Towards Vandalism Detection in Knowledge Bases: Corpus Construction and Analysis

Stefan Heindorf[1] Martin Potthast[2] Benno Stein[2] Gregor Engels[1]

[1]University of Paderborn
<last name>@upb.de

[2]Bauhaus-Universität Weimar
<first name>.<last name>@uni-weimar.de

ABSTRACT

We report on the construction of the Wikidata Vandalism Corpus WDVC-2015, the first corpus for vandalism in knowledge bases. Our corpus is based on the entire revision history of Wikidata, the knowledge base underlying Wikipedia. Among Wikidata's 24 million manual revisions, we have identified more than 100,000 cases of vandalism. An in-depth corpus analysis lays the groundwork for research and development on automatic vandalism detection in public knowledge bases. Our analysis shows that 58% of the vandalism revisions can be found in the textual portions of Wikidata, and the remainder in structural content, e.g., subject-predicate-object triples. Moreover, we find that some vandals also target Wikidata content whose manipulation may impact content displayed on Wikipedia, revealing potential vulnerabilities. Given today's importance of knowledge bases for information systems, this shows that public knowledge bases must be used with caution.

Categories and Subject Descriptors: H.3.2 [Information Storage and Retrieval]: Information Storage

General Terms: Documentation, Measurement

Keywords: Corpus; Data Quality; Knowledge Bases; Vandalism

1. INTRODUCTION

Information systems increasingly rely on structured knowledge to answer queries. This pertains particularly to question answering systems that reason about complex question queries, such as IBM Watson and Wolfram Alpha, but also to web search engines which display helpful information for all kinds of queries alongside the traditional "ten blue links." In fact, the additional information provided by the latter often suffices to answer many informational queries straight away, relieving users of browsing search results themselves. Given that, according to Broder [2], up to 48% of web search queries fall into this category, a lot depends on the accuracy of a search engine's knowledge base, or else users are misled.

Many search engines incorporate publicly available knowledge bases, which may be constructed, to various degrees, manually or automatically. Figure 1 overviews and organizes a selection of well-known knowledge bases in terms of their domain specificity and their primary construction principle. Both automatically and manually constructed knowledge bases face quality problems, albeit for different reasons. While automatically constructed knowledge bases solely rely on the validity of the heuristics applied for information extraction from their respective source of information, manually constructed knowledge bases rely on the trustworthiness of their volunteer contributors. Hence, the latter often impose rigid review processes to maintain integrity. Reviewing up to millions of changes every month is a tedious and time-consuming task, and if pending reviews are the reason for publication delays of new contributions, volunteers get discouraged. Not reviewing new contributions, however, may result in periods of time when vandalism, i.e., deliberate attempts to reduce data quality, is publicly visible.

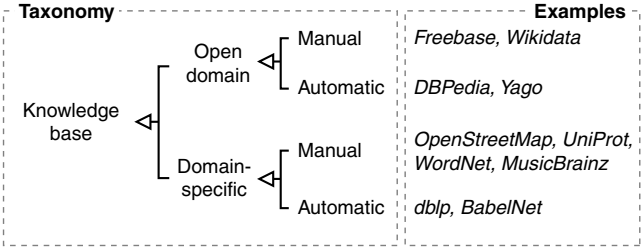

Figure 1: Taxonomy of knowledge bases with respect to their domain specificity and their primary construction principle along with well-known examples for each category.

An automatic solution to aid in reviewing would be of great benefit for the community of a knowledge base. However, vandalism in knowledge bases received little attention until now. The prevalence of the problem is unclear as well as what typical vandalism looks like. This gap of knowledge prevents in-depth research and development. To close this gap, we compile the first standardized vandalism corpus for structured knowledge bases and report on its analysis. As a basis for our corpus, we analyze Wikidata [11], the open-domain knowledge base of the Wikimedia Foundation which increasingly serves as a backend for Wikipedia. Despite its fairly recent launch, Wikidata attracts contributions from a large user base. Our contributions are twofold: (1) we compile the first large-scale corpus of Wikidata vandalism called Wikidata Vandalism Corpus WDVC-2015,[1] and, (2) we conduct a corpus analysis to investigate what content is vandalized and who are the vandals.

In what follows, Section 2 reviews related work, Section 3 gives a short introduction to Wikidata, Section 4 details the corpus construction methodology, and Section 5 reports on our corpus analysis. Section 6 concludes the paper.

[1]The corpus can be found at http://www.uni-weimar.de/medien/webis/corpora/

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

DOI: http://dx.doi.org/10.1145/2766462.2767804.

2. RELATED WORK

To the best or our knowledge, only two papers have tackled the detection of vandalism and low-quality content in *structured* knowledge bases so far: Neis et al. [4] develop a rule-based approach to detect malicious changes to OpenStreetMap, a domain-specific knowledge base for map data. For evaluation, a data set is constructed by analyzing whether or not OpenStreetMap users who have been blocked are vandals. The second paper by Tan et al. [9] focuses on Freebase, an open-domain crowdsourced knowledge base operated by Google. A machine learning-based approach to detect low-quality content is proposed, and for its evaluation, a data set is constructed using a so-called "test of time"-heuristic: basically, all additions that have been deleted within 4 weeks after their submission are considered low-quality. However, the authors of both papers provide little insight into the nature of low-quality content in their subjects of study, nor do the latter assess the performance of their heuristic. While the data sets are available to other researchers, they have not been published as dedicated corpora. Moreover, the usefulness of the Freebase data set may be diminished by the fact that Google is currently shutting down Freebase, whereas its contents will be transferred to Wikidata.

Besides structured knowledge bases, a large body of work can be found on detecting vandalism in what may be called an *unstructured* knowledge base, namely Wikipedia [1, 7, 10, 12, 13]. Many of these approaches have been proposed as a result of two shared tasks on Wikipedia vandalism detection [8, 6]. The evaluation corpus for these shared tasks has been constructed by Potthast [5] by means of manual annotation of a small sample of the revision history of Wikipedia via Amazon Mechanical Turk. The corpus is deemed high quality, yet, it comprises only 2400 revisions labeled vandalism. Therefore, Tran and Christen [10] as well as West et al. [13] construct their own evaluation data to achieve larger scales by means of automatic labeling methods. Both analyze the comments left by editors removing vandalism so as to identify prior revisions where vandalism has been introduced. Except West et al. [13], who analyze the spatio-temporal properties of vandalism, none of the above authors analyze their data sets in-depth. Our corpus complements theirs, but for the domain of structured knowledge bases, and, we provide a corpus analysis from the start.

3. WIKIDATA

Wikidata was launched in October 2012 by the Wikimedia Foundation. Based on the observation that a lot of factual knowledge appears redundantly throughout Wikimedia's projects, the goal of Wikidata is to collect and organize this knowledge in a central place [11]. Just like Wikipedia, anyone can edit Wikidata and the project quickly gained traction: since December 2013 there are constantly over 4000 editors with more than 5 edits per month [16].

Like other knowledge bases, Wikidata basically consists of subject-predicate-object triples. These triples are grouped by subject and each group is called an item. Currently, Wikidata stores about 17 million items and more than 50 million triples [17]. Every item is represented on its own Wikidata page which is divided into five parts: a unique label, a short plain text description, a set of aliases, a set of statements, and a set of sitelinks. For example, the item labeled "Barack Obama" is described as "44th President of the United States" and has aliases such as "Barack Hussein Obama II" and "Barack Obama II." Statements where this item forms the subject include "instance of: human" and "date of birth: 4 August 1961," and its sitelinks refer to all Wikipedia articles in different languages that cover Barack Obama.[2] Labels, descriptions, and aliases are multilingual and can have different values in every language. We refer to labels, descriptions, and aliases as textual content, whereas statements and sitelinks are structured content. Every edit of an item is recorded and assigned a unique ID within the revision history of Wikidata.

[2] http://www.wikidata.org/wiki/Q76

4. CORPUS CONSTRUCTION

Our corpus is designed with three goals in mind: it allows for the analysis of vandalism in knowledge bases, it allows for the evaluation of machine-learning approaches that detect vandalism, and it is as far as possible compatible to related corpora from the literature to allow for cross-domain evaluations (e.g., applying Wikipedia vandalism detection approaches to the structured data of Wikidata). In particular, we label revisions as vandalism or not. The following subsections detail the raw data of Wikidata, and how we apply an automatic labeling process based on the revision history.

4.1 Wikidata Database Dump

Our corpus is based on a database dump of the full revision history of Wikidata until November 7, 2014, which is provided by the Wikimedia Foundation free of charge [14]:

Revisions made on Wikidata	167,802,227	100 %
Revisions made on meta pages	1,211,628	1 %
Revisions made on special items	11,167	0 %
Revisions made automatically	142,574,999	85 %
Revisions made manually	24,004,433	14 %

Not all revisions enclosed in the dump are on Wikidata items, but on meta pages (e.g., user pages, discussion boards, etc.) and test items (sandbox items and Wikidata Tours items). We filter these revisions for their irrelevance to our corpus. From the remainder, about 85% of revisions are made automatically by bots approved by Wikidata's community. Basically, they are rule-based scripts that simplify tedious tasks. However, bots cannot fully replace the careful, manual curation of the data. As we are interested in detecting ill-intentioned contributions by humans, and not errors in bots, we base our corpus on the 24 million manual revisions.

4.2 Automatic Revision Labeling

Since vandalism detection is a classification task, we label all manual revisions as vandalism or not. While manually labeling such a large quantity of revisions is infeasible, we resort to automatic labeling and manual validity checks instead. The goal is to label as much vandalism as possible in a way that maintains precision, while being robust against vandal interference. Two of Wikidata's editing facilities are exploited for this purpose, namely *rollback* operations and *undo/restore* operations. While the former have previously been used by West et al. [13] to label vandalism in Wikipedia, neither has been manually validated until now.

Rollback. There are about 200 administrators and privileged users on Wikidata who are entitled to use the rollback facility: with a single click, a rollback reverts all consecutive revisions of the last editor of a given item. According to the Wikidata help, a "rollback should only be used to revert vandalism and test edits" [15]. Hence, all revisions that are reverted in a rollback can be considered vandalism. The use of the rollback facility is automatically logged in the comment of the resulting revision, so that identifying preceding revisions where vandalism was introduced is straightforward.

Undo/Restore. Like rollbacks, the undo/restore facility allows for reverts: the undo command reverts a single revision and the restore command restores an item to a previous state, undoing all intermediate revisions. Unlike rollbacks, however, the undo/restore facility is available to everyone, including unregistered users. The Wikidata help does not explicitly mention specific situations for which using this facility is reserved. Furthermore, users are free

to change the automatically suggested comment. Hence, it is not possible to identify all undo/restore uses as reliable as rollbacks, however, users will often not bother to change the comment.

Of the 24 million manual revisions made on Wikidata, a total of 103,205 have been reverted via rollbacks, and 64,820 via undo/restore. Based on our below validity analysis, we label rollback revisions as vandalism, whereas this cannot be done with confidence for undo/restore revisions. Further, the natural level of granularity underlying Wikidata's database dumps is revisions, but vandals often strike the same item more than once in a row. Such events are quite different from isolated instances of vandalism by different users. Therefore, we also group consecutive revisions by the same user and annotate those revision groups. Nevertheless, our corpus keeps revisions as the level of granularity to maintain compatibility with related corpora from the literature.

4.3 Corpus Validity

To validate our labeling strategy, we manually review a random sample of 1000 rollback revisions, 1000 undo/restore revisions, and 1000 inconspicuous revisions. For each revision, the reviewer was presented with a visualization of the difference between the item states before and after the revision took place in the form of the well-known Wikipedia diff pages. To base the reviewer's decision solely on the content of a revision, meta information such as user names, IP addresses, dates, etc. were suppressed. We found 86%±3% (95% confidence level) of the rollback revisions to be vandalism, and 62%±3% of the undo/restore revisions. Furthermore, the reviewer annotated 4% (42) of the 1000 inconspicuous revisions as vandalism. We have analyzed these revisions more closely and found that only 1% (11) constitute actual vandalism. In many cases, only the context of a revision, which was invisible to our reviewer, reveals that it is not actually vandalism.

From a machine learning perspective, these results are encouraging: they show that, despite automatic labeling, the rollback revisions have a very low noise ratio, whereas the undo/restore revisions are on the borderline of being useful training data. For example, a biased SVM and weighted logistic regression can deal with up to 40% noisy labels [3]. Furthermore, the confidence in our labels is more easily obtained than that of Wikipedia vandalism corpora: more than 50% of revisions on Wikipedia require at least 9 reviewers to be labeled with a 2/3 majority vote [5]. In our case, two reviewers—the Wikidata admin and our reviewer—are sufficient to arrive at a reasonable quality. Altogether, the rollback revisions make up for a strong signal of vandalism on Wikidata, and this signal cannot be manipulated by vandals, unless they first infiltrate the senior staff of Wikidata. Regarding the undo/restore revisions, we leave denoising efforts, e.g., by means of statistical analyses and crowdsourcing, as future work.

5. CORPUS ANALYSIS

This section reports selected results of an exploratory analysis of our corpus. Particularly, we shed light on the following questions: (1) What is vandalized on Wikidata? (2) Who vandalizes Wikidata? But before going into details about these questions, we give key figures and statistics about our corpus, and Wikidata in general.

5.1 Corpus Statistics

Figure 2 visualizes the genesis of Wikidata in terms of manual revisions made per month from its launch in October 2012 until October 2014. Wikidata's growth rate significantly increased twice in March 2013 and again in May 2014. To determine the cause for these events, we analyzed all revisions with regard to the type of content affected. To determine the content type of a revision,

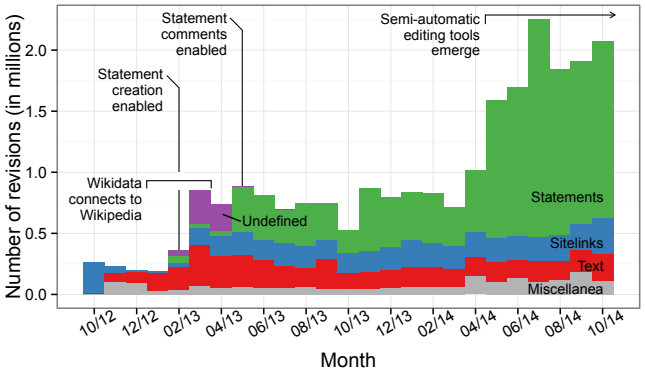

Figure 2: Manual revisions on Wikidata per month. Revisions affecting textual content (labels, descriptions, and aliases) are distinguished from revisions affecting structural content (statements and sitelinks). Major growth events are labeled.

we parse the automatically generated comments by the Wikibase software. The miscellaneous revisions includes merging of items, item creation and revisions with empty comments.

The first jump of growth rate was caused by enabling statement creation for first time. In the months around this event, Wikidata was connected to the Wikipedias in various languages, adding millions of statements and sitelinks. At that time, revisions affecting statements often had no automatically generated comment ("Undefined" in Figure 2). The second growth rate increase is due to the emergence of semi-automatic editing tools for Wikidata, most notably the Wikidata Game. This game with a purpose displays part of a Wikipedia page and asks the editor, for example, whether it is about a person and what their demographics are. In this way, editors can make a large amount of contributions in a short time.

5.2 What is vandalized on Wikidata?

Our answer to this question is twofold; we investigate (1) which categories of items are vandalized, and (2) which content types are vandalized (e.g., textual content or structural content).

Vandalized Item Categories. The distribution of vandalism on Wikidata is flat. Table 1 (left) shows the top vandalized items on Wikidata, where the top vandalized item accumulates only 47 of all vandalism cases (in terms of revision groups), whereas 70% (44,646 of 63,455) of the cases are the only ones in their items' revision histories. To learn which kinds of items are affected most by vandalism, we categorized the top 1000 vandalized items, and, the top 1000 items with the most revision groups (regardless whether they contain vandalism). Table 1 (right) contrasts the results. The distribution of attention over categories differs wildly when comparing vandals and all editors: disregarding categories with less than 1% share, the least vandalized category Places gets almost 4 times as much attention by all editors (31%), making it the most edited category. The categories People and Nature are on par, but the categories Culture, Society, Meta items, and Technology each get roughly double as much attention by vandals than overall. The focus of vandals deviates significantly from typical editors, giving rise to category-specific detection approaches.

The category Meta items comprises items that are used to automatically generate parts of important Wikipedia pages. By manipulating these items, a vandal may have tremendous impact. Apparently, many vandals attempt to exploit this potential vulnerability, since these items are otherwise of no particular interest (i.e., many contain only sitelinks). Moreover, while categorizing the revision samples, we noticed that 11% of the vandalized items concerned India, cross-cutting all categories, compared to 0.5% overall. The reason for this is still unclear, but merits further investigation.

Table 1: Top vandalized items (left); top vandalized categories and top edited categories in a sample of 1000 items each (right).

Cases	Item title
47	Cristiano Ronaldo
43	Lionel Messi
43	One Direction
41	Portal:Featured content
34	Justin Bieber
33	Barack Obama
29	English Wikipedia
29	Selena Gomez

Category	Vandalism	All
Culture	20%	12%
People	20%	21%
Society	16%	9%
Nature	14%	15%
Meta items	13%	8%
Technology	9%	4%
Places	8%	31%
Other	1%	1%

Vandalized Content Types. Table 2 (row 1) overviews which of the five content types of an item are most often vandalized. About 57% of the vandalism happens in textual content like labels, descriptions, and aliases; and about 40% happens in structural content like statements and sitelinks. The remaining 2% of miscellaneous vandalism includes merging of items and indecisive cases.

An explanation for the large portion of vandalism in textual content can be explained by the fact that textual content is more convenient to vandalize: it appears at the top of an item page and Wikibase allows for unrestricted plain text content. This renders the adaptation of Wikipedia vandalism detection approaches from the literature to detecting vandalism on Wikidata feasible. Nevertheless, up to 40% of vandalism can be found in structured contents of Wikidata items. Vandalism in these contents fulfills all the formatting constraints but still represents knowledge that must be considered malicious with regard to the item in question. Detecting such forms of vandalism requires the development of entirely new approaches to vandalism detection. A vandalism detection solution for Wikidata cannot work without also addressing this problem, since vandals are quick to notice weaknesses and they will adjust their behavior to exploit them.

5.3 Who vandalizes Wikidata?

About 86% (88,592 of 103,205) of vandalism on Wikidata originates from anonymous users. But simply considering all these revisions vandalism yields only 0.20 F_1 at 0.12 precision (88,592 of 768,027) and 0.86 recall—a baseline, but no justification to abolish anonymous editing. Only about 41% (35,979 of 88,592) of anonymous vandalism originates from an IP that has been previously used for vandalism. On average, one IP address is used for 1.7 (88,592 / 52,613) vandalism revisions. However, spot checks revealed that the IP addresses of some vandals on Wikidata also had a recent track record on Wikipedia, but a systematic analysis of cross-wiki vandalism is left for future work. For comparison, when vandals register before vandalizing, they tend to use their accounts an average 6.4 times (14,613 / 2,273) to vandalize and about 84% (12,340 of 14,613) of vandalism comes from accounts that have been previously used to vandalize. The available data does not reveal if vandals then switch accounts or just leave.

Finally, rows 2 and 3 of Table 2 break down vandalism from unregistered users and registered users by content type. Unregistered users primarily vandalize textual content and sitelinks, whereas registered users primarily vandalize statements and sitelinks. An explanation for this difference eludes our analysis so far, but it can be used to train different models via machine learning algorithm for each case, improving overall classification performance.

6. CONCLUSION AND OUTLOOK

In this paper, we report on the construction of a new resource for studying low-quality content in knowledge bases, namely vandalism. Vandalism has been one of the primary problems of Wikipedia, and with the growing success of Wikidata, this project will become a bigger target for vandals as well. Vandalism in knowledge bases can have severe consequences, if information systems rely on its integrity and return facts extracted from the knowledge base without double-checking.

Table 2: Vandalism by content type and user anonymity.

Vandalism	Item content type				Revisions
	Text	Statements	Sitelinks	Misc.	
Overall	57%	8%	32%	3%	103,205
Unregistered	63%	5%	31%	1%	88,592
Registered	19%	29%	39%	13%	14,613

To enable future research in this domain, and to gain insights into the nature of vandalism in knowledge bases, we have compiled the first standardized corpus of knowledge base vandalism. Our corpus encompasses the entire history of Wikidata, and vandalism has been reliably identified using an automatic labeling strategy that cannot be easily manipulated. We have analyzed the vandalism found with regard to the mostly affected items, the affected types of content within items, and with regard to typical user behavior. Based on these insights, our future work will be the development a new approach to detect vandalism in knowledge bases.

Acknowledgement

This work was supported by the German Research Foundation (DFG) within the Collaborative Research Center "On-The-Fly Computing" (CRC 901).

References

[1] B. Adler, L. de Alfaro, S. M. Mola-Velasco, P. Rosso, and A. G. West. Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. CICLing 2011.
[2] A. Broder. A Taxonomy of Web Search. *SIGIR Forum* 36(2) 2002.
[3] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. NIPS 2013.
[4] P. Neis, M. Goetz, and A. Zipf. Towards Automatic Vandalism Detection in OpenStreetMap. *ISPRS Int. J. Geo-Inf.* 1(3) 2012.
[5] M. Potthast. Crowdsourcing a Wikipedia Vandalism Corpus. SIGIR 2010.
[6] M. Potthast and T. Holfeld. Overview of the 2nd International Competition on Wikipedia Vandalism Detection. CLEF 2011.
[7] M. Potthast, B. Stein, and R. Gerling. Automatic Vandalism Detection in Wikipedia. ECIR 2008.
[8] M. Potthast, B. Stein, and T. Holfeld. Overview of the 1st International Competition on Wikipedia Vandalism Detection. CLEF 2010.
[9] C. H. Tan, E. Agichtein, P. Ipeirotis, and E. Gabrilovich. Trust, but Verify: Predicting Contribution Quality for Knowledge Base Construction and Curation. WSDM 2014.
[10] K.-N. Tran and P. Christen. Cross Language Learning from Bots and Users to detect Vandalism on Wikipedia. *IEEE TKDE* 27(3), 2014.
[11] D. Vrandečić and M. Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57(10), 2014.
[12] W. Y. Wang and K. R. McKeown. "Got You!": Automatic Vandalism Detection in Wikipedia with Web-based Shallow Syntactic-semantic Modeling. COLING 2010.
[13] A. G. West, S. Kannan, and I. Lee. Detecting Wikipedia Vandalism via Spatio-temporal Analysis of Revision Metadata? EUROSEC 2010.
[14] Wikimedia Foundation. Wikidata XML Dump. http://dumps.wikimedia.org/wikidatawiki/20141106/wikidatawiki-20141106-pages-meta-history.xml.bz2, 2014.
[15] Wikimedia Foundation. Wikidata Rollbackers Help. http://www.wikidata.org/wiki/Wikidata:Rollbackers, 2015.
[16] Wikimedia Foundation. Wikidata Statistics. http://stats.wikimedia.org/wikispecial/EN/TablesWikipediaWIKIDATA.htm, 2015.
[17] Wikimedia Foundation. Wikidata Statistics. https://tools.wmflabs.org/wikidata-todo/stats.php, 2015.