Wikipedia Academy 2012, Berlin, Germany, July 2012.
http://wikipedia-academy.de

On the Evolution of Quality Flaws and the Effectiveness of Cleanup Tags in the English Wikipedia

Maik Anderka Benno Stein Matthias Busse

Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

ABSTRACT

The improvement of information quality is a major task for the free online encyclopedia Wikipedia. Recent studies targeted the analysis and detection of specific quality flaws in Wikipedia articles. To date, quality flaws have been exclusively investigated in current Wikipedia articles, based on a snapshot representing the state of Wikipedia at a certain time. This paper goes further, and provides the first comprehensive breakdown of the *evolution* of quality flaws in Wikipedia. We utilize cleanup tags to analyze the quality flaws that have been tagged by the Wikipedia community in the English Wikipedia, from its launch in 2001 until 2011. This leads to interesting findings regarding (1) the development of Wikipedia's quality flaw structure and (2) the usage and the effectiveness of cleanup tags. Specifically, we show that inline tags are more effective than tag boxes, and provide statistics about the considerable volume of rare and non-specific cleanup tags. We expect that this work will support the Wikipedia community in making quality assurance activities more efficient.

Keywords: Wikipedia, Cleanup Tags, Quality Flaws, Information Quality, Quality Flaw Evolution

1. INTRODUCTION

The assessment of information quality in Wikipedia is a topic of enormous interest. This is witnessed by a large body of relevant research. A good deal of the existing research targets the classification of Wikipedia articles into predefined quality schemes, for instance, "Is an article featured or not?" [11, 17, 10, 21, 5, 6, 13]. However, these approaches are not designed to provide a rationale governing the respects in which an article violates Wikipedia's featured article criteria, and hence they provide virtually no support for Wikipedia's quality assurance activities. More recent studies address this issue by targeting the analysis and the automated detection of specific *quality flaws* in Wikipedia [2, 3, 4]. These approaches take advantage of the fact that Wikipedia users who encounter some flaw can tag the article with a so-called *cleanup tag*, see Figure 1. The existing cleanup tags can be used to compile the set of quality flaws that have so far been tagged by Wikipedia users. It has been shown that one in four English Wikipedia articles is tagged to contain at least one quality flaw [4]. These findings, however, relate to the state of Wikipedia at a certain time, because the previous studies investigate quality flaws only in the "current" revisions of Wikipedia articles. Though these studies give interesting insights into Wikipedia's quality flaw situation, they show only a very small snapshot of Wikipedia's history.

This paper gives the first comprehensive breakdown of the evolution of quality flaws in Wikipedia. In contrast to previous work, we do not restrict our analysis to a particular snapshot, but investigate the entire history of the English Wikipedia. In particular, we analyze the occurrence of cleanup tags in the revisions of Wikipedia articles. The benefits of this approach are twofold: First, it shows how the incidence and the extent of (tagged) quality flaws have evolved. Second, it shows how the way the Wikipedia community perceives and handles quality flaws has changed over time. We address the following concrete research questions:

RQ1. *When did the first cleanup tags emerge, and how have the number and the kind of tags changed over time?* The age, the number, and the kind of cleanup tags give some indication of the importance and the scope of the respective quality flaws. Moreover, we expect the absolute number of cleanup tags to become stable at some (future) point, when each possible flaw is covered by some tag.

RQ2. *Has the frequency, the type, and the distribution of tagged quality flaws changed over time?* This question relates to the usage of cleanup tags. We expect that certain cleanup tags have been widely used, whereas others have been used infrequently or even not at all. This gives some indication of the benefit of a certain tag for the Wikipedia community.

RQ3. *How fast get tagged quality flaws fixed?* We expect that certain flaws get fixed faster than others; consider for instance a broken link, in contrast, to an article that is not written from a neutral point of view. In this regard, the mean fix time (or the mean survival time respectively) of a flaw gives some indication of the flaw's complexity.

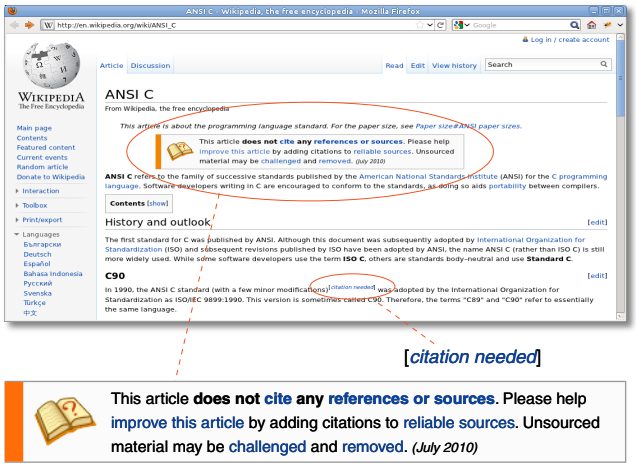


Figure 1: The Wikipedia article "ANSI C" with two cleanup tags. The tag box *Unreferenced* refers to the whole articles and the inline tag *Citation needed* refers to a particular claim. [4]

To answer the above research questions we investigate the revision history of the English Wikipedia from its launch in January 2001 until September 2011. Our analysis comprises all 412 477 496 revisions of the 24 931 064 pages that existed in September 2011. The contents of all revisions sum up to 7.3TB, which requires a sophisticated approach to access and process the relevant data. We use the Wikipedia database backup dumps provided by Wikimedia as the basis for our analyses, and process the dumps on a Hadoop cluster using Google's MapReduce.

The contributions of this paper can be summarized as follows:

1. We present an comparative overview of different approaches to access the Wikipedia database. The overview may serve as a reference for researchers and practitioners to help them choose the proper approach for the respective task at hand.
2. We reveal five cleanup tags that have not been used at all, and 15 cleanup tags that have been used less than once per year. These tags should be deleted because the benefit for the Wikipedia community is marginal.
3. We also reveal ten cleanup tags that have been used, but the tagged flaws have never been fixed. These tags should be redefined or replaced by others tags because they are either too unspecific or too complex.
4. We show that inline tags are more effective than tag boxes. The percentage of inline tags increased since 2005, and the average fixing time of a flaw that has been tagged with an inline tag is 73,35% compared to a flaw that has been tagged with a tag box.

We are confident that our findings provide valuable insights for the Wikipedia community and help to make future quality assurance activities more effective. Moreover, our findings can prove beneficial not only to Wikipedia but also to other Wiki-based projects and to user-generated content in general.

The reminder of this paper is organized as follows. Section 2 gives some background on quality assessment in Wikipedia and discusses related work. Section 3 gives a breakdown of Wikipedia's current quality flaw situation. Section 4 gives an overview of different approaches to access the Wikipedia database and describes the methods we used to process the database dumps and to identify tagged revisions. Section 5 presents the results of our analyses and discusses them with respect to our research questions. Finally Section 6 concludes this paper and gives an outlook on future work.

2. RELATED WORK AND BACKGROUND

There is a large body of research targeting the assessment of information quality in Wikipedia. Information quality is a multi-dimensional concept and combines criteria such as accuracy, reliability, timeliness, objectivity, completeness, and relevance. Wikipedia has been compared to other encyclopedias with respect to individual quality criteria, including accuracy [9, 14] and formality of language [7]. However, most of the studies investigate only a small sample of articles or target a particular topic. A widely accepted interpretation of information quality is the "fitness for use in a practical application" [20], i.e., the assessment of information quality requires the consideration of both context and use case. In Wikipedia the context is well-defined, namely by the encyclopedic genre, and the information quality ideal of the English Wikipedia has been formalized within the so-called featured article criteria.[1]

A large part of existing research on quality assessment in Wikipedia deals with featured article identification. Therefore, articles are analyzed with respect to the number of edits and editors [11, 21], the mutual dependency between article quality and author authority [10], the number of words [5], the character trigrams distribution [13], and particular combinations of several article features [6, 17]. As already motivated, the practical support for Wikipedia's quality assurance process is marginal, as these approaches provide no indication of the shortcomings of non-featured articles.

Recently, the usage of cleanup tags has been proposed to compile the set of quality flaws that exist in Wikipedia [2]. This approach takes advantage of the fact that Wikipedia users who encounter some flaw can tag the article with a cleanup tag, see Figure 1. The authors of [2] investigate a specific subset of 70 cleanup tags in a snapshot of the English Wikipedia from January 16, 2010. They find among other that 8.52% of the nearly 3 million articles in the snapshot have been tagged to contain at least one of the 70 flaws. Moreover, the tagged articles are used as a source of human-labeled data, which is then exploited by a machine learning approach to predict flaws in untagged articles. In a follow-up paper the authors refine their quality flaw detection approach and evaluate its accuracy using the English Wikipedia snapshot from January 15, 2011 [3]. In this paper, we do not target the detection of quality flaws. Nevertheless, the existing detection approaches can benefit from our findings; for instance, by using only recently tagged article revisions to train the respective machine learning approaches. This avoids the problem of "outdated" training date, particularly if a flaw with a high fixing time is to be detected.

The work most closely related to this paper is [4]. The authors give a comprehensive breakdown of quality flaws in the English Wikipedia, based on cleanup tags. They identify 388 quality flaws and analyze the distribution of tagged articles over Wikipedia's namespaces, main topics, and flaw types among others. However, in contrast to this paper, the analyses in [4] are restricted the a snapshot of Wikipedia. Gaio et al. [8] were the first who investigate the usage and the effectiveness of a cleanup tag over a certain time period. Their analyses target the cleanup tag *Complex* in SimpleWiki, a relative small language edition of Wikipedia that is written in basic English. In a follow-up study the authors apply similar analyses to the English Wikipedia and investigate the cleanup tag *NPOV* (neutral point of view) [15]. They find among others that editing increases after an article has been tagged. However, the findings of Gaio et al. relate only to a single quality flaw. Our analyses comprise all existing cleanup tags, and hence the entire set of quality flaws that have so far been tagged by Wikipedia users.

3. CURRENT QUALITY FLAW SITUATION

Before we go on to deal with the evolution of quality flaws in Wikipedia it is useful to give a breakdown of the current quality flaw situation. This helps to understand the methods, and also provides the basis for the subsequent analyses.

We employ the same approach as in [4] to compile the set of cleanup tags from a Wikipedia snapshot. We use the English Wikipedia snapshot from September 1, 2011, which is provided by the Wikimedia Foundation.[2] The snapshot comprises most of the tables of the Wikipedia database in the form of SQL dumps, totaling about 40GB. (We will discuss the dumps in more detail in the next section.) In a preprocessing step, we create a local (partial) copy of the Wikipedia database by importing the SQL dumps into a MySQL database. The local copy allows for efficient analy-

[1]Featured article criteria: `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria`.

[2]`http://dumps.wikimedia.org/enwiki/20110901`.

Table 1: Quality flaw situation in the English Wikipedia snapshot from September 11, 2011. Each row contains the number of flaws that belong to the respective flaw type along with the number of articles that have been tagged with these flaws. The values are given separately for the two scopes: article flaws (first multicolumn) and inline flaws (second multicolumn); the third multicolumn summarizes the values.

Flaw type	Article flaws		Inline flaws		$\sum$	
	Flaws	Articles	Flaws	Articles	Flaws	Articles
Verifiability	43	501 387	53	298 953	96	730 580
Wiki tech	23	178 538	2	14 096	25	191 927
General cleanup	19	72 874	0	0	19	72 874
Expand	12	64 963	4	244	16	65 197
Unwanted content	35	54 063	8	2 578	43	56 376
Style of writing	48	25 446	24	20 433	72	45 092
Neutrality	29	17 569	10	1 919	39	19 342
Merge	6	14 545	0	0	6	14 545
Specific subject	42	7 330	6	2 415	48	9 702
Structure	14	7 784	0	0	14	7 784
Time-sensitive	6	5 695	5	1 698	11	7 345
Miscellaneous	47	2 022	8	107	55	2 127
$\sum$	324	952 216	120	342 443	444	1 020 052

ses, without causing traffic on the Wikimedia servers. The cleanup tags are compiled in two-steps. At first, an initial set of cleanup tags is extracted from a Wikipedia administration category, which comprises templates that are used for tagging articles as requiring cleanup,[3] and from a Wikipedia meta page, which comprises a manually maintained listing of templates that may be used to tag articles as needing cleanup.[4] Afterwards the initial set is further refined by resolving redirects and discarding subtemplates as well as meta-templates. Altogether we extract a set of 444 cleanup tags, each of which defines a particular quality flaw.

To breakdown the current quality flaw situation we manually divided the 444 flaws into 12 general quality flaw types, which have been proposed in [4]. Moreover, we quantify the scope of a flaw, and distinguish *article flaws* that refer to the whole article and *inline flaws* that refer to a certain text fragment (see Figure 1). The respective criteria to divide the flaws into the types and to quantify the scope of a flaw are the same as in [4]. Table 1 shows the results of our breakdown. Altogether 1 020 052 articles have been tagged with at least one of the 444 quality flaws, which corresponds to 27.17% of all articles in the snapshot. The majority (71.62%) of the tagged articles have been tagged with a flaw that belongs to the flaw type *Verifiability*. The 96 flaws that belong to this type refer among others to unsourced statements and to articles with inadequate and invalid references. Though a relative small number of 25 flaws belong to the type *Wiki tech*, a considerable amount (18.82%) of the tagged articles have been tagged with one of this flaws. The flaw type *Wiki tech* targets technical aspects of an article, including categorization issues, syntactical problems, and connectivity in terms of Wikipedia-internal links. Note that the individual frequencies of tagged articles per flaw type do not sum up to the respective total frequencies in the last row of Table 1. This is due to the fact that some articles are tagged with multiple flaws (multi-labeling). Analogously, the individual frequencies of tagged articles per scope do not sum up to the respective total frequencies in the last column of the table. From the 444 flaws, 324 are article flaws and 120 are inline flaws. The flaw type *Verifiability* is the only type with more inline flaws than article flaws. Altogether, 93.35% of tagged articles are tagged with an article flaw and 33.57% are tagged with an inline flaw. (Note that some articles are tagged with both article flaws and inline flaws.)

[3]Category "Cleanup templates": `http://en.wikipedia.org/wiki/Category:Cleanup_templates`.

[4]Page "Template messages/Cleanup": `http://en.wikipedia.org/wiki/Wikipedia:Template_messages/Cleanup`.

4. METHOD

Analyzing the evolution of quality flaws requires an investigation of Wikipedia's revision history. The English Wikipedia has been launched in January 2001, and since then it has received more than 524 million edits.[5] Each edit produces a new revision of the edited page. The complete revision history is stored in the Wikipedia database. The huge amount of data and the fact that a direct unrestricted access to the database is in general not available pose particular challenges for a comprehensive and efficient analysis. We therefore start with a comparative overview of different approaches to access the Wikipedia database (Section 4.1). The most promising approach for the analyses in this paper is using the database dumps provided by Wikimedia. We describe the relevant dumps and our procedure to process the dumps (Section 4.2). Afterwards we present our approach to identify those revisions that have been tagged with some cleanup tag (Section 4.3).

4.1 Accessing the Wikipedia Database

In the following we target the English Wikipedia, however, the described approaches apply to other language versions as well. Figure 2 depicts different possibilities to access the Wikipedia database. The MediaWiki software provides three basic approaches: First, the standard Web interface, which is implemented in the main script of the MediaWiki software `index.php`.[6] Second, the MediaWiki Web service API, which is implemented in `api.php` and provides access for scripts and bots.[7] Third, the page *Special:Export*, which uses `dumpBackup.php` to export pages to XML.[8] The database backup dumps are another (indirect) way to access the Wikipedia database.[9] Backup dumps are compiled in regular time intervals, and correspond to a snapshot of the Wikipedia database at a certain time. Further the Wikimedia Toolserver provides access to a replication of the Wikipedia database.[10] The Toolserver is operated by Wikimedia Deutschland and hosts various software tools that are related to the Wikimedia projects.

Since each of the five approaches has advantages and disadvantages, we have compiled seven criteria to assess and compare the approaches, see Figure 2 (right-hand side). All approaches provide read access to the Wikipedia database, however, only the Web interface and the MediaWiki API also provide write access. The three MediaWiki methods are suitable to access smaller amounts of data, e.g., to get certain meta information on a particular set of pages. These methods do not scale when a huge amount of data is required, e.g., the revision history of all pages. The Wikimedia servers are not meant to handle this kind of queries, and moreover, there are several limitations that make it difficult to query large amounts of data, e.g., the maximum number of page revisions that are returned by the MediaWiki API is 500 for a single query. The database dumps are most suitable for comprehensive analyses of Wikipedia, which concern for instance the revision history or the link graph.

[5]List of official Wikipedias: `http://meta.wikimedia.org/wiki/List_of_Wikipedias`.

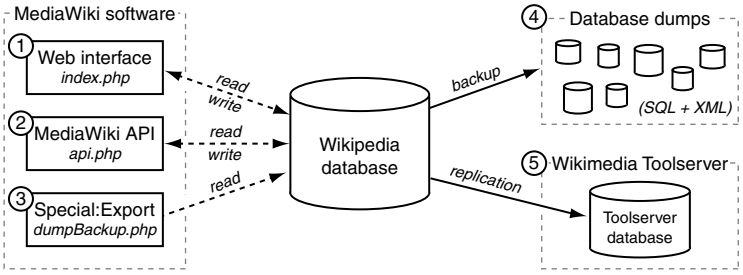
[6]`http://www.mediawiki.org/wiki/Manual:Index.php`.

[7]`http://www.mediawiki.org/wiki/API:Main_page`.

[8]`http://meta.wikimedia.org/wiki/Help:Export`.

[9]Wikimedia downloads: `http://download.wikimedia.org`.

[10]Wikimedia Toolserver: `http://www.toolserver.org`.



Criterion	Approaches				
	(1)	(2)	(3)	(4)	(5)
Read access	✓	✓	✓	✓	✓
Scalable	×	×	×	✓	○
Page content	✓	✓	✓	✓	×
Open access	✓	✓	✓	✓	○
Different formats	×	✓	✓	×	×
Current data	✓	✓	✓	×	○
Write access	✓	✓	×	×	×

Figure 2: The figure on the left-hand side shows five different approaches to access the Wikipedia database: the Web interface, the MediaWiki API, the page *Special:Export*, the database dumps, and the Wikimedia Toolserver. The table on the right-hand side compares the five approaches using seven criteria, which are quantified as follows: the criteria applies (✓), it applies only partially (○), and it does not apply (×). The first four criteria are relevant for the analyses in this paper.

The dumps can be processed locally using any scalable technology, like MapReduce and Hadoop. The scalability of the Toolserver is given only partially because the resources must be shared with all users. All approaches provide access to the content of the Wikipedia pages except the Toolserver; the table "text" is not considered in the Toolserver's replication process. The MediaWiki methods and the database dumps are publicly available, whereas using the Toolserver requires an account that has to be requested providing a justification and need to be approved by Toolserver staff. The MediaWiki API and the *Special:Export* page provide various output formats, including JSON, serialized PHP, and XML, whereas the other approaches have a predefined output format. The MediaWiki methods provide access to the current Wikipedia data. As already mentioned, the database dumps represent the Wikipedia database at a certain time. The Toolserver database can not be considered a real-time copy of the original Wikipedia database because of replication lags (typically only a few seconds), so that the Toolserver database may represent a state at some point in the past.

In this paper we use the database dumps. Since we want to analyze the content of each page revision in the English Wikipedia, scalability and the availability of the page content are the most determining criteria. Moreover, the open access to the data ensures the reproducibility of our results. A specific output format is not required, and we need neither access to the latest Wikipedia data nor write access to the database.

4.2 Processing the Database Dumps

Analyzing Wikipedia's entire revision history using the database dumps requires an efficient approach to process the huge amount of data. A Wikimedia snapshot contains two kinds of database dumps, SQL dumps and XML dumps. We already used the SQL dumps in Section 3 to create a local (partial) copy of the Wikipedia database, which is used to analyze the current Wikipedia data. However, the SQL dumps contain no revision history information because the relevant Wikipedia database tables "revision" and "text" are not dumped directly.[11] The data of these two tables is provided as XML dumps, which use the same XML wrapper format that *Special:Export* produces for individual pages.[12] In order to augment our local copy of the Wikipedia database, we retrieve the relevant information to create the table "revision" from the XML dump *stub-meta-history*. The dump contains meta information for each revision, including the page identifier, the size of the revision in byte, the user who performed the edit, and an optional editor comment. The size of the uncompressed dump amounts to 120.9GB. We processed the dump on a Hadoop cluster with 39 nodes using MapReduce. In particular, we implement a respective XML parser in the Map phase, which parses the relevant revision information from each *page* element.

We also need to analyze the revisions' content, to identify tagged revisions (this will be described in the next subsection). The XML dump *pages-meta-history* contains the raw wikitext markup of every revision. The size of the uncompressed dump amounts to 7.3TB. We processed this dump on the Hadoop cluster as well. However, we had to adapt our parsing approach because there are pages with more than 100 000 revision, which cannot be handled efficiently by a single Map job. We therefore implemented a dedicated Wikipedia revision input format, so that one revision is processed by a single Mapper. The Mapper parses the wikitext markup and extracts the relevant information. Parsing the 7.3TB lasts about five hours. Due to the large amount of data and because we are only interested in tagged revision, we do not create the complete "text" table, but store only the relevant information in our local copy of the Wikipedia database.

4.3 Identifying Tagged Revisions

Our goal is to analyze the occurrence of cleanup tags in the revisions of Wikipedia articles. The analyses are based on the 444 cleanup tags that have been identified in the articles' current revisions (see Section 3). We do not consider cleanup tags that may have existed at some point in the past and that have been deleted.

In a first step, we identify for each cleanup tag those revisions where the tag occurs. We therefore process the *pages-meta-history* dump as described above. Cleanup tags are included in a Wikipedia page using the template syntax:[13]

```
{{template name|parameter 1|parameter 2|...}}
```

The template name corresponds to the page title of the respective cleanup tag. However, a cleanup tag may have several alternative titles linking to it through redirects. For example, the cleanup tag *Unreferenced* has the redirects *Unref*, *Noreferences*, and *No refs* among others. We resolve all redirects using the tables *redirect* and *page* of the local Wikipedia database, and parse the wikitext markup of each revision using regular expressions. Altogether, 97.4 million tagged revisions are identified, which corresponds to 23.61% of all revisions. The majority (92.4%) of the tagged revisions belongs to articles. As motivated previously, this paper targets quality flaws in articles, and hence we discard the revisions of non-article pages.

[11]For details refer to the page `http://meta.wikimedia.org/wiki/Data_dumps#What_happened_to_the_SQL_dumps.3F`.

[12]Schema for the XML dumps: `http://www.mediawiki.org/xml/export-0.3.xsd`.

[13]Help page for templates: `http://en.wikipedia.org/wiki/Help:Template`.

We also identify sequences of tagged revisions, which is of particular interest to investigate research question RQ3. We therefore traverse the revisions of a single article in chronological order using the timestamps provided by the *revision* table. Given the set of tagged revisions identified above, we investigate for each article and for each cleanup tag: the time at which the article has been tagged, the time at which the tag has been removed (i.e., the respective flaw has been fixed), and the duration that the article remained tagged. This approach, however, is prone to vandalism edits. Though vandalism is repaired relatively quickly by the Wikipedia community [19], the vandalized revisions distort our analyses. A common type of vandalism is, for example, the (partial) deletion of the article content. If an existing cleanup tag is also deleted, our approach incorrectly assumes that the vandalism edit has fixed the respective quality flaw. To prevent such issues, we do not consider empty revisions in our analyses. Moreover, we discard revisions that stem from the following types of edits:

- *Anonymous edits*. It has been shown that the majority of vandalism edits is caused by anonymous editors, whereas the amount of serious anonymous edits is relatively small [1, 16]. Whether an edit has been made by an anonymous or an registered user is stored in the *user* field of the *revision* table, which contains either an IP address (anonymous) or an user name (registered).
- *Vandalism edits identified by users*. It has become common practice in the Wikipedia community to use dedicated keywords in the editor comment if a vandalism edit has been reverted. These keywords are "revert", "rv" (revert), "rvv" (revert due to vandalism), "vandalism", "spam", "undid", and "rollback". We consider an edit as vandalism if the editor comment of the subsequent revision contains one of the mentioned keywords or some combination respectively. A similar approach is used in [18]. The editor comments are contained in the *revision* table.
- *Vandalism edits identified by anti-vandalism bots*. Anti-vandalism bots are programs that operate under a common Wikipedia user account and repair certain types of vandalism autonomously. We consider an edit as vandalism if it has been reverted by one of the eleven bots listed in a respective Wikipedia category[14]. These edits are identified by the bot's user names.

We are confident that we identify the majority of vandalized revisions using the mentioned heuristics. Note in this respect that vandalism detection is a separate research area, see for instance [12], and that the detection of vandalism in general is beyond to scope of this paper. Without considering vandalism edits, we identify 125.7 million article revisions, of which 50.5 million have been tagged with at least one of the 444 cleanup tags. These revisions are the basis for the analyses described in the next section.

5. ANALYSIS AND RESULTS

This section presents and discusses the results of our analysis, focusing on the research questions stated in the introduction. We will not show the individual results for each of the 444 cleanup tags, this is beyond the scope of this paper and would also be less insightful. Instead, we use the organization of the cleanup tags into scopes and quality flaw types, which is described in Section 3. Nevertheless, we will discuss interesting particularities of a few individual cleanup tags where it is appropriate. As already mentioned, our analysis comprises the time period from January 2001 until September 2011. Note that in those cases where we present averaged values over a whole year, the data for 2011 is incomplete (the respective values are written in italics). Each of the following subsections addresses one of our three research questions.

[14]Wikipedia anti-vandalism bots: `http://en.wikipedia.org/wiki/Category:Wikipedia_anti-vandal_bots`.

5.1 Evolution of Cleanup Tags

RQ1. *When did the first cleanup tags emerge, and how have the number and the kind of tags changed over time?*

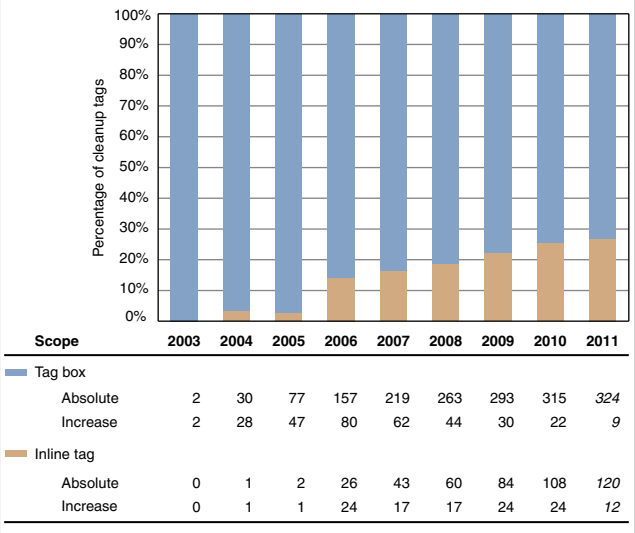
At first we investigate how the number of cleanup tags has evolved. The first cleanup tags were *POV* and *Disputed*, which have been created in December 2003. The former stands for "point of view" and refers to missing neutrality, the later refers to an article's factual accuracy. Since 2003 the number of cleanup tags grows steadily, see Table 2. A reason for the significant increase in 2006 might be the increasing popularity of cleanup tags due to the creation of the page *Template messages/Cleanup* in the previous year. This page comprises a manually maintained listing of cleanup tags; it was already used in Section 3 to compile the existing tags. More than the half of the 444 tags already existed in 2007. Since 2007, the number of newly created cleanup tags per year declines. In 2011 only 11 new cleanup tags were created until September 1. A possible explanation for this development might be that the majority of quality flaws are already covered by the existing cleanup tags. If this trend continues, a relative stable number of cleanup tags is to be expected within the next few years.

Table 2: The number of existing cleanup tags per year (absolute) and the number of newly created cleanup tags per year (increase).

	2003	2004	2005	2006	2007	2008	2009	2010	2011
Absolute	2	31	79	183	262	323	377	423	*444*
Increase	2	29	48	104	79	61	54	46	*11*

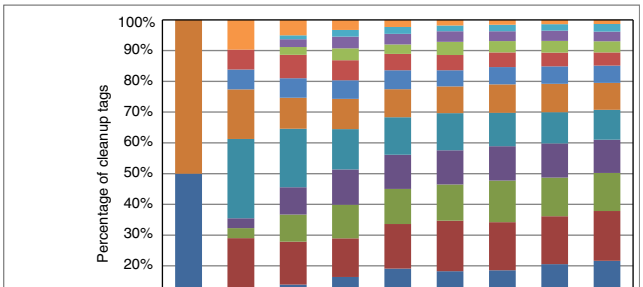
Figure 3 shows a break down into the scope of the cleanup tags. Though there were already 77 tag boxes at the end of 2005, there were only 2 inline tags at this time. The first inline tags were *Dubious* and *Citation_needed*, which have been created in July 2004 and June 2005 respectively. The former tag is the inline version of the tag box *Disputed*, and refers to a specific statement or alleged fact that is subject to dispute. Since 2006 the number of newly created inline tags per year is nearly constant, whereas the number of newly created tag boxes declines. Consequentially, the percentage of inline tags has increased over the last years. This development might indicate that inline tags are more appropriate than tag boxes for tagging quality flaws. In general, inline tags are more specific than tag boxes. For instance, the tag box *Unreferenced* states that the article does not cite any references or sources. By contrast, the inline tag *Citation needed* gives a direct indication of a claim that needs to be referenced (see Figure 1). However, some flaws refer to the whole article per definition and hence it is not appropriate to use inline tags for these flaws.

Figure 4 shows the development of cleanup tags broken down into the 12 flaw types. In 2005 there is already at least one cleanup tag for each flaw type. Already in 2006 certain flaw types emerge that comprise more cleanup tags than other, including *Verifiability*, *Unwanted content*, *Style of writing*, *Neutrality*, *Specific subject*, and *Miscellaneous*. The first four types are of particular interest, as they correspond to fundamental properties of an encyclopedia. The relative high number of cleanup tags in these four types might be an indication that the respective quality flaws are considered as quite serious by the Wikipedia community.



Scope	2003	2004	2005	2006	2007	2008	2009	2010	2011
Tag box									
Absolute	2	30	77	157	219	263	293	315	*324*
Increase	2	28	47	80	62	44	30	22	*9*
Inline tag									
Absolute	0	1	2	26	43	60	84	108	*120*
Increase	0	1	1	24	17	17	24	24	*12*

Figure 3: Number of existing cleanup tags per year broken down into the two scopes *tag box* and *inline tag*. The plot shows the percentages; the table below shows the absolute values along with the increase compared to the respective previous year.



Flaw type	2003	2004	2005	2006	2007	2008	2009	2010	2011
Merge	0	3	4	6	6	6	6	6	*6*
Time-sensitive	0	0	1	4	6	6	8	9	*11*
Structure	0	0	2	7	9	11	12	14	*14*
Expand	0	0	2	7	8	14	14	16	*16*
General cleanup	0	2	6	12	14	16	18	19	*19*
Wiki tech	0	2	5	11	16	17	21	24	*25*
Neutrality	1	5	8	18	24	28	35	39	*39*
Unwanted content	0	8	15	24	32	39	41	43	*43*
Specific subjects	0	1	7	21	29	36	42	47	*48*
Miscellaneous	0	1	7	20	30	38	51	53	*55*
Style of writing	0	7	11	23	38	53	59	66	*72*
Verifiability	1	2	11	30	50	59	70	87	*96*

Figure 4: Number of existing cleanup tags per year broken down into the 12 flaw types. The plot shows the percentages; the table below shows the absolute values.

5.2 Evolution of Tagged Quality Flaws

RQ2. *Has the frequency, the type, and the distribution of tagged quality flaws changed over time?*

So far we have analyzed the number and the kind of cleanup tags. Now we analyze the usage of cleanup tags and thus the extent of (tagged) quality flaws. Table 3 shows the number of tagged quality flaws per year. The first flaw has been tagged in Mai 2004, in the article *Stage name* using the cleanup tag *Merge*. The number of tagged quality flaws grew steadily until 2011. Analog to the number of cleanup tags, the number of tagged flaws increased significantly in 2006. As already mentioned, one possible explanation for this development might be the creation of the page *Template messages/Cleanup*. In 2008 there were more than 1 million tagged quality flaws, however, this year had the lowest increase since 2004. Though the values for 2011 are incomplete because our analysis ends in September 1, 2011, these values show an interesting trend: The number of tagged quality flaws has decreased after the first nine months of 2011. This indicates that the number of tagged quality flaws that have been fixed was greater than the number of quality flaws that have been newly tagged. This also gives a first indication towards the usefulness of tagging, as it shows that tagged flaws are actually fixed at some time. (We will discuss this in more detail in the next subsection.) Altogether, more than 6 million quality flaws have been tagged since 2004.

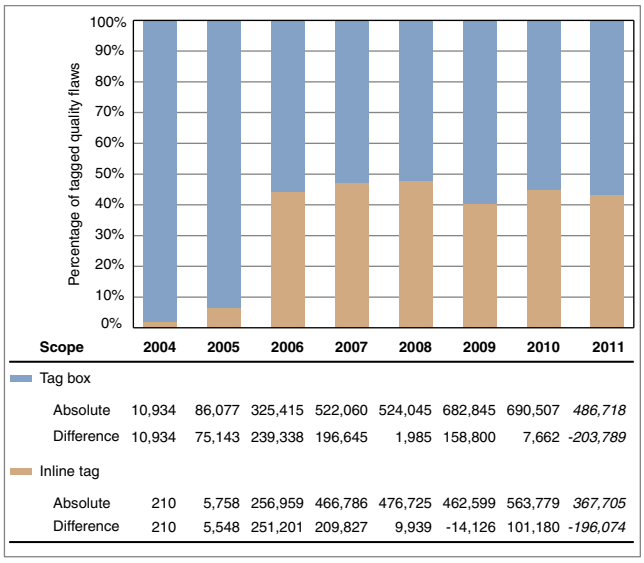
Figure 5 breaks down the development of tagged quality flaws into the scope of the tags. In 2004 and 2005 the vast majority of tagged flaws have been tagged using tag boxes, which is due to the fact that at this time only two inline tags existed (see Figure 3). In 2006 there was still a relative small number of 26 inline tags, compared to the 157 tag boxes that existed in this year, however, the percentages of tagged inline flaws rapidly increased to 44.12%. Since 2006 the percentage of inline tags remains relative constant between 40% and 45%. The absolute number of tagged flaws per year was subject to high variations, for both tag boxes and inline tags. The numbers of tagged flaws rapidly increased in 2006 and 2007. However, in 2008 the differences compared to the last year were relative small. Tag boxes rapidly increased again in 2009, but remained nearly constant in 2010. The absolute number of tagged inline flaws even decreased in 2009, but increased again in 2010. The small increases as well as the decrease must not mean that tagging activities declined in the respective years, instead this might be an indicator that a relative high number of tagged flaws have been fixed.

Figure 6 shows the development of tagged quality flaws broken down into the 12 flaw types. As expected, the number of tagged flaws per year increases in the majority of flaw types, following the overall trend. Most notably is the flaw type *Verifiability*; the number of tagged quality flaws has rapidly increased in 2006 and nearly doubled in 2007. Moreover, the *Verifiability* type contains the majority of tagged flaws per year since 2006. The flaw type *Expand* also shows an interesting development, the number of tagged flaws have more than tripled in 2010, although at this time a relative small number of 16 respective cleanup tags existed (see Figure 4). The flaw type *Expand* is also the only type with an increasing number of tagged flaws in the first nine month of 2011. Another rapid increase that pertains the flaw type *Wiki tech* occurred in 2009. The annual number of tagged flaws that belong to the type *Merge* decreases since 2007. This might indicate that the organization of the existing articles improved steadily, and thus fewer and fewer content need to be merged and reorganized.

Up to now we have analyzed the usage of cleanup tags by the number of tagged quality flaws averaged over the tags' scope and over the type of the respective flaw. Table 4 shows usage statistics for individual cleanup tags, whereas only the most widely used and the most least used tags are listed. The most common cleanup tag is *Citation_needed*, it has been used nearly 2 million times in the 2 268 days after its creation, which corresponds to an average ratio of 871.64 usages per day. This means that on average one in 200 revisions has been tagged with this cleanup tag. That a tag exists for a long time must not mean that it has been used more frequently com-

Table 3: The absolute number of tagged quality flaws per year and the difference to the respective previous year.

	2004	2005	2006	2007	2008	2009	2010	2011
Tagged flaws (absolute)	11 144	91 835	582 374	988 846	1 000 770	1 145 444	1 254 286	*854 423*
Difference to last year	+11 144	+80 691	+490 539	+406 472	+11 924	+144 674	+108 842	*-399 863*



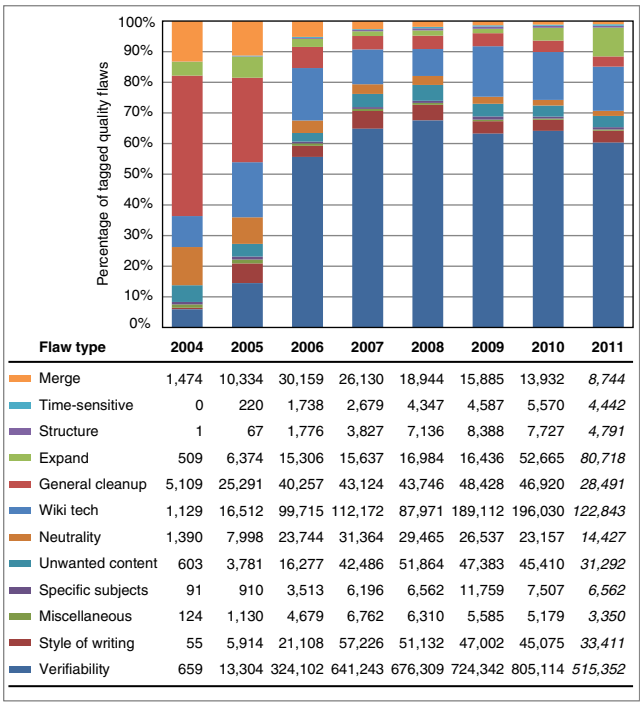
Scope	2004	2005	2006	2007	2008	2009	2010	2011
Tag box								
Absolute	10,934	86,077	325,415	522,060	524,045	682,845	690,507	*486,718*
Difference	10,934	75,143	239,338	196,645	1,985	158,800	7,662	*-203,789*
Inline tag								
Absolute	210	5,758	256,959	466,786	476,725	462,599	563,779	*367,705*
Difference	210	5,548	251,201	209,827	9,939	-14,126	101,180	*-196,074*

Figure 5: Number of tagged quality flaws per year broken down into the two scopes *tag box* and *inline tag*. The plot shows the percentages; the table below shows the absolute values and the differences to the respective previous year.

pared to a younger tag. Consider for instance the cleanup tags *Refimprove* and *Uncategorized*. The former exists since nearly three and a half year and has been used on average 149.02 times per day, whereas the latter exists since more than six years and have been used only 122.27 times per day. However, the 40 cleanup tags with the highest number of usages per day are all older than three years. Those cleanup tags that have been created before 2008 account for more than 95% of the ever tagged quality flaws. The lower part of Table 4 shows the other extreme. There are five cleanup tags that have not been used at all. Moreover, 15 cleanup tags have been used less than once per year, and additional 131 cleanup tags have been used less than once per month (not listed in the table).

Table 4: The most widely used and the most least used cleanup tags, along with the respective total number of instances, the age of the tag in days (from its creation until now), and the ratio of instances per day.

Cleanup tag	Instances	Age	Ratio
Citation_needed	1 976 868	2 268	871.64
Unreferenced	608 468	2 406	252.90
Refimprove	242 155	1 625	149.02
Dead_link	239 576	1 676	142.95
Uncategorized	296 755	2 427	122.27
...			
Title_incomplete	0	580	0.0
Time_references_needed	0	102	0.0
ShadowsCommons	0	1 883	0.0
Convert_to_SVG_and_copy_to_Wikimedia_Commons	0	1 483	0.0
Cat_nomore	0	705	0.0



Flaw type	2004	2005	2006	2007	2008	2009	2010	2011
Merge	1,474	10,334	30,159	26,130	18,944	15,885	13,932	*8,744*
Time-sensitive	0	220	1,738	2,679	4,347	4,587	5,570	*4,442*
Structure	1	67	1,776	3,827	7,136	8,388	7,727	*4,791*
Expand	509	6,374	15,306	15,637	16,984	16,436	52,665	*80,718*
General cleanup	5,109	25,291	40,257	43,124	43,746	48,428	46,920	*28,491*
Wiki tech	1,129	16,512	99,715	112,172	87,971	189,112	196,030	*122,843*
Neutrality	1,390	7,998	23,744	31,364	29,465	26,537	23,157	*14,427*
Unwanted content	603	3,781	16,277	42,486	51,864	47,383	45,410	*31,292*
Specific subjects	91	910	3,513	6,196	6,562	11,759	7,507	*6,562*
Miscellaneous	124	1,130	4,679	6,762	6,310	5,585	5,179	*3,350*
Style of writing	55	5,914	21,108	57,226	51,132	47,002	45,075	*33,411*
Verifiability	659	13,304	324,102	641,243	676,309	724,342	805,114	*515,352*

Figure 6: Number of tagged quality flaws per year broken down into the 12 flaw types. The plot shows the percentages; the table below shows the absolute values.

5.3 Fix Time of Quality Flaws

RQ3. *How fast get tagged quality flaws fixed?*

So far, it was sufficient to investigate the date at which the instances of the cleanup tags have been placed in the articles. For the analyses in this section, we investigate also the date at which the instances have been removed, i.e., when the respective flaws have been fixed by the Wikipedia community. However, not all of the tagged flaws have yet been fixed. Table 5 shows the total number of fixed flaws as well as the percentage of fixed flaws (FTR) broken down into the 12 flaw types. Altogether, more than 4 million flaws have been fixed, which corresponds to 69.7% of the flaws that have ever been tagged. The time period between tagging date and fixing date is considered as the flaw's fixing time. We analyze the fixing time for the instances of all tagged quality flaws, which gives us the average fixing time (AFT) for each flaw type. The overall AFT is 146.8, i.e., the 4 132 524 flaws have been fixed on average after 146.8 days. However, 33.9% of the flaws have been fixed within the first week, which is indicated by the respective FWR value. Quality flaws that belong to the type *Expand* have the highest average fixing time and the lowest FWR value. This and the relative low FTR value might be an indication that these flaws are too complex. The majority (60.25%) of the fixed flaws belong to the flaw type *Verifiability*, whose AFT and FWT values nearly match the respective averaged values over all types. Quality flaws that belong to the types *Wiki tech* and *Miscellaneous* have been fixed relatively quickly, and nearly 40% within the first week.

Table 5: For each of the 12 flaw types: the total number of times the respective flaws have been fixed, the percentage of fixed flaws relative to the number of tagged flaws (fixed tagged ratio, FTR), the average fixing time in days (AFT), and the percentage of fixed flaws that have been fixed within the first week (first week ratio, FWR).

Flaw type	Fixed flaws	FTR	AFT	FWR
Expand	100 624	49.2	220.7	24.9
General cleanup	201 116	71.5	203.3	26.9
Specific subject	27 922	64.8	175.2	25.1
Style of writing	203 901	78.1	164.4	29.1
Verifiability	2 489 970	67.3	155.8	33.5
Time-sensitive	15 329	65.0	154.7	28.3
Structure	25 442	75.5	137.2	33.0
Unwanted content	175 639	73.5	122.1	37.4
Merge	109 728	87.4	111.9	31.2
Neutrality	136 622	86.4	107.2	43.7
Wiki tech	615 575	74.6	97.5	38.8
Miscellaneous	30 656	92.6	96.3	38.7
$\sum$	4 132 524	69.7	146.8	33.9

Table 6: For each of the 12 flaw types and for the two scopes article flaws and inline flaws: the total number of times the respective flaws have been fixed and the average fixing time in days (AFT).

Flaw type	Article flaws		Inline flaws	
	Fixed flaws	AFT	Fixed flaws	AFT
Expand	99 037	222.6	1 587	101.3
General cleanup	201 116	203.4	0	0
Specific subject	21 560	185.8	6 362	139.1
Style of writing	154 607	181.4	49 294	111.4
Verifiability	741 600	231.3	1 748 370	123.7
Time-sensitive	12 812	156.7	2 517	144.7
Structure	25 442	137.3	0	0
Unwanted content	152 823	127.6	22 816	85.8
Merge	109 728	111.8	0	0
Neutrality	118 668	108.1	17 954	101.4
Wiki tech	588 858	97.1	26 717	108.4
Miscellaneous	28 740	99.6	1 916	46.6
$\sum$	2 254 991	167.0	1 877 533	122.5

Table 6 breaks down the average fixing time as well as the total number of fixed flaws into the scope of the cleanup tags. The total number of fixed flaws differs only slightly for the two scopes, article flaws account for 55.6% of all fixed flaws. However, the vast majority (93.12%) of fixed inline flaws belongs to the flaw type *Verifiability*. Nevertheless, the average fixing time of inline flaws is significantly smaller for the majority of flaw types, except for the type *Wiki tech*, which has a larger AFT value for inline flaws. The *Verifiability* type has the largest difference regarding the average fixing time of article flaws and inline flaws. As already mentioned, inline flaws are more specific, so that a potential corrector has a concrete indication of what need to be done to fix the flaw. This applies especially for flaws that belong to the *Verifiability* type. It is, for instance, easier to find a reference for a certain statement, than for a latent idea that is mentioned somewhere in an article. Altogether, inline flaws have been fixed faster than article flaws, the average fixing time is 122.5 days and 167 day respectively.

Table 7 shows the fixing statistics for individual quality flaws. The tag box *Hoax* states that an article contains false facts. The fact that this flaw has been fixed after 6.4 day on average might indicate that hoaxes are considered quite serious by the Wikipedia community. However, a high AFT must not mean that the respective flaw is regarded as unimportant. The AFT can be considered as measure for a flaw's complexity. Consider for instance the article flaws *Uncategorized* and *Orphan* (states that an article has too few incoming links). It is relative easy to find a category for an uncategorized article, but finding related articles for an orphaned article and defining reasonable links is much more complicated. In some cases, those related articles are yet to be created. Hence the AFT is relative high for the *Orphan* flaw. Despite its high complexity, this flaw is considered as important, which is witnessed by the large number of 113 112 fixed flaws. A possible way to increase the AFT of such complex, but still important, flaws is to provide respective tools that support potential correctors. In the case of the *Orphan* flaw this might be a tool that retrieves articles with similar content. There are also ten flaws that have never been fixed at all: *Ideail*, *List_years*, *Kmposts*, *Lacking_overview*, *Off_topic_sentence*, *Clarify-span*, *Tertiary*, *Author_incomplete*, *Cleanup-lang*, and *RJL*. A possible explanation is that these flaws are either too unspecific or too complex.

Table 7: The quality flaws with the shortest and the longest average fixing time in days (AFT) respectively, along with the total number of times the flaw has been fixed. Only those flaws are listed that have been fixed more than 1 000 times.

Flaw name	AFT	Fixed flaws
Hoax	6.4	1 528
Not English	10.1	3 156
Image requested	19.4	2 335
Uncategorized	23.2	294 706
Disambiguation cleanup	23.3	8 066
...		
Expand section	257.6	75 442
Citations missing	263.4	11 164
Orphan	268.7	113 112
Cleanup FJ biography	324.8	1 574
Cleanup-school	325.1	1 036

6. CONCLUSIONS

This paper investigates quality flaws in the English Wikipedia, whereas, for the first time, the entire revision history of all articles is considered. We analyze the time period from January 2001 until September 2011, which comprises 412 477 496 revisions, whose content sum up to 7.3TB. A comparative analysis of different approaches to access the Wikipedia database reveals that the Wikimedia database dumps are most appropriate for analyzing the evolution of quality flaws. We use the SQL dumps to establish a local copy of the Wikipedia database, which provides the relevant meta information. Moreover, we process the XML dumps on a Hadoop cluster using MapReduce to identify those revisions that have been tagged with certain cleanup tags. Our analysis is based on the 444 cleanup tags that existed in September 2011.

Our findings yield the following conclusions for the Wikipedia community:

1. If possible, users should use inline tags rather than tag boxes to tag quality flaws, because inline flaws are fixed faster. According to this, it should be investigated whether existing tag boxes can be redefined as inline tags.
2. Those cleanup tags that have been used not at all or very infrequently should be deleted, because they provide no benefits and may distract users.

3. Those cleanup tags that have never been fixed should be redefined, because they are too unspecific or too complex.

4. Several cleanup tags have a high fixing time but still address relevant flaws. Dedicated tools should be developed that support potential correctors, so that these flaws can be fixed faster.

Our analysis also reveals that the number of newly created cleanup tags per year declines since 2006, which indicates that a stable set of cleanup tags, which cover all relevant quality flaws, is to be expected within the next few years.

Part of our current research is the analysis of how the editing behavior changes after an article has been tagged, and what users are involved in tagging and fixing quality flaws. With respect to future work, it should be investigated how vandalism effects our analysis. We have applied simple heuristics to discard vandalism edits, a more sophisticated vandalism detection approach would be desirable.

7. REFERENCES

[1] B. T. Adler, L. de Alfaro, and I. Pye. Detecting Wikipedia vandalism using WikiTrust: lab report for PAN at CLEF'10. *Notebook Papers of CLEF 2010 LABs and Workshops*, 2010.

[2] M. Anderka, B. Stein, and N. Lipka. Towards automatic quality assurance in Wikipedia. In *Proceedings of the 20th conference on World Wide Web (WWW'11)*, pages 5–6, 2011.

[3] M. Anderka, B. Stein, and N. Lipka. Predicting quality flaws in user-generated content: the case of Wikipedia. In *Proceedings of the 35th international ACM conference on research and development in information retrieval (SIGIR'12)*, 2012.

[4] M. Anderka and B. Stein. A breakdown of quality flaws in Wikipedia. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality'12)*, pages 11–18, 2012.

[5] J. Blumenstock. Size matters: word count as a measure of quality on Wikipedia. In *Proceedings of the 17th conference on World Wide Web (WWW'08)*, pages 1095–1096, 2008.

[6] D. Dalip, M. Gonçalves, M. Cristo, and P. Calado. Automatic quality assessment of content created collaboratively by Web communities: a case study of Wikipedia. In *Proceedings of the 9th ACM/IEEE-CS joint conference on digital libraries (JCDL'09)*, pages 295–304, 2009.

[7] W. Emigh, S. Herring. Collaborative authoring on the Web: a genre analysis of online encyclopedias. In *Proceedings of the 38th annual Hawaii international conference on system sciences (HICSS'05)*, 2005.

[8] L. Gaio, M. den Besten, A. Rossi, and J. Dalle. Wikibugs: using template messages in open content collections. In *Proceedings of the 5th symposium on wikis and open collaboration (WikiSym'09)*, pages 14:1–14:7, 2009.

[9] J. Giles, Internet encyclopaedias go head to head. In *Nature*, 438(7070), 2005.

[10] M. Hu, E. Lim, A. Sun, H. Lauw, and B. Vuong. Measuring article quality in Wikipedia: models and evaluation. In *Proceedings of the conference on information and knowledge management (CIKM'07)*, pages 243–252, 2007.

[11] A. Lih. Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th symposium on online journalism*, pages 16–17, 2004.

[12] M. Potthast, B. Stein, and T. Holfeld. Overview of the 1st international competition on Wikipedia vandalism detection. In *Notebook Papers of CLEF 10 Labs and Workshops*, 2010.

[13] N. Lipka and B. Stein. Identifying featured articles in Wikipedia: writing style matters. In *Proceedings of the 19th conference on World Wide Web (WWW'10)*, pages 1147–1148, 2010.

[14] L. H. Rector. Comparison of Wikipedia and other encyclopedias for accuracy breadth, and depth in historical articles. In *Reference services review*, 36(1):7–22, 2008.

[15] A. Rossi, L. Gaio, M. den Besten, and J. Dalle. Coordination and division of labor in open content communities: the role of template messages in Wikipedia. In *Proceedings of the 38th annual Hawaii conference on system sciences (HICSS'10)*, pages 1–10, 2010.

[16] K. Smets, B. Goethals, and B. Verdonk. Automatic vandalism detection in Wikipedia: towards a machine learning approach. In *Proceedings of the workshop on Wikipedia and artificial intelligence: an evolving synergy (WikiAI'08)*, pages 43–48, 2008.

[17] B. Stvilia, M. Twidale, L. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *Proceedings of the conference on information quality (ICIQ'05)*, pages 442–454, 2005.

[18] B. Suh, G. Convertino, E. Chi, and P. Pirolli. The singularity is not near: slowing growth of Wikipedia. In *Proceedings of the 5th symposium on wikis and open collaboration (WikiSym'09)*, pages 1–10. 2009.

[19] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'04)*, pages 575–582, 2004.

[20] R. Wang and D. Strong. Beyond accuracy: what data quality means to data consumers. In *Journal of management information systems* 12(4):5–33, 1996.

[21] D. Wilkinson and B. Huberman. Cooperation and quality in Wikipedia. In *Proceedings of the 3rd symposium on wikis and open collaboration (WikiSym'07)*, pages 157–164, 2007.