Provenance Information in a Collaborative Knowledge Graph: an Evaluation of Wikidata External References

Alessandro Piscopo, Lucie-Aimée Kaffee, Chris Phethean, and Elena Simperl

University of Southampton,
Southampton, United Kingdom
{A.Piscopo, Kaffee, C.J.Phethean, E.Simperl}@soton.ac.uk

Abstract. Wikidata is a collaboratively-edited knowledge graph; it expresses knowledge in the form of subject-property-value triples, which can be enhanced with references to add provenance information. Understanding the quality of Wikidata is key to its widespread adoption as a knowledge resource. We analyse one aspect of Wikidata quality, provenance, in terms of relevance and authoritativeness of its external references. We follow a two-staged approach. First, we perform a crowdsourced evaluation of references. Second, we use the judgements collected in the first stage to train a machine learning model to predict reference quality on a large-scale. The features chosen for the models were related to reference editing and the semantics of the triples they referred to. 61% of the references evaluated were relevant and authoritative. Bad references were often links that changed and either stopped working or pointed to other pages. The machine learning models outperformed the baseline and were able to accurately predict non-relevant and non-authoritative references. Further work should focus on implementing our approach in Wikidata to help editors find bad references.

Keywords: Wikidata, provenance, collaborative knowledge graph

1 Introduction

Wikidata is a collaborative knowledge graph started in 2012 by the Wikimedia foundation. It supplies data to other Wikimedia projects (including Wikipedia), as well as anyone else who wants to use it, under a public license. Wikidata already has a broad coverage, with data covering more than 24M abstract and concrete entities, gathered by a user pool of around 17,000 monthly active users. This data has already been encoded in RDF and connected to the Linked Data Web [7]. All these features have drawn the attention of researchers and practitioners alike. Following its elder sister Wikipedia, Wikidata requires all information to be verifiable, but goes a step further. It is a secondary database and as such its aim is not to state facts about the world, but to report claims from primary sources [26]. Each claim must be supported by a source and linked to it.

While most work around Wikidata focuses on the quality of its triples [4] or analyses its community dynamics [20], to the best of our knowledge no studies have investigated provenance quality. Little is known about the quality of the sources included to support claims, although this is a key issue for Wikidata. Provenance facilitates the reuse of data by improving error-detection and decision-processes based on the information source [16]. The lack of provenance information or the use of poor sources may affect its trustworthiness and hinder the reuse of its data for business and other purposes [12]. Additionally, the availability of provenance information can increase trust in the project, as noted in Wikipedia [17]. On a practical side, a method to detect bad external sources would support editors in maintaining Wikidata knowledge graph.

This paper proposes an approach to analyse quality of provenance information in Wikidata. We define quality in terms of relevance and authoritativeness of the external sources used in Wikidata references. To assess these, we use two complementary methods: microtask crowdsourcing and machine learning. Crowdsourcing is used to collect judgements about relevance and authoritativeness of sources. These judgements are successively utilised to train machine learning models to predict problematic references on a large-scale.

2 Background and Related Work

Wikidata consists of *items*—representing concrete (such as the Colosseum) or abstract things (e.g. humans)—and *properties* which express relationships between items or between items and values. Items and properties are identified by URIs, starting respectively with 'Q' or 'P'. Relationships are encoded via *claims*, which can be enriched by adding qualifiers (to provide contextual information) and/or references to form a *statement*. Statements are maintained by the Wikidata community. Beyond human editors, who can be registered or contribute anonymously, pieces of software called *bots* carry out a broad range of tasks, adding and maintaining content. Whereas human editors are the main contributors to the conceptual structure of Wikidata [18], bots perform more when it comes to adding and modifying content and can often add large batches of statements in one go [24]. No study has examined yet the differences between bot and human contributions in Wikidata in terms of quality. This would be relevant, considering the difference between the contribution patterns of these two user types. Bots author the majority of revisions in Wikidata and the sources on which their edits are based belong to a predetermined set of domains, thus they focus on fewer types of statements compared to human editors. We analysed external reference quality on Wikidata overall and separately by bot or human editor to provide insights into the outcome of the work of each user type.

2.1 Provenance in Wikidata

Provenance information may be either recorded at the moment of data creation (*eager* approach) or computed upon request (*lazy* approach) [11]. Wikidata adopts the former approach and editors are asked to add sources to the

statements that they create. Provenance in Wikidata can be added by enriching claims with references. Most types of statements require a reference, otherwise they are deemed unverified and should be removed [28]. However, community-generated policies define some statement types that are exempt from this rule, such as undisputed claims (e.g. *Earth*, *instance of*, *planet*) [27]. The sources used as references can either be internal (an item already in Wikidata), or external, i.e. linking to a URL [27]. Statements that are not exempt must be verifiable by consulting a referenceable primary source. This must be **accessible** 'by at least some' Wikidata contributors to confirm the source firsthand [28]. A good reference must also be **relevant**—it must provide evidence for the claim it is linked to. Additionally, good references must be **authoritative** or 'deemed trustworthy, up-to-date, and free of bias for supporting a particular statement' [28].

2.2 Authoritativeness in Wikidata

Wikidata defines authoritative sources by describing suitable types of publisher and author. This is also the approach of Wikipedia, whose policy Wikidata redirects to. Specifically, the term 'source' has three meanings in Wikipedia [29]: the *type of work* itself, the *author* of the work, and the *publisher* of the work. Wikidata's policy specifies types of sources that are authoritative: books; academic, scientific and industry publications; policy and legislation documents; news and media sources. These must have a corresponding entity in Wikidata, linked to claims through property P248 (*stated in*). Databases and web pages may also be authoritative. Databases require a corresponding property already defined in the knowledge graph, pointing to an entry in the database. Authoritativeness of web pages, referenced through property P854 (*reference URL*), depends on their author and publisher type. Authors may be *individuals* (one or more identifiable persons), *organisations*, or *collective* (a number of individuals who often utilise a username and whose contribution is voluntary). Sources whose author is unknown should be avoided, as well as user-generated sources, e.g. forums or social review sites. Regarding publishers, sources with no editorial oversight and relying on rumours and personal opinions are not generally considered authoritative. Authoritative publishers are government agencies, companies and organisations, and academic institutions [28]. Self-published sources are generally not accepted, nor are websites with promotional purposes or those affected by political, financial, or religious bias. Wikipedia pages are not good references because they are not primary sources and are collectively created. Table 1 shows publisher types. Combinations of author and publisher types are in Table 2.

2.3 Evaluating Provenance

Literature about authoritativeness on the Web can be roughly divided into two approaches. The first uses automated methods to analyse the hyperlinked structure of the Web to generate link-based metrics to gauge the authoritativeness of pages. As an example, in [15] authority measures are generated using inter-links within sub-graphs of the Web. A similar approach is followed by works that

Table 1. Types of publisher in the classification used. On the right column, sub-types or, when these are missing, definitions of higher-level types.

Academic and scientific organisations	*Academic and research institutions* (e.g. universities and research centres, but not museums and libraries); *Academic publishers*; *Other academic organisations*
Companies or organisations	*Vendors and e-commerce companies*; *Political or religious organisations*; *Cultural institutions*; *Other types of company*
Government agencies	Any governmental institution, national or supranational
News and media outlets	*Traditional news and media* (e.g. news agencies, broadcasters); *Non-traditional news and media* (e.g. online magazines, platforms to collaboratively create news)
Self-published sources	Any sources that does not belong to any organisation/company, maintained by authors themselves

Table 2. Authoritativeness of sources (ticks indicate authoritative)

Author / **Publisher**	**Individual**	**Organisation**	**Collective**
Academic and research institution	✓	✓	✗
Academic publisher	✓	✓	✗
Other academic	✓	✓	✗
Government agency	✓	✓	✗
Vendor or e-commerce company	✗	✗	✗
Political or religious organisation	✗	✗	✗
Cultural institution	✓	✓	✗
Other type of company	✓	✓	✗
Traditional news and media	✓	✓	✗
Non-traditional news and media	✓	✗	✗
Self-published source	✗	✗	✗

investigate automatic source retrieval. DeFacto [16] uses machine learning and NLP to produce scores about the likelihood of a web page to contain specific pieces of information and about its trustworthiness. Fetahu *et al.* also apply machine learning to assess web pages and find sources that are authoritative and relevant for statements within Wikipedia articles [8]. Other methods focus on evaluating provenance through similarity and distance metrics computed across different databases [5]. These models did not apply to Wikidata as this quantitative approach differs from the focus on principles such as type, author, and publisher that Wikidata follows. Furthermore, Wikidata external sources have diverse formats including web pages, PDFs, or csv files, which may be problematic to evaluate for completely automated systems such as [8] or [16]. DeFacto's measure of trustworthiness would need extensive testing in order to understand how it matches the definition of authoritativeness used by Wikidata.

The second group of approaches, followed by Wikidata, manually identifies principles to define credible and authoritative web sources. A small sample of

Wikipedia citations have been evaluated by analysing their author, publisher, and document types in [9]. Crowdsourcing has been used to evaluate page relevance with faster completion times compared to expert-run experiments or online surveys, and low cost, whilst yielding high quality results [2].

3 Methods

We developed an approach that evaluates Wikidata references in terms of relevance and authoritativeness. We aimed to carry out a large-scale scale evaluation of Wikidata provenance, and adopted a two-staged approach relying on two complementary methods: microtask crowdsourcing and machine learning. Because of the advantages outlined above, we performed a crowdsourced evaluation of references, which was used to train a machine learning model to predict their quality. Machine learning can be easily applied on a large-scale and is virtually costless. We evaluated only external references, which were 6% of the total. In our analysis, we distinguished between bots and people because of their different roles in maintaining the knowledge graph. We posed these research questions:

RQ1 To what extent are Wikidata external references relevant?
RQ2 To what extent are Wikidata external reference authoritative?
RQ3 To what extent can non-relevant and non-authoritative references be predicted in Wikidata?

3.1 Source Evaluation

We designed three crowdsourcing tasks to assess reference quality, which were carried out on CrowdFlower [1]. All tasks included one type of microtask, except one, which included two. In order to increase the clarity of microtasks, we refined their design by launching test runs of small samples (between 50–100) of references to be evaluated. User behaviour (number of missed questions and completion time) was observed to understand microtask clarity.

Relevance. The first task (**T1**) was designed to assess relevance by asking users to find the pieces of information composing a statement within its source. Each microtask in T1 evaluated a reference, i.e. a statement with its attached source. In order to decrease the cognitive burden on workers, we structured microtasks along three questions, one for each element of a statement (subject, property, object). For each of these, we asked whether the source provided information about it. Users were prompted the successive questions only if they responded positively to the prior one (we asked about the property of a statement only if evidence about its subject was found in the source). English labels were shown for each statement's part, instead of their URIs. In the case of pages not working or requiring a log in, or for pages not in English, users could select the appropriate responses. Figure 1 illustrates an example of T1 microtask.

[1] https://www.crowdflower.com/.

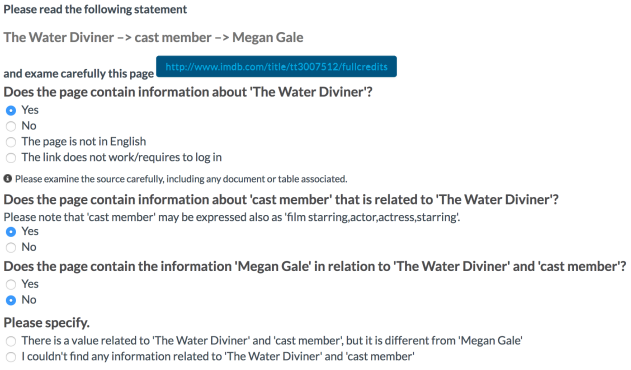
Authoritativeness. A similar concept to authoritativeness—credibility—is consistently assessed under a positive bias by web users [13]. Hence, instead of directly questioning users about the authoritativeness of a source, which would have likely given overly subjective responses, we tested whether sources matched the types specified by Wikidata policy and asked the crowd to classify them, similar to the approach followed for Wikipedia in [9].

Author type was assessed in **T2**. Microtasks in T2 asked users to indicate the most appropriate author type for a source. Users were shown only the source, rather than the whole reference. Therefore, T2 included only unique web pages.

Task 3 (T3) evaluated publisher type. We assumed that pages belonging to the same domain had the same publisher. Hence, we collected judgements for unique domains, rather than for each single reference. **T3.A** included only higher-level types of publisher: *academic and scientific organisations*, *companies and organisations*, *government institutions*, *news and media*, and *self-published sources*. It consisted of a multiple choice question to select the most appropriate type of publisher. **T3.B** collected judgements related to the sub-types in Table 1. In T3.B users were asked whether the publisher type obtained from the previous task was appropriate for the source, in order to test contributors' performance and verify the results of T3.A. If users answered positively, they were asked to classify the sub-type of the source publisher. User pools of T3.A and T3.B were independent from each other. Responses for pages not working or requiring log in, or pages not in English were included in T2, T3.A, and T3.B.

Quality Assurance. Crowdsourcing is vulnerable to users who perform poorly due to lack of skills, malicious behaviour, or distraction [6]. We adopted various strategies to tackle this issue. We added gold standard questions to tasks and excluded workers whose performance fell under a certain threshold, which we set to 80% in all tasks. Tasks were structured in pages, each containing a number of microtasks varying according to the task. Workers were first required to pass a test consisting of a page of test questions with an accuracy above or equal to the threshold set. Additionally, a test question was included in each page of work. Users had to keep an accuracy above the minimum threshold throughout their

Fig. 1. A HIT from T1



contribution. We followed previous research regarding the experimental design of workers' qualification, granularity of task, and monetary rewards (see Table 3). Based on observations collected during test runs of the tasks, we accepted only workers with a previous accuracy rate of 85%—the highest allowed by CrowdFlower[2]—to select highly performing users [6]. Payments per microtask were determined according to [23]. Correct answers were selected by majority voting over five assignments per microtask, following [1]. Information on how to complete the task and links to clarifying examples[3] were available on each page.

Table 3. Crowdsourcing experiment design

	T1	**T2**	**T3.A**	**T3.B**
Worker qualification	$\geq 85\%$	$\geq 85\%$	$\geq 85\%$	$\geq 85\%$
Granularity (microtasks per page)	10	8	8	8
Monetary reward (per microtasks)	\$0.08	\$0.06	\$0.05	\$0.05
Assignments	5	5	5	5
Min. worker accuracy	80%	80%	80%	80%

3.2 Automatic Evaluation Model

We used a machine learning classifier to identify not relevant or not authoritative sources. We trained a supervised algorithm for each outcome variable, using the labels obtained through the crowdsourcing experiment. Both relevance and authoritativeness models included features concerning the source itself and the semantics and editing activity of the statement it referred to. We assumed more frequently used sources are more likely to be checked by several users and therefore to be trusted. Regarding statement semantics, the rationale was that if a reference is good for a statement, it might be good for similar statements as well. Activity metrics were added as users with a larger number of edits may be more trustworthy, according to previous findings [21]. We included the same features in both models, as they could contribute to various extents to their accuracy.
URL reference uses. Number of times a URL has been used as a reference.
Domain reference uses. Number of times a domain has been used.
Source HTTP code. HTTP response code given by the source link.
Statement property. The property used in the statement.
Statement item. The item subject of the statement, represented as a vector of its structured components, i.e. labels and aliases were excluded.
Statement object. The object of the statement represented as a vector.
Subject parent class. Item parent class, i.e. object of property P279 (*subclass of*) or P31 (*instance of*).
Property parent class. Property parent class, i.e. object of P279 or P31.
Object parent class. Item parent class, i.e. object of P279 or P31.

[2] http://crowdflowercommunity.tumblr.com/post/108559336035/new-performance-level-badge-requirements.
[3] Examples were provided for T2, T3.A, and T3.B: https://wdref-author-evaluation.000webhostapp.com/, https://wdref-evaluation.000webhostapp.com/.

Author type. Anonymous, bot, or registered human.
Author activity. Total number of revisions carried out by the reference creator, prior to adding it.
Author activity concerning references. Proportion between number of reference edits and total number of edits carried out by the author of the reference. Editors who are more active on references are more likely to add good sources.

We tested three different algorithms that previously performed well for different tasks, Naive Bayes, SVM, and Random Forests. Models were trained using the python library scikit-learn [19].

4 Evaluation

4.1 Data

Wikidata Corpus. We used the complete edit history of Wikidata updated to the 1st October 2016. We extracted all statements containing external references, excluding those pointing to a Wikimedia link and those not requiring any reference, according to Wikidata policy [27]. This gave $1,629,102$ references, of which $1,449,295$ pointed to two domains (`uniprot.org` and `ebi.ac.uk`). Around 98% of these were added by one bot and each of their domains was successively assigned a database property, therefore we removed them from the sample. Only references in English were selected; we dropped all references whose source did not have an international top-level domain or one from an English-speaking country[4]. $83,215$ references remained, from which we extracted 2586 (99% confidence level, 2.5% margin of error; further details in Table 4). We wanted our sample to reflect the different subject-object relations supported by references. Therefore, the sample was drawn in order to reflect the proportion of property uses from the larger dataset. We automatically tested the validity of each link by querying its HTTP code with the python library `requests`. Pages that returned a 404 code or timed out were flagged as not working and not submitted to the crowd. One link[5], used in several references (512, 19.8%), redirected to another page which did not contain the data initially hosted by the link and was judged as not relevant. Two more links[6] (282 uses, 11% of the total) pointed to csv files that were automatically checked. Both links were classified as relevant and not submitted to the crowd. Other pages belonged to research projects which explicitly stated their authors. We labelled their author type as 'individual' and did not submit them for evaluation. After this filtering, the datasets submitted to the crowd included 1701 references (T1), 1178 unique URLs (T2), and 335 unique domains (T3.A and T3.B)[7].

[4] We kept the following top-level domains: `tv`, `au`, `gov`, `com`, `net`, `org`, `info`, `edu`, `uk`, `mt`, `eu`, `ca`, `mil`, `wales`, `nz`, `ph`, `euweb`, `ie`, `id`, `info`, `ac`, `za`, `int`, `london`, `museum`.
[5] http://www.census.gov/popest/data/counties/totals/2013/files/CO-EST2013-Alldata.csv
[6] https://figshare.com/articles/GRID_release_2015_12_14/2010108, https://figshare.com/articles/GRID_release_2016-05-31/3409414.
[7] Data and code available at https://github.com/Aliossandro/WD_references_analysis.

Table 4. Sample characteristics. Humans include registered and anonymous users.

	Total instances	Total items	Total statements	Total properties	Total URLs	Total domains	Avg. domains per property	Avg. edits per reference
All	2586	2372	2583	182	1674	345	3	1.03
Added by bots	1175	1108	1,175	30	486	38	3	1
Added by humans	1411	1269	1408	173	1189	325	2.7	1.2

Crowdsourcing Gold Standard. Two of the researchers independently created the gold standard for each task, manually labelling random samples of each of the datasets submitted to the crowd. The size of the annotated samples were determined to ensure that workers would not respond twice to the same question (sample size: T1:333; T2:116; T3.A and T3.B:67). Inter-rater agreement of gold standard questions (using Cohen's kappa) was between moderate and substantial for the four tasks: T1:0.447; T2:0.802; T3.A:0.587; T3.B:0.545. Divergent judgements were settled by mutual agreement. Furthermore, sources assessed in T1 had varying levels of difficulty. In some the information sought could be easily found, whereas others were very technical or contained long text. To better assess the crowd's performance, we labelled each reference in T1 gold standard as 'easy' or 'hard'. We found 239 easy and 94 hard references.

Machine Learning Data. We aimed to build binary classifiers to predict relevance and authoritative of sources. Hence, we converted the judgements collected into binary labels for each of these two outcome variables, i.e. relevant vs. non-relevant and authoritative vs. non-authoritative. We followed Wikidata and Wikipedia verifiability policies to identify the combinations of author and publisher types corresponding to authoritative sources (Table 2). Wikidata contemplates exceptions for sources generally considered as 'bad', e.g. self-published sources are acceptable in references regarding their author. For the purpose of analysis, we classified these types of sources as always not authoritative. We deemed not relevant, nor authoritative, sources with non-working links or that required log in as these were not accessible. We also excluded all references classified as not in English by crowdworkers. After this filtering, the dataset used to train the models had 2550 instances (1781 relevant vs. 769 non-relevant; 1610 authoritative vs. 940 non-authoritative).

4.2 Metrics

Crowdsourcing Experiment. CrowdFlower provides a full report for each task, which includes every response, plus several details about workers, e.g. id, country, and their previous accuracy rate. We extracted from this data the metrics we used to evaluate the performance of crowdworkers. For each task, we measured the percentage of correct answers to test questions, inter-rater agreement (measured as Fleiss' kappa [1]), and completion time.

Predictive models We evaluated the performance of the predictive models by comparing them to a baseline. For the relevance model, the baseline was generated by matching English labels of subject and object of a statement in the source text. A match of both would correspond to a relevant source. In case of labels composed of several words, if any of them were found in a page, we considered that a match. For authoritativeness, a blacklist of deprecated domains has been compiled within the *primary sources tool* project [25]. This list is currently used to exclude non-authoritative sources, thus we judged it as a meaningful term of comparison for an approach assessing reference authoritativeness. We deemed not authoritative all sources whose domain was not included in this blacklist.

4.3 Crowdsourcing Experiment Evaluation

The accuracy of trusted workers, i.e. contributors whose accuracy did not drop under 80%, was higher than that threshold (around 90%) and their responses had Fleiss' kappa between 0.335 and 0.534, indicating fair to moderate agreement. These figures suggest that judgements collected had good quality (see Table 5).

More than half of participants who worked on T1 were discarded due to a low accuracy rate. However, this was the task with the highest rate of microtasks completed per hour (37), i.e. the average number of microtasks successfully completed by the minimum number of workers (5) per hour. Furthermore, workers' accuracy was high on both easy (91.5%) and hard (89.7%) references.

T2 took longer to complete (90h), although not by microtasks/hour (13). The accuracy rate of all contributors to T2 was lower than T1 (72% vs. 75%). Task 3.A appeared to be the most difficult. The accuracy of its overall user pool (including trusted and non-trusted workers) was the lowest, with 66% of correct responses to test questions. Consequently, a high number of contributors were expelled from the task, leading to very long completion times. However, responses to T3.A had a moderate inter-rater agreement (0.435). 94.8% of the responses were confirmed by the first question of T3.B.

4.4 Relevance Evaluation

The ensuing sections report the findings of the reference evaluation. The results presented include both references assessed through crowdsourcing and those previously evaluated by the researchers (see section 4.1).

The majority (67.2%) of sources evaluated in T1 were relevant (see Table 6) (**RQ1**). Non-relevant sources (23.8%) primarily did not support the subject of the statements (20.9% of the total). 7.5% of the pages assessed were not working. Only 1.5% of sources were found to not be in English, meaning that the approach followed to select only English-language pages worked well. Registered and anonymous human users were counted together, as the number of references added by anonymous users in our sample was not sufficient to draw sound conclusions. Overall, human editors added more relevant references than bots (90% vs. 43.8%). Evaluation results by type of user are shown in Table 6.

Table 5. Task statistics (includes test questions)

Task	Microtasks	Total judgements	Trusted judgements	Total workers	Trusted workers	Trusted workers accuracy	Fleiss' k	Time	Cost
T1	1701	13,330	9671	457	218	0.335	87.4%	45h	$858
T2	1178	14,340	9170	749	322	0.534	80.9%	90h	$500
T3.A	345	4325	1950	322	60	0.435	76.9%	81h	$116
T3.B	345	3622	2555	239	116	0.391	68.2%	24h	$119

Table 6. Percentage of relevant sources by type of user

	Humans	Bots	All Users
Relevant	90	30.3	67.1
Not relevant	3.1	58.5	23.9
Page not working	4.9	11.1	7.5
Page not in English	2	0.1	1.5

4.5 Authoritativeness Evaluation

Concerning publisher type, the majority of references pointed to sources published by government agencies (37.5%). Academic institutions were the second most common type (around 24%). This changes if we look at the occurrences of unique web domains. In this case, government agencies slip to 5.8%, whereas 'other companies or organisations' become the most used sources with 19.9%. Regarding editor types, governments were still the most common among both bots and humans. However, the situation differs depending on whether all references are considered or unique domains. This is common to other publisher types and affects especially bot-added sources. Table 7 shows percentages of publisher type by user type, for all references and unique domains.

Organisation staff were by far the most common author type (78%) overall, and both among bot- and human-added references (see Table 8). Sources created by identifiable individuals followed (7.9%) and appear to be reused less often than those authored by organisation (12.5% of unique URLs). Collectively-authored sources represented only 2.9% of our sample. Whereas these were only 0.2% of bot-added pages, they were 5.3% of those created by human users. Finally, applying the criteria in Table 2, 63.7% of the references were classified as authoritative (**RQ2**). We summarised results about reference quality in Table 9.

4.6 Quality Prediction Models

The trained models were binary classifiers aiming to predict non-relevant and non-authoritative references. We used stratified 10-folds cross-validation to estimate the algorithms' performance. Stratified cross-validation ensures outcome classes have the same distribution in the subsets selected in each fold and improves the comparability of different algorithms [10]. The F_1 measure was computed on true and false positive over all folds, providing a more unbiased estimate compared to other methods [10]. We used Matthews correlation coefficient

Table 7. Percentage of sources by type of publisher

	Sources			Unique Domains		
	Humans	Bots	All Users	Humans	Bots	All Users
Governmental agencies	32.7	44.4	37.5	34.2	1.5	5.8
Other companies & organisations	15.3	12.6	14.4	17.6	27	19.9
Academic & research institutions	13	12.6	12.4	15.3	28.2	7.8
Other academic organisations	10.3	12.6	11.2	0.4	1.2	1.2
Cultural institutions	7.7	11.9	15	8.6	28.8	15
Vendors & e-commerce companies	7.3	1.8	5.4	8.6	1	15.9
Non-traditional news & media	3.7	1.2	2.5	4.3	2.9	10.1
Self-published	3	0.2	1.6	2.5	0.1	5.4
Traditional news & media	2	0	1.1	2.4	0	5.2
Political or religious institutions	0.9	4.6	1.2	0.9	4.6	1.7
Academic publishers	0.4	0	0.2	0.5	0	1.1
Others	0.1	0	0.1	0.1	0	0.3

Table 8. Percentage of sources by type of author

	Sources			Unique Domains		
	Humans	Bots	All Users	Humans	Bots	All Users
Organisation	75.7	81.4	78.2	72.4	50.5	65.8
Individual	10.8	4.5	7.9	11.8	13.1	12.5
Collective	5.3	0.2	2.9	6.1	0.6	4.5
Page not working	3.9	0.2	2.1	4.9	0.6	3.7
Page not in English	4.3	13.7	3.7	4.9	35.2	13.4

(MCC) to estimate the level of agreement between predicted and observed labels. MCC has values between -1 and $+1$, with higher values indicating better agreement [3]. Class unbalance was addressed by adjusting prediction weights in SVM and Random Forest [19]. Further details about implementation and hyperparameters of the models are provided in the above cited GitHub repository.

The **relevance** baseline was good at predicting non-relevant sources ($F_1 = 0.84$, MCC $= 0.68$), although it was outperformed by all models. Random Forest provided the best scores. The **authoritativeness** baseline gave worse results ($F_1 = 0.53$, MCC$= 0.15$). All trained models outperformed the baseline, with Random Forest yielding the highest F_1 (0.89) and MCC (0.83). Results for both models are shown in Table 10.

5 Discussion

The crowdsourced experiment provided accurate results, as shown by the level of agreement between workers and the percentage of correct responses to test questions. Task completion times differed greatly, probably due to the task type. T1 asked users to find a piece of information within a web page and seemed to be straightforward. Conversely, the classification tasks T3.A and T3.B were harder. This may be due to the classification system used appearing unclear for workers, or clashing with their prior knowledge, leading to erroneous responses, similar to what has been noted before in taxonomy creation tasks [14]. Nevertheless, the judgements collected in T3.B largely confirmed T3.A.

The majority of references examined included relevant sources, although those added by humans and bots diverged considerably. This (see Table 6) may

Table 9. Percentage of sources by relevance and authoritativeness

	Humans	Bots	All Users
Relevant & authoritative	78.2	41.1	60.8
Relevant & not authoritative	14	2.5	9
Not relevant & authoritative	3.7	2.3	0.7
Not relevant & not authoritative	4.1	55.7	27.8

Table 10. Performance of prediction models for relevance and authoritativeness

		P	R	F_1	AUC-PR	MCC
Relevance	Baseline	0.88	0.83	0.84	0.81	0.68
	Naive Bayes	0.94	0.94	0.90	0.92	0.86
	Random Forest	0.95	0.95	**0.92**	0.94	**0.89**
	SVM	0.94	0.94	0.91	0.94	0.87
Authoritativeness	Baseline	0.71	0.65	0.53	0.62	0.16
	Naive Bayes	0.90	0.90	0.86	0.88	0.78
	Random Forest	0.93	0.92	**0.89**	0.93	**0.83**
	SVM	0.90	0.90	0.89	0.90	0.79

have been caused by a link to a US census dataset that was redirected to another page, which did not contain relevant data anymore. We believe this is not an isolated case. Bots add large numbers of statements in batch, including references. References pointing to invalid URLs may become outdated or invalid. Continuous control from the community is required—the eyeballs required to make all bugs shallow [22]—or a method to automatically check sources.

Government agencies are the most common publisher type, both among human- and bot-added references. Sources are generally authored by organisation staff and not by individuals. Two classes of publisher showed large differences between percentage of references and percentage of unique domains (Table 7). In both categories, the skewness is likely to be determined by the massive automatic generation of statements by bots. This led us to hypothesise that typical bot editing patterns may result in a lower degree of diversity of source types. The data confirmed this: in spite of similar numbers of references by bots and humans (46.3% vs 53.6%), bots used 36 web domains, compared to 295 by humans. This analysis should increase awareness about the current limitations of using bots to add references, and in turn help design bots that follow a more nuanced approach to reference selection.

The distribution of author and publisher types for references did not match Wikipedia [9], despite the partial overlap of the two communities [20]. Almost no news sources are used as sources in Wikidata, compared to the online encyclopedia. Whereas Wikidata recommends primary sources as references, Wikipedia asks editors to use secondary sources and officially disapproves of primary ones, in line with the rule that the encyclopedia cannot contain original research.

Sources are generally split between 'good' and 'bad' (Table 9). Few references are relevant but not authoritative; even fewer are not relevant but authoritative. Accessibility was required, therefore several were classified as neither relevant nor authoritative because they were not working or required to log in. Some pages redirected to a new one, which often was not relevant. These were possibly valid

at the time of addition, but subsequently changed. A frequent check of URL validity may be effective to spot those that have become bad.

The predictive models for relevance and authoritativeness performed well, which may support our intuition that that sources from a website that are good for a type of statement, i.e. using a determined property with defined domain and range, are likely to be good for similar statements. Another explanation may regard the characteristics of references in Wikidata. From a total of around 2000 properties, only about 200 have references. Sources from the same web domain tend to have the same level of quality. On the other hand, the number of domains per property is low. As a consequence, the algorithm may find 'easy' to assess combination of properties and domains. If the number of properties with references and the diversity of web domains used will increase, further research should evaluate how this affects the performance of predictive models of reference quality. It should also seek to understand how to adapt these models to be implemented in Wikidata, to help editors find bad references.

6 Conclusions and future work

The contribution of this paper is twofold. First, this is the first study to evaluate provenance quality in Wikidata. Second, we tested a two-staged approach to evaluate Wikidata references, combining microtask crowdsourcing and machine learning. Crowdsourcing provided accurate evaluation of external references, which were mostly relevant and authoritative. A continuous check by users may be needed to address the issue of links becoming non-valid. Models to predict non-relevant or non-authoritative references may also be useful. With respect to that, our results were encouraging. Our models outperformed the baseline, which motivates towards further work to integrate them in Wikidata. Future work should validate whether our results hold true for non-English sources. Besides using outgoing links, Wikidata expresses provenance by means of internal connections, which were not examined in this study. These are a substantial part of Wikidata references and should be examined in the future, in order to achieve a comprehensive evaluation of provenance quality in Wikidata.

7 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).

References

1. Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., Lehmann, J.: Crowdsourcing linked data quality assessment. In: The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II. Lecture Notes in Computer Science, vol. 8219, pp. 260–276. Springer (2013)

2. Alonso, O., Rose, D.E., Stewart, B.: Crowdsourcing for relevance evaluation. SIGIR Forum 42(2), 9–15 (2008)
3. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16(5), 412–424 (2000)
4. Brasileiro, F., Almeida, J.P.A., de Carvalho, V.A., Guizzardi, G.: Applying a Multi-Level Modeling Theory to Assess Taxonomic Hierarchies in Wikidata. In: Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume. pp. 975–980 (2016)
5. Dai, C., Lin, D., Bertino, E., Kantarcioglu, M.: An Approach to Evaluate Data Trustworthiness Based on Data Provenance. In: Secure Data Management, 5th VLDB Workshop, SDM 2008, Auckland, New Zealand, August 24, 2008, Proceedings. Lecture Notes in Computer Science, vol. 5159, pp. 82–98. Springer (2008)
6. Eickhoff, C., de Vries, A.P.: Increasing cheat robustness of crowdsourcing tasks. Inf. Retr. 16(2), 121–137 (2013)
7. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandecic, D.: Introducing Wikidata to the Linked Data Web. In: The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I. Lecture Notes in Computer Science, vol. 8796, pp. 50–65. Springer (2014)
8. Fetahu, B., Markert, K., Nejdl, W., Anand, A.: 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. pp. 337–346. ACM (2016)
9. Ford, H., Sen, S., Musicant, D.R., Miller, N.: 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. pp. 9:1–9:10 (2013)
10. Forman, G., Scholz, M.: Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. SIGKDD Explorations 12(1), 49–57 (2010)
11. Hartig, O.: Provenance Information in the Web of Data. In: Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009, Madrid, Spain, April 20, 2009. CEUR Workshop Proceedings, vol. 538. CEUR-WS.org (2009)
12. Hartig, O., Zhao, J.: Using Web Data Provenance for Quality Assessment. In: Proceedings of the First International Workshop on the role of Semantic Web in Provenance Management (SWPM 2009), collocated with the 8th International Semantic Web Conference (ISWC-2009), Washington DC, USA, October 25, 2009. CEUR Workshop Proceedings, vol. 526. CEUR-WS.org (2009)
13. Kakol, M., Jankowski-Lorek, M., Abramczuk, K., Wierzbicki, A., Catasta, M.: On the subjectivity and bias of web content credibility evaluations. In: 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume. pp. 1131–1136. International World Wide Web Conferences Steering Committee / ACM (2013)
14. Karampinas, D., Triantafillou, P.: Crowdsourcing taxonomies. In: The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings. Lecture Notes in Computer Science, vol. 7295, pp. 545–559. Springer (2012)
15. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM 46(5), 604–632 (1999)

16. Lehmann, J., Gerber, D., Morsey, M., Ngomo, A.N.: DeFacto - Deep Fact Validation. In: The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I. Lecture Notes in Computer Science, vol. 7649, pp. 312–327. Springer (2012)
17. Lucassen, T., Schraagen, J.M.: Trust in Wikipedia: how users trust information from an unknown source. In: Proceedings of the 4th ACM Workshop on Information Credibility on the Web, WICOW 2010, Raleigh, North Carolina, USA, April 27, 2010. pp. 19–26. ACM (2010)
18. Müller-Birn, C., Karran, B., Lehmann, J., Luczak-Rösch, M.: Peer-production system or collaborative ontology engineering effort: What is Wikidata? In: Proceedings of the 11th International Symposium on Open Collaboration, San Francisco, CA, USA, August 19-21, 2015. pp. 20:1–20:10. ACM (2015)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
20. Piscopo, A., Phethean, C., Simperl, E.: 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) (2017)
21. Potthast, M., Stein, B., Gerling, R.: Automatic Vandalism Detection in Wikipedia. In: Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings. Lecture Notes in Computer Science, vol. 4956, pp. 663–668. Springer (2008)
22. Raymond, E.S.: The cathedral and the bazaar - musings on Linux and open source by an accidental revoltionary (rev. ed.). O'Reilly (2001)
23. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In: 2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 254–263. ACL (2008)
24. Steiner, T.: 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. pp. 25:1–25:7. ACM (2014)
25. Tanon, T.P., Vrandecic, D., Schaffert, S., Steiner, T., Pintscher, L.: 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. pp. 1419–1428 (2016)
26. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM 57(10), 78–85 (2014)
27. Wikidata: Wikidata:Sources — Wikidata, the free knowledge base. https://www.wikidata.org/wiki/Help:Sources (2017), [Online; accessed 09-April-2017]
28. Wikidata: Wikidata:Verifiability — Wikidata, the free knowledge base. https://www.wikidata.org/wiki/Wikidata:Verifiability (2017), [Online; accessed 07-April-2017]
29. Wikipedia: Wikipedia:Verifiability — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Wikipedia:Verifiability (2017), [Online; accessed 07-April-2017]