atomic' type of information to which a knowledge graph needs to link. On the other hand, the high number of common sources, intended as web domains, suggests that a big part of the information available from the two platforms may reflect similar points of view. How these are close may be subject of future investigation.

Finally, sources in Wikidata seem to be less Anglo-American centric than in Wikipedia. This might be a sign that Wikidata is on the right track to achieve its goal to actively promote diversification of knowledge across several languages, challenging any cultural monopoly of information. Our study is a starting point for further research aiming to evaluate cultural diversity and the circulation of information between these two fascinating projects.

ACKNOWLEDGEMENT

This project is supported by funding received from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 642795 (WDAqua ITN).

We thank the Wikidata development team at Wikimedia Deutschland for the support received in retrieving the data.

REFERENCES

1. Michael Färber, Basil Ell, Carsten Menne, and Achim Rettinger. 2015. A Comparative Survey of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web Journal, July* (2015).

2. Besnik Fetahu, Katja Markert, Wolfgang Nejdl, and Avishek Anand. 2016. Finding News Citations for Wikipedia. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*. ACM, 337–346.

3. Fabian Flöck, Denny Vrandecic, and Elena Simperl. 2011. Towards a diversity-minded Wikipedia. In *Web Science 2011, WebSci '11, Koblenz, Germany - June 15 - 17, 2011*. 5:1–5:8.

4. Heather Ford, Shilad Sen, David R. Musicant, and Nathaniel Miller. 2013. Getting to the source: where does Wikipedia get its information from?. In *Proceedings of the 9th International Symposium on Open Collaboration, Hong Kong, China, August 05 - 07, 2013*, Ademar Aguiar and Dirk Riehle (Eds.). ACM, 9:1–9:10.

5. Paul T. Groth. 2013. The Knowledge-Remixing Bottleneck. *IEEE Intelligent Systems* 28, 5 (2013), 44–48.

6. Isto Huvila. 2010. Where does the information come from? Information source use patterns in Wikipedia. *Inf. Res.* 15, 3 (2010).

7. Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the 2007 Conference on Human Factors in Computing Systems, CHI 2007, San Jose, California, USA, April 28 - May 3, 2007*, Mary Beth Rosson and David J. Gilmore (Eds.). ACM, 453–462.

8. Martin Körner, Tatiana Sennikova, Florian Windhäuser, Claudia Wagner, and Fabian Flöck. 2016. Wikiwhere: An interactive tool for studying the geographical provenance of Wikipedia references. *CoRR* abs/1612.00985 (2016).

9. Markus Krötzsch, Sebastian Schaffert, and Denny Vrandečić. 2007. Reasoning in semantic wikis. In *Reasoning Web*. Springer, 310–329.

10. Brendan Luyt and Daniel Tan. 2010. Improving Wikipedia's credibility: References and citations in a sample of history articles. *JASIST* 61, 4 (2010), 715–722.

11. Finn Årup Nielsen. 2007. Scientific citations in Wikipedia. *First Monday* 12, 8 (2007).

12. Yelena Pancheshnikov. 2007. A Comparison of Literature Citations in Faculty Publications and Student Theses as Indicators of Collection Use and a Background for Collection Management at a University Library. *The Journal of Academic Librarianship* 33, 6 (2007), 674 – 683.

13. Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web* 8, 3 (2017), 489–508.

14. Alessandro Piscopo, Christopher Phethean, and Elena Simperl. 2017. Wikidatians are Born: Paths to Full Participation in a Collaborative Structured Knowledge Base. In *50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*. AIS Electronic Library (AISeL).

15. Richard Rogers. 2009. *The end of the virtual: Digital methods*. Vol. 339. Amsterdam University Press.

16. Thomas Steiner. 2014. Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux): A Global Study of Edit Activity on Wikipedia and Wikidata. In *Proceedings of The International Symposium on Open Collaboration, OpenSym 2014, Berlin, Germany, August 27 - 29, 2014*. ACM, 25:1–25:7.

17. Thomas Pellissier Tanon, Denny Vrandecic, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From Freebase to Wikidata: The Great Migration. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*. ACM, 1419–1428.

18. Denny Vrandecic. 2013. The Rise of Wikidata. *IEEE Intelligent Systems* 28, 4 (2013), 90–95.

19. Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

20. Wikidata. 2017a. Wikidata:Sources — Wikidata, the free knowledge base. `https://www.wikidata.org/wiki/Help:Sources`. (2017). [Online; accessed 09-April-2017].

21. Wikidata. 2017b. Wikidata:Statistics — Wikidata, the free knowledge base. `https://www.wikidata.org/wiki/Wikidata:Statistics`. (2017). [Online; accessed 09-April-2017].

22. Wikidata. 2017c. Wikidata:Verifiability — Wikidata, the free knowledge base. `https://www.wikidata.org/wiki/Wikidata:Verifiability`. (2017). [Online; accessed 07-April-2017].

23. Wikipedia. 2017a. List of Wikipedias — Wikipedia, The Free Encyclopedia. `http://en.wikipedia.org/w/index.php?title=List%20of%20Wikipedias&oldid=773693902`. (2017). [Online; accessed 09-April-2017].

24. Wikipedia. 2017b. Wikipedia:Verifiability — Wikipedia, The Free Encyclopedia. `http://en.wikipedia.org/w/index.php?title=Wikipedia%3AVerifiability&oldid=775109734`. (2017). [Online; accessed 22-April-2017].

25. T. D. Wilson. 1994. Information needs and uses: fifty years of progress. Information seeking behaviour, Behavior, Information needs, Information use. In *Fifty years of information progress: a Journal of Documentation review*, B. C. Vickery (Ed.). Aslib, London, 15–51.