Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux): A Global Study of Edit Activity on Wikipedia and Wikidata

Extended Version of the WWW2014 Web Science Track Short Paper

Thomas Steiner*
Google Germany GmbH
ABC-Str. 19
20354 Hamburg, Germany
tomac@google.com

ABSTRACT

Wikipedia is a global crowdsourced encyclopedia that at time of writing is available in 287 languages. Wikidata is a likewise global crowdsourced knowledge base that provides shared facts to be used by Wikipedias. In the context of this research, we have developed an application and an underlying Application Programming Interface (API) capable of monitoring realtime edit activity of all language versions of Wikipedia and Wikidata. This application allows us to easily analyze edits in order to answer questions such as "Bots *vs.* Wikipedians, who edits more?", "Which is the most anonymously edited Wikipedia?", or "Who are the bots and what do they edit?". To the best of our knowledge, this is the first time such an analysis was done for Wikidata *and* for really *all* Wikipedias—large and small. According to our results, all Wikipedias *and* Wikidata together are edited by about 50% bots and by about 23% anonymous users. Wikidata alone accounts for about 48% of the totally observed edits. If we do *not* consider Wikidata, *i.e.*, if we *only* look at all Wikipedias, about 15% of all edits are made by bots and 26% of all edits are made by anonymous users. Overall, we found a stabilizing number of 274 active bots during our observation period. Our application is available publicly online at the URL `http://wikipedia-edits.herokuapp.com/`, its code has been open-sourced under the Apache 2.0 license.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services

General Terms

Human Factors, Languages, Measurement, Experimentation

*Thomas Steiner's second affiliation is *Université de Lyon, CNRS Université Lyon 1, LIRIS, UMR5205, F-69622*

OpenSym '14, August 27–29 2014, Berlin, Germany

http://dx.doi.org/10.1145/2641580.2641613 ...$15.00.

Keywords

Wikipedia, Wikidata, realtime monitoring, study, Web app

1. INTRODUCTION

How Wikipedia Came to Life.

The fundamental shift from book-based encyclopedias to CD-ROM-based encyclopedias to finally Web-based encyclopedias happened in the course of the 90ies and started a new era of freely and openly available knowledge accessible to everybody. The free online encyclopedia Wikipedia[1] [19] was formally launched on January 15, 2001 by Jimmy Wales and Larry Sanger, albeit the fundamental wiki technology and the underlying concepts are older. Wikipedia's direct predecessor was Nupedia [19], a similarly free online encyclopedia, however, that was exclusively edited by experts following a strict peer-review process. Wikipedia's initial role was to serve as a collaborative platform for draft articles for Nupedia. What happened in practice was that Wikipedia rapidly overtook Nupedia as there was no peer-review burden and it is now a globally successful Web encyclopedia available in 287 languages with overall more than 30 million articles.[2]

International Expansion.

The international expansion began early on in the project's existence, with the first two non-English Wikipedias both started on March 16, 2001 being the German and the Catalan ones, followed briefly afterward by (Romanized) Japanese. What followed was a wave of new languages, with French, Chinese, Dutch, Esperanto, Hebrew, Italian, Portuguese, Russian, Spanish, Swedish, Arabic, Hungarian, Afrikaans, Norwegian, and Serbian all being rolled out in the first year.

The First Wikipedia Bots.

Wikipedia bots are computer programs with the purpose of automatically editing Wikipedia. After occasional smaller-scale tests, the first large-scale bot operation was started in October 2002 by Derek Ramsey,[3] who created a bot to add a large number of articles about United States towns based on tabular information stemming from U.S. census data. The generated articles used a uniform text template,

[1]Wikipedia: `http://www.wikipedia.org/`
[2]Wikipedia statistics: `http://stats.wikimedia.org/`
[3]History of Wikipedia bots: `http://bit.ly/History_Bots`

so that all articles followed the same writing style. Today, bots are not only used to generate articles, but also to fight vandalism and spam, to correct typographic errors, to improve references, and many more automatable tasks.[4]

The Knowledge Base Wikidata.

As Wikipedia is a truly global effort, sharing non-language-dependent facts like population figures centrally in a knowledge base makes a lot of sense to facilitate international article expansion. Wikidata[5] [23] is a free knowledge base that can be read and edited by both humans and bots. The knowledge base centralizes access to and management of structured data, such as references between Wikipedias and statistical information that can be used in articles. Controversial facts such as borders in conflict regions can be added with multiple values and sources, so that Wikipedia articles can, dependent on their standpoint, choose preferred values.

Contributions.

The contributions of this paper are twofold. On the engineering side, first, we have developed an application and released its source code as open-source that allows for realtime monitoring of all 287 Wikipedias and Wikidata. Second, we have permanently made available a publicly useable Application Programming Interface (API) that our application is based upon and that we invite other interested parties to use. On the research side, during the observation period from November 4 to November 6, 2013, we have monitored exactly 3,805,185 Wikipedia and Wikidata edits, out of which exactly 1,918,378 ($\sim$ 50.4%) were made by bots. From the 3,805,185 total edits, 1,837,146 ($\sim$ 48.3%) can be allotted to Wikidata. We have compiled global and local statistics on the impact of bot edits *vs.* human Wikipedian edits, on anonymous human edits *vs.* logged-in human edits, and on bot activity in general. In continuation, we have deep-dived into the data and looked at the most *bot-edited*, *human-edited*, and *anonymously-edited* Wikipedias and Wikidata. The current paper is an extended version of the short paper *Bots vs. Wikipedians, Anons vs. Logged-Ins* [20] that appeared as a poster in the Web Science track at the World Wide Web conference 2014 in Seoul, Korea. The short paper focused on the methodology, the present long paper additionally focuses on the obtained results.

2. METHODOLOGY AND TOOLS

In our application, we make use of two enabling technologies, namely the Wikipedia Recent Changes Internet Relay Chat (IRC) feed and a push API called Server-Sent Events.

Wikipedia Recent Changes.

Whenever a human or bot changes an article of any of the 287 Wikipedias,[6] a change event gets communicated by a chat bot over the Wikimedia IRC server (`irc.wikimedia.org`),[7] so that parties interested in the data can listen to the changes as they happen [21]. For each language version, there is a specific chat room following the pattern `"#" + language + ".wikipedia"`. For example, changes to Catalan Wikipedia articles will be streamed to the room `#ca.wikipedia`. An exception from this pattern is the room `#wikidata.wikipedia` for the language-independent knowledge base Wikidata [23]. A sample original chat message with the components separated by the asterisk character `'*'` announcing a change to an article can be seen in the following. `"[[Keep Calm and Carry On]] http://en.wikipedia.org/w/index.php?diff=585806152&oldid=585805943 74.197.171.148 * (+14) /* Parodies */"`. The message components are *(i)* article name, *(ii)* revision URL, *(iii)* Wikipedia editor handle, and *(iv)* change size and description.

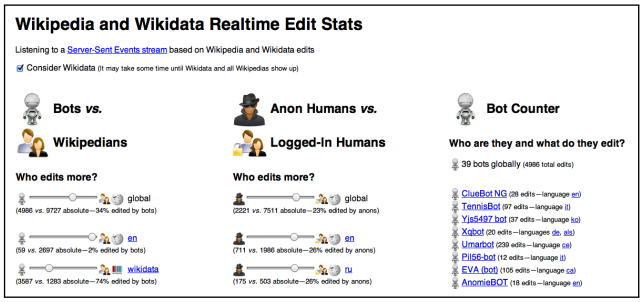

Figure 1: Screenshot of the application available at `http://wikipedia-edits.herokuapp.com/` (edited for legibility), the source code is available under the terms of the Apache 2.0 license

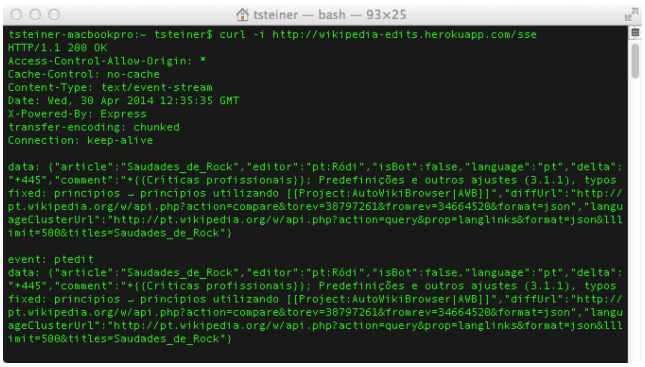

Figure 2: Output of the command line tool cURL showing the MIME type `text/event-stream`, the CORS header, and two edits when connecting to the API `http://wikipedia-edits.herokuapp.com/sse/`

Server-Sent Events.

Server-Sent Events [10] defines an API for using an HTTP connection to receive push notifications from a server in the form of DOM events. Therefore, on the server side, a script generates messages of the MIME type `text/event-stream` in an event stream format that can be seen in Listing 1. The required event payload is in the `data:` field, events can optionally be typed via a proceeding `event:` field and be uniquely identified via a proceeding `id:` field. The `data:` field allows no line breaks, however, multiline messages are possible by prepending a separate `data:` string to each line. Lines that start with a colon are comments and are therefore ignored. Consecutive events are separated by two line

[4]Wikipedia bots by purpose: `http://en.wikipedia.org/wiki/Category:Wikipedia_bots_by_purpose`
[5]Wikidata: `http://www.wikidata.org/`
[6]List of Wikipedias by size: `http://meta.wikimedia.org/wiki/List_of_Wikipedias`
[7]Raw IRC feeds of recent changes: `http://meta.wikimedia.org/wiki/IRC/Channels#Raw_feeds`

breaks. The `EventSource` interface enables Web applications to listen to pushed events from a server over the HTTP protocol. On the client side, using this API consists of creating an `EventSource` object and registering event listeners, as can be seen in Listing 2. When the client loses connection, it automatically reconnects, *i.e.* without manual interaction.

Implementation Details.

Our application is based on a Server-Sent Events API that we have implemented in Node.js, a server side JavaScript software system designed for writing scalable Internet applications. Programs are created using event-driven, asynchronous input/output operations to minimize overhead and maximize scalability. Using Martyn Smith's Node.js IRC library,[8] we listen for Wikipedia and Wikidata edit events and send Server-Sent Events whenever we detect one. We detect bots by checking if the bot flag is set—and as this is not the case for all bots—additionally by checking if any of the editor's name components contain the pre- or suffix "bot".[9] Our API is available publicly online at the URL `http://wikipedia-edits.herokuapp.com/sse` and open for third parties to use. On the client side, in the actual application, we have registered generic event handlers for events pushed by the API and keep track of edit statistics over time. This application can be tested at `http://wikipedia-edits.herokuapp.com/`. A screenshot of the application can be seen in Figure 1, the raw output of the command line tool cURL when connecting to the API can be seen in Figure 2.

[8]Node IRC: `https://github.com/martynsmith/node-irc`

```
: // Embedded JSON data without line breaks

id: 1386885965813
event: enedit
data: {
  "article": "Golden_Globe_Award_for_Best_↩
      Actress_-_Motion_Picture_Musical_or_Comedy",
  "editor": "en:86.150.237.133",
  "isBot": false,
  "language": "en",
  "delta": "+9",
  "comment": "/* 2010s */",
  "diffUrl": "http://en.wikipedia.org/w/api.php?↩
      action=compare&torev=585820379&fromrev=↩
      585776128&format=json",
  "languageClusterUrl": "http://en.wikipedia↩
      .org/w/api.php?action=query&prop=↩
      langlinks&format=json&lllimit=500&↩
      titles=Golden_Globe_Award_for_Best_↩
      Actress_-_Motion_Picture_Musical_or_Comedy"
}
```

Listing 1: Server-Sent Event of type "enedit" (formatted for legibility, `data:` allows no line breaks)

```
// connect to an SSE stream at relative URI /sse
var source = new EventSource('/sse');

// generic event listener for any Wikipedia edit
source.addEventListener('message', function(e) {
  var data = JSON.parse(e.data);
  console.log(e.lastEventId + ' ' + data.article);
}, false);

// listener for English Wikipedia edits
source.addEventListener('enedit', function(e) {
  var data = JSON.parse(e.data);
  console.log(e.lastEventId + ' ' + data.article);
}, false);
```

Listing 2: Creation of an `EventSource` object and registration of two event listeners

3. RESULTS

We have observed all 287 Wikipedias and Wikidata during the observation period from November 4 to November 6, 2013. While this may not sound like a long time, already during this short period overall exactly 3,805,185 edit events occurred. Reinoso *et al.* have shown in [18] that accessing and editing patterns to Wikipedia articles in different languages are strongly dependent on the weekday and special seasons (*e.g.*, academic vacations), so our results should be regarded with their observations in mind. Our application updates in realtime, which allows us to detect when relative figures, *i.e.*, percentages of bots *vs.* Wikipedians and anonymous *vs.* logged-in humans start to converge. This was the case after about half of the observation period. At the end of the observation, from all 287 Wikipedias and Wikidata, exactly 260 ($\sim$ 90.3%) were edited, which, given the long-long-tail of Wikipedias with very few articles, justifies our observation period duration. Our application and underlying API being publicly available, interested parties can run longer analyses at will. In the following, we will zoom in on some interesting data points. As a side-remark—in consistence with Section 2—we treat Wikidata like a language when describing our results.

Linearity of Number of Bots and Edits.

After the above-mentioned half of the observation period, the number of bots *vs.* the number of edits started to grow in a linear way, as can be seen in Figure 3, which supports the fact that relative figures from there on converged.

The Hockey Stick of Most Active Bots.

Figure 4 shows a classic hockey stick curve of the most active bots. Dexbot[10] and ValterVBot[11] lead the field, what follows next is a long-tail of slowly declining bot activity that, starting from KrBot, is roughly linear. Dexbot and ValterVBot are both heavily active on Wikidata, KrBot[12] is active on Wikidata and the Russian and Ukrainian Wikipedias.

The Few Linguistic Geniuses.

Figure 5 shows that only ten bots are active in five languages or more, the clear leaders being タチコマ robot[13] with

[9]Bot detection code: `https://github.com/tomayac/wikipedia-edits-server-sent-events/blob/master/server.js#L111-L113`
[10]Dexbot: `http://en.wikipedia.org/wiki/User:Dexbot`
[11]ValterVBot: `http://bit.ly/ValterVBot`
[12]KrBot: `https://ru.wikipedia.org/wiki/%D0%A3%D1%87%D0%B0%D1%81%D1%82%D0%BD%D0%B8%D0%BA:KrBot`
[13]タチコマ robot: `http://en.wikipedia.org/wiki/User:`

102 languages and EmausBot[14] with 96 languages. Both bots are consistently flagged as global bots.[15] A large majority of 86% of all bots are only active in one language.

The Linguae Francae of Wikipedia and Wikidata.

Given the previous results that only very few bots act in more than one language, we asked ourselves what were the languages that attracted the most bots. Figure 6 shows that the clear winners are the English Wikipedia and Wikidata, and, starting from the French Wikipedia, a slow almost linear decline in the number of active bots can be seen. Overall only 5% of all languages attracted more than ten bots.

The Relatively Most Bot-Edited Languages.

Requiring a minimum threshold of 1,000 bot edits, we looked at what were the relatively (*i.e.*, not absolutely) most bot-edited languages. Figure 7 shows the results, with the fact that the Sindhi Wikipedia, during the observation period, was 100% bot-edited, followed by the Sorani Wikipedia with 92% and Wikidata with 88% bot activity.

The Relatively Most Human-Edited Languages.

Requiring a minimum threshold of 1,000 human edits, we looked at what were the relatively most human-edited languages. According to our results that can be seen in Figure 8, during the observation period, the Galician, the Slovenian, and the Javanese Wikipedias were to 100% human-edited, followed by a very smoothly descending stairway curve with 31 languages being to 90% or more human-edited.

The Relatively Most Anonymously-Edited Languages.

Looking at only human editors, we analyzed the relatively most anonymously-edited languages, requiring a minimum number of 1,000 anonymous edits in order to be counted. Our results depicted in Figure 9 show an almost linearly declining curve with the Thai, the Korean, and the Simple English Wikipedias being at the top with 41% and twice 38% anonymous editing activity in the observation period.

The Relatively Most Logged-In-Edited Languages.

Figure 10 shows the languages that had the most logged-in edits, requiring a minimum number of 1,000 logged-in edits. The Javanese and the Esperanto Wikipedia together with Wikidata mark the top logged-in-edited languages, with the other languages declining approximately linearly.

4. DISCUSSION

Interpreting and analyzing our obtained results, there are some interesting discussion points. We will deep-dive into some of them and provide interpretations in the following.

The Long-Tail of Bots.

Figure 4, Figure 5, and Figure 6 all show a tendency toward hockey stick curves. This implies that few outliers dominate the long-tail rest. Our results are in accordance with the work of Geiger *et al.* [7], who have examined Wikipedia quality control in the absence of the anti-vandalism

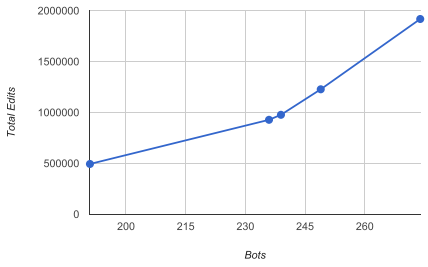

Figure 3: Number of bots *vs.* number of total edits (all languages) showing linear growth

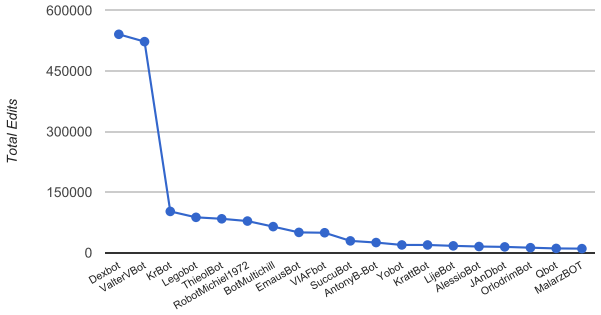

Figure 4: Bots with more than 10,000 edits

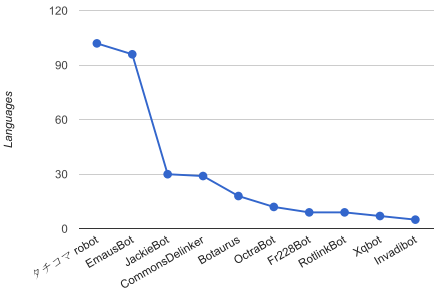

Figure 5: Multilingual bots active in ≥ 5 languages

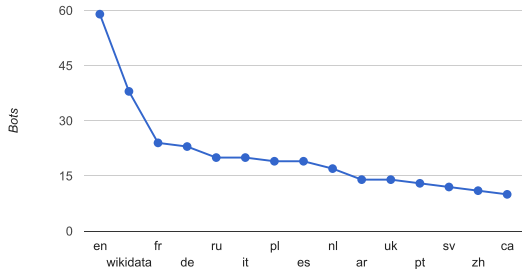

Figure 6: Languages with ≥ 10 active bots

%E3%82%BF%E3%83%81%E3%82%B3%E3%83%9E_robot

[14]EmausBot: http://en.wikipedia.org/wiki/User:EmausBot

[15]Bots with global flag: http://en.wikipedia.org/w/index.php?title=Special:GlobalUsers/Global_bot

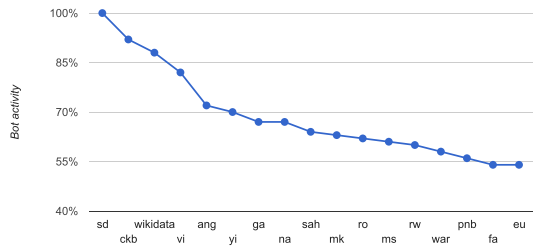

Figure 7: Percentage of bot activity $\geq 50\%$ per language with $\geq 1,000$ bot edits

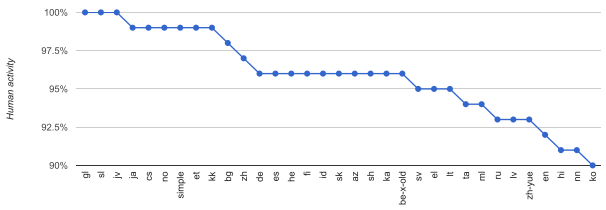

Figure 8: Percentage of human activity $\geq 90\%$ per language with $\geq 1,000$ human edits

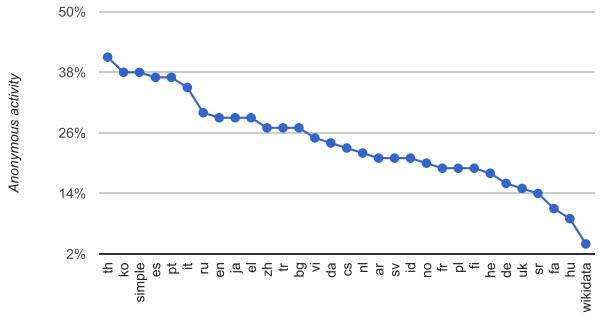

Figure 9: Percentage of anonymous activity per language with $\geq 1,000$ anonymous edits

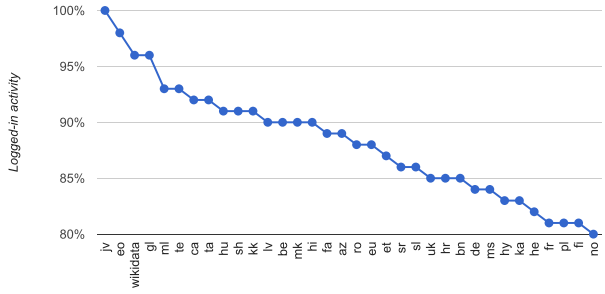

Figure 10: Percentage of logged-in activity $\geq 80\%$ per language with $\geq 1,000$ logged-in edits

bot ClueBot NG[16]—according to our results the 20$^{\text{th}}$-most-active bot and, at a global view, still part of the hook of the hockey stick. Their results suggest that a long-tail of bots take over pending corrective tasks, however, at a slower rate, which acknowledges the importance of the long-tail.

Bots vs. Wikipedians.

Bots on Wikipedia and Wikidata are used to make repetitive automated edits that would be extremely tedious to do manually. It is interesting to compare the scales of the x- and y-axes of Figure 7 and Figure 8. In the prior case (most bot-edited), the y-scale goes from 40% to 100% and the x-scale lists 17 languages. In the latter case (most human-edited), the y-scale goes from 90% to 100% and the x-scale lists 33 languages. What this implies is that the amount of languages dominantly edited by humans is a lot higher than the amount of languages dominantly edited by bots. This suggests that while bots can reach quite impressive qualities at contributing to Wikipedia [8], overall humans still excel.

Anons vs. Logged-Ins.

It is a well-known fact that the number of Wikipedia editors has been declining in recent years [9]. Looking at Figure 9 and Figure 10, we can observe a similar trend as in the paragraph above. The y-axis for the anonymously-edited languages ranges from 2% to 50% with 31 languages listed on the x-axis, compared to the y-axis for the most logged-in-edited languages ranging from 80% to 100% with the x-axis listing 34 languages, which implies that the amount of logged-in activity greatly outperforms the amount of anonymous activity. We leave comparing logged-in edits to anonymous edits open for future work, the hypothesis being that anonymous activity is more likely to either be vandalism and spam [2] or smaller corrective edits, whereas logged-in activity is more likely going to produce quality content.

5. RELATED WORK

The Wikimedia Foundation themselves provide statistics for all Wikipedias with ten or more articles and edits in the previous month [28], as well as statistics for Wikidata [27], a regular Wikipedia article on Wikipedia statistics [25], and a special auto-updating page on global high-level Wikipedia statistics [24]. Statistics from [27] and [28] are based on recent database dump files and contain a note that *"the lengthy dump process (many weeks) means [that] a delay in publishing these statistics is always to be expected"*. We accessed the statistics on December 13, 2013, where the data was processed up until October 31, 2013. Based upon the fact that these database dumps of all Wikipedias are also publicly available, in 2005 Voss [22] gave an overview on Wikipedia research and analyzed articles, authors, edits, and links, as well as content and quality. A similar study on Wikipedia research dating from 2006 has been conducted by Ayers [3]. Unlike the methods described in [3, 22, 27, 28], our approach works in realtime. In [5], Ciampaglia and Vancheri perform an empirical analysis of user participation excluding bots in five large Wikipedias. Adler *et al.* study in [1] various means to measure user contributions to Wikipedia. As a part of their study, they analyze bot editors that, if they were included, created massive outliers in the intro-

[16]ClueBot NG: `http://en.wikipedia.org/wiki/User:ClueBot_NG`

duced measures. In contrast to the approaches [1, 5], we do not exclude bots for our analysis, but rather include them as equal citizens, while at the same time still allowing for dynamically disregarding them if need be. Geiger and Halfaker study in [7] bot statistics and the effect of the absence of a popular anti-vandalism bot on Wikipedia's quality control network. The tool *WikiChecker*[17] allows for near-realtime statistics about logged-in *vs.* anonymous users and more in the English, French, Russian, and Japanese Wikipedias. The application *Wikipulse* shows realtime absolute edit statistics for 36 Wikipedias, Wikimedia Commons, and Wikidata. We provide detailed realtime statistics for all Wikipedias and Wikidata. The *Wikipedia Recent Changes Map* tool, by LaPorte and Hashemi [13], inspired by a similar tool called *WikipediaVision* [11] by László Kozma, takes the IP addresses of anonymous editors of some of the biggest Wikipedias and maps them in realtime to geolocations via a lookup service. We make the same information available for all Wikipedias and Wikidata, however, do not geolocate the IP addresses. *Listen to Wikipedia* [12] is another project by LaPorte and Hashemi that makes Wikipedia and Wikidata edits audible and visualizes them in realtime. Depending on the size, the kind of edit, or the editor status, different sounds get played and colored circles displayed. The application *Wikistream*[18] by Ed Summers visualizes edits on all Wikipedias and Wikidata in realtime and uses recently uploaded images from Wikimedia Commons as background images, but does not keep track of edit statistics. *WikiTrip* [15] is an application developed by Massa *et al.* that provides *ad hoc* visualizations over time of several kinds of information about the Wikipedians who edited a selected page of any of the Wikipedias: their location in the world, their gender, and the distribution of logged-in *vs.* anonymous users. Several tools provide processed page view statistics based on database dumps, examples are *Wikipedia article traffic statistics*[19] by "User:Henrik", *Wikipedia article traffic statistics*[20] by "User:Emw", or finally *Wikipedia Page Views*[21] by Hannes Mühleisen. Mestyán *et al.* [16] measure and analyze the activity level of editors and viewers in order to predict the box office success of movies. We do not focus on page view statistics based on database dumps, but realtime article edit statistics instead. In [4], Boukhelifa *et al.* classify statistics tools in the categories global and local. According to this classification, local tools focus on individual articles or users and therefore require time-consuming on-the-fly computations. In contrast, global tools show the evolution of aggregated data for all Wikipedias, but without acquiring realtime data. According to this classification, our approach can be classified as global, however, with realtime support.

6. CONCLUSIONS AND FUTURE WORK

We have introduced an application and underlying API for the realtime monitoring of all 287 Wikipedias *and* Wikidata. To the best of our knowledge, this is the first study to look at Wikidata editing statistics. Via this application, which was also open-sourced under the Apache 2.0 license, we have collected more than 3.8 million edit events that we have analyzed in order to get a better understanding of Wikipedia and Wikidata editors. This allowed us to get a feeling of the relations of logged-in *vs.* anonymous edits and edits made by bots *vs.* edits made by humans globally and locally. We have looked at bots and their languages and created several analyses based thereon.

Figure 11: High-quality, geo-referenced multimedia content for disaster response (`http://commons.wikimedia.org/wiki/File:TerryPkwyGasStation.jpg`)

Future wok has several possible directions. On the engineering side, we want to improve the application such that it creates the diagrams shown in this paper directly. Further, we envision a global bot health check system that, based on the regular bot behavior patters, tries to detect outliers caused by potentially rogue bots or failing bots, so that we can alert administrators early on *"when the levee breaks"* [7]. On the research side, an interesting step to take is to look at qualitative differences between bot edits and human edits, and for the latter, logged-in and anonymous edits. While we now have a feeling for relative and absolute numbers, looking into the contents of the actual edits—which our API allows, as for each edit event it sends the `diffUrl` and `languageClusterUrl` (see Listing 1)—opens many research opportunities ranging from spam and vandalism detection [2] to realtime article tracking for monitoring events as they happen on Wikipedia [21]. Like with Twitter's Streaming APIs[22] that people have found creative uses for [17], our API can facilitate interesting use cases for data from Wikipedia and Wikidata like the detection of edit wars [26] reflecting real-world conflicts. In future versions of the Server-Sent Events API, we plan to allow developers to specify certain terms, articles, languages, categories, multimedia uploads, *etc.*, so that, rather than subscribe to the full firehose edit stream, they can track only a subset of articles they are interested in. This will be very useful for use cases like disaster response, where interested parties could subscribe to all sorts of edit events where the corresponding articles are categorized as describing disasters, or where high-quality, geo-tagged multimedia data as in Figure 11 gets added, which can help prioritize efforts of disaster responders. Another use case is brand perception monitoring, where brand owners could subscribe to all edit events mentioning the relevant brand terms. We also envision a feature similar to Web hooks [14], where rather than using the Server-Sent Events API directly, developers specify URLs where they want to be notified by the system once edit events occur. Finally, we want to implement support for the PubSubHubbub protocol [6], which allows for a scalable and distributed publish-subscribe architecture.

[17]WikiChecker: `http://en.wikichecker.com/`
[18]Wikistream: `http://wikistream.wmflabs.org/`
[19]Wikipedia article traffic stats: `http://stats.grok.se/`
[20]Wikipedia article traffic stats: `http://toolserver.org/~emw/wikistats/`
[21]Wikipedia Page Views: `http://wikistats.ins.cwi.nl/`
[22]Twitter Streaming APIs: `https://dev.twitter.com/docs/streaming-apis`

Concluding, we have contributed a useful Wikipedia and Wikidata monitoring tool as well as an open API and have performed an initial global study with both interesting and surprising insights that expectedly will be the first in a series of many more future studies and applications by us and hopefully others. The developed sample application and its openly licensed source code can serve as a reference for other academic researchers and even parties with commercial interests to build upon.

7. REFERENCES

[1] B. T. Adler, L. de Alfaro, I. Pye, and V. Raman. Measuring Author Contributions to the Wikipedia. In *Proceedings of the 4th International Symposium on Wikis*, WikiSym '08, pages 15:1–15:10, New York, NY, USA, 2008. ACM.

[2] E. Alfonseca, G. Garrido, J.-Y. Delort, and A. Peñas. WHAD: Wikipedia Historical Attributes Data. *Language Resources and Evaluation*, 47(4):1163–1190, 2013.

[3] P. Ayers. Researching Wikipedia – Current Approaches and New Directions. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–14, 2006.

[4] N. Boukhelifa, F. Chevalier, and J. Fekete. Real-time Aggregation of Wikipedia Data for Visual Analytics. In *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 147–154, Oct. 2010.

[5] G. L. Ciampaglia and A. Vancheri. Empirical Analysis of User Participation in Online Communities: the Case of Wikipedia. In *International AAAI Conference on Weblogs and Social Media*, pages 219–222, May 2010.

[6] B. Fitzpatrick, B. Slatkin, M. Atkins, and J. Genestoux. PubSubHubbub Core 0.4 – Working Draft, June 2013. https://pubsubhubbub.googlecode.com/git/pubsubhubbub-core-0.4.html.

[7] R. S. Geiger and A. Halfaker. When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes? In *Proceedings of the 9th International Symposium on Open Collaboration*, WikiSym '13, pages 6:1–6:6. ACM, 2013.

[8] L. Guldbrandsson. Swedish Wikipedia surpasses 1 million articles with aid of article creation bot, June 2013. http://blog.wikimedia.org/2013/06/17/.

[9] A. Halfaker, R. S. Geiger, J. T. Morgan, and J. Riedl. The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist*, 57(5):664–688, 2013.

[10] I. Hickson. Server-Sent Events. Candidate Recommendation, W3C, Dec. 2012. http://www.w3.org/TR/eventsource/.

[11] L. Kozma. WikipediaVision (beta), May 2013. http://www.lkozma.net/wpv/faq.html.

[12] S. LaPorte and M. Hashemi. Listen to Wikipedia, July 2013. http://blog.hatnote.com/post/56856315107/listen-to-wikipedia.

[13] S. LaPorte and M. Hashemi. Wikipedia Recent Changes Map, May 2013. http://blog.hatnote.com/post/49342528753/wikipedia-recent-changes-map.

[14] J. Lindsay. Web hooks to revolutionize the web, May 2007. http://progrium.com/blog/2007/05/03/web-hooks-to-revolutionize-the-web/.

[15] P. Massa, M. Napolitano, F. Scrinzi, and M. Ferron. WikiTrip: Animated Visualization over Time of Geo-location and Gender of Wikipedians Who Edited a Page. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, pages 40:1–40:9, New York, NY, USA, 2012. ACM.

[16] M. Mestyán, T. Yasseri, and J. Kertész. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. *PLoS ONE*, 8(8), 2013.

[17] S. Petrović, M. Osborne, and V. Lavrenko. Streaming First Story Detection with Application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computer Linguistics*, pages 181–189. Association for Computational Linguistics, 2010.

[18] A. Reinoso, J. Gonzalez-Barahona, R. Muñoz-Mansilla, and I. Herraiz. Temporal Characterization of the Requests to Wikipedia. In C. Lai, G. Semeraro, and E. Vargiu, editors, *New Challenges in Distributed Information Filtering and Retrieval*, volume 439 of *Studies in Computational Intelligence*, pages 71–89. Springer Berlin Heidelberg, 2013.

[19] L. Sanger. The Early History of Nupedia and Wikipedia: A Memoir. In C. DiBona, M. Stone, and D. Cooper, editors, *Open Sources 2.0: The Continuing Evolution*, pages 307–38. O'Reilly Media, 2005.

[20] T. Steiner. Bots vs. Wikipedians, Anons vs. Logged-Ins. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 547–548, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[21] T. Steiner, S. van Hooland, and E. Summers. MJ No More: Using Concurrent Wikipedia Edit Spikes with Social Network Plausibility Checks for Breaking News Detection. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 791–794. ACM, 2013.

[22] J. Voss. Measuring Wikipedia. In *10th International Conference of the International Society for Scientometrics and Informetrics*, July 2005.

[23] D. Vrandečić. Wikidata: A New Platform for Collaborative Data Collection. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 1063–1064. ACM, 2012.

[24] Wikipedia. Statistics, Dec. 2013. http://en.wikipedia.org/wiki/Special:Statistics.

[25] Wikipedia. Wikipedia:Statistics, Dec. 2013. http://en.wikipedia.org/wiki/Wikipedia:Statistics.

[26] T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertész. Dynamics of Conflicts in Wikipedia. *PLoS ONE*, 7(6), 2012.

[27] E. Zachte. Statistics Wikidata, Nov. 2013. http://stats.wikimedia.org/wikispecial/EN/TablesWikipediaWIKIDATA.htm.

[28] E. Zachte. Wikipedia Statistics, Nov. 2013. http://stats.wikimedia.org/EN/Sitemap.htm.