ABSTRACT
Ensuring the quality of the content provided in online settings is an important challenge today, for example, for social media or news. The Wikipedia community has ensured the high-quality standards for an online encyclopaedia from the beginning and has built a sophisticated set of automated, semi-automated, and manual quality assurance mechanisms over the last fifteen years. The scientific community has systematically studied these mechanisms but one mechanism has been overlooked — edit filters. Edit filters are syntactic rules that assess incoming edits, file uploads or account creations. As opposed to many other quality assurance mechanisms, edit filters are effective before a new revision is stored in the online encyclopaedia. In the exploratory study presented, we describe the role of edit filters in Wikipedia’s quality assurance system. We examine how edit filters work, describe how the community governs their creation and maintenance, and look into the tasks these filters take over. Our goal is to steer researchers’ attention to this quality control mechanism by pointing out directions for future studies.

CCS CONCEPTS
• Human-centered computing; • Information systems → Wikis;

KEYWORDS
Wikipedia, edit filters, algorithmic quality assurance, vandalism

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
OpenSym 2020, August 25–27, 2020, Virtual conference, Spain © 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-8779-8/20/08... $15.00 https://doi.org/10.1145/3412569.3412580

1 INTRODUCTION
The public heatedly debated so-called “upload filters” in the context of the EU copyright law reform in 2019. Upload filters are conceived as a form of copyright protection. They should check for possible infringements of copyrighted material even before a contribution is published online — and the contributions affected can be automatically blocked and thus effectively prevented. Central arguments put forward by the upload filters’ sympathizers were the desire to improve the quality of the content provided and defend creators’ rights. Some questions which arose are how platform providers should implement and use such filters, to what extent such filters can help assure quality and how the risk of surveillance of platform users can be minimised. Even though we cannot answer all these questions, we would like to take the opportunity to look into a concrete filter mechanism that has been conceptualised, developed, and maintained by an online community: the Wikipedia community.

Wikipedia is an open online encyclopaedia; in other words, everybody can participate. It is created by a peer production community that consists of 38 million users. Wikipedia is currently comprised of more than 50 million articles in more than 300 different languages. Research has often explained its success and productivity by the sophisticated social organisation of its human editors, but it

2However, such participation is possible only if a person has an internet connection, knowledge of wiki syntax and the willingness to understand all the policies and guidelines, among other things.
Figure 2: Quality assurance mechanisms on Wikipedia organised by their degree of automation and mean response time.

has also shown that algorithmic governance realised by software-based mechanisms is crucial [4, 7, 8, 11, 14]. Over the past 15 years, the community has built an effective quality assurance system that includes cognitive and algorithmic mechanisms [11]. How is this quality assurance system explained on a general level? When a person edits Wikipedia, bots evaluate the contribution within seconds [6]. Subsequently, humans check the edit with the help of semi-automated tools [6] and finally, Wikipedians evaluate changes manually [4]. We discuss different mechanisms and their effect on an edit. However, there is one mechanism in Wikipedia’s quality assurance system which has so far been hardly discussed — edit filters. Unlike bots or semi-automated tools that are employed after an edit has been stored in Wikipedia, edit filters operate in advance. They can disallow certain editors’ activities from the beginning but also warn editors or tag specific activities.

Focusing on the English Wikipedia, we discuss edit filters and their role among existing quality control mechanisms. We describe the technical functioning of edit filters and how they are governed. Based on the data available, we carried out a content analysis to typify edit filters’ tasks and describe how filters interact with the user. Furthermore, we examine how the use of filters has evolved over time. This exploratory study provides first insights into the technical functioning of filters, which tasks filters perform in Wikipedia’s quality assurance system and how these tasks are differentiated according to the context of use.

We have organised our article into four sections. Firstly, we highlight existing research on quality assurance mechanisms on Wikipedia. Secondly, we describe edit filters: how they work technically and how the Wikipedia community governs the usage of this software. We then introduce the data we collected in order to understand the types of edit filters and how we labelled them. Finally, we show some descriptive analysis of the data and discuss our insights in the last section. We especially suggest possible directions for future research.

2 RELATED WORK

We have structured the different mechanisms of Wikipedia’s quality assurance system by their degree of automation, which ranges from fully automated, where a decision is realised by a computer with no human intervention, to the lowest level that describes a human who carries out the whole process manually. Semi-automated mechanisms support human decisions in one way or the other, in any case, the final decision is made by the human. We provide an overview of these mechanisms in Figure 2 ordered by their mean response time regarding the arrival of new edits.

2.1 Automated Mechanisms

Automated mechanisms denote software that works in a fully automated fashion without human intervention. Previous research has discussed following automated mechanisms: Wikipedia’s SpamBlacklist, fully automated bots, such as ClueBot NG, and the machine learning framework ORES.

West et al. discuss link spam on Wikipedia and have identified the SpamBlacklist — consisting of a list of external links that are not allowed to be added — as the first mechanism to be activated in the spam removal pipeline [18]. The mechanism disallows the addition of said external links on the spot.

Geiger and Ribes investigate the collaboration of different mechanisms (bots such as ClueBot and HBV AIV helperbot) and humans using semi-automated tools, such as Huggle and Twinkle, who worked together towards the banning of a malicious user [8]. Halfaker and Riedl offer a historical review of bots and semi-automated tools and their involvement in vandal fighting [11], assembling a comprehensive list of tools and touching on their working principle (rule vs. machine learning based). They also develop a bot taxonomy, classifying bots in one of the following three groups according to their task area: content injection, monitoring or curating; augmenting MediaWiki functionality; and protection from malicious activity. In [6], Geiger and Halfaker conduct an in-depth analysis of ClueBot NG, ClueBot’s machine learning based successor, and its place within Wikipedia’s vandal fighting infrastructure.
concluding that quality control on Wikipedia is a robust process and most malicious edits are eventually reverted even with some of the actors (temporarily) inactive, although at a different speed. Geiger [5] discusses the repercussions of so-called “bespoke code” (code that runs alongside the core software), such as bots, and finds it more flexible but also less reliable than software which is part of the server-side codebase4.

The Wikipedian machine learning framework ORES is announced and described by the “ORES Paper” [12]. It uses learning models to predict a quality score for each article and edit based on edit/article quality assessments manually assigned by Wikipedians. Potentially damaging edits are highlighted, which allows editors who engage in vandal fighting to examine them in greater detail. Halfaker and Geiger [9] scrutinise more closely ORES’s potential to decouple following key tasks in algorithmic quality assurance: training data curation, building of learning models, and development of interfaces/tools that use the predictions.

2.2 Semi-Automated Mechanisms

Semi-automated mechanisms allow people to eventually make a decision, whereby this decision-making is supported by the software to a varying extent (e.g., by showing alternatives or making recommendations). In the following, we briefly describe Wikipedia’s PageProtection mechanism and the tools Huggle, Twinkle and STiki, which have all been scrutinised by the scientific community.

Hill and Shaw [13] study PageProtection and find it highly configurable: available in more than ten varieties including the most popular “full protection” (only administrators can edit) and “semi-protection” (only registered, autoconfirmed users can edit). Moreover, it is found that pages are protected for various reasons: exemplarily, to prevent edit warring or vandalism and to enforce a policy or the law; it is an established process to protect articles on the front page. The researchers also look into the historical development of protected pages on Wikipedia and discuss the repercussions of the mechanism for the users affected [13]. We have classified PageProtection as a semi-automated mechanism, since an administrator has to actively protect a page.

Geiger and Ribes [8] discuss Huggle and Twinkle: Huggle is a fully assisted quality control tool which compiles lists of potentially problematic edits and automates workflows, such as the edit revert and warning off disrupting editors. Twinkle, on the other hand, is a less automated web browser extension which adds contextual links to Wikipedia and, thus, facilitates the execution of complex workflows, such as the roll back of multiple edits, nominating software to a varying extent (e.g., by showing alternatives or making recommendations). Such bots, and finds it more flexible but also less reliable than software which is part of the server-side codebase4.

4 Note that part of these observations are not relevant anymore since a lot of bots are run on infrastructure (ToolForge) provided by Wikimedia itself.

Figure 3: The filter pattern of Edit filter 50 "Shouting".

2.3 Manual Mechanisms

When using manual mechanisms, software offers users no assistance in decision-making but might provide necessary data. Geiger and Ribes maintain that Wiki software is also suitable for manual quality control work: representing information changes via diffs allows editors to quickly spot content that deviates from its immediate context [8]. Asthana and Halfaker give us a hint regarding what type of quality control work humans take over: less obvious and less rapid, requiring more complex judgement [1].

In the following, we investigate another mechanism – edit filters. We describe the role of edit filters in Wikipedia’s quality assurance system by examining how edit filters work, how the community governs their creation and maintenance, and look into the tasks these filters take over.

3 EDIT FILTERS ON WIKIPEDIA

On 10 August 2020 at 8:17 PM User:GandalfGray decides to edit the article “Sea otter” on English Wikipedia5. They click on “edit source”, add the string 57SJ7JWHWYBJ3QAG5XCO to the top of the page and press “save". The edit, however, is not simply saved. Instead, User:GandalfGray gets a warning message (cf. Figure 1) which identifies the edit as probably non-constructive and encourages them to revise it before saving. The message also points out possible resources for getting help and provides the user with the opportunity to report a potential false positive. The editor User:GandalfGray has just triggered the edit filter 506.

3.1 From a Technical Perspective

Wikipedia’s edit filters are a fully automated quality control mechanism implemented by a MediaWiki extension7 that allows every edit (and some other editor’s actions) to be checked against a specified pattern before it is published on Wikipedia. Filter patterns bear some similarity to regular expressions and also allow for checking the value of some system variables, such as the character count

4https://en.wikipedia.org/wiki/Sea_ottter
5https://en.wikipedia.org/wiki/Special:AbuseFilter/50
6The MediaWiki extension is called AbuseFilter. After some discussion, the community has come to use the term EditFilter for all its user-facing parts in order to avoid false positives being characterised as “abuse” and, thus, good faith editors who are striving to improve the encyclopedia feeling alienated.
of the contribution, the user group(s) of the editor and the page namespace. If there is a match, the edit in question is logged and, potentially, additional actions, such as tagging the edit summary, issuing a warning or disallowing the edit, are invoked. We show an example of such a pattern (filter 50) in Figure 3. It is triggered if a non-confirmed user (line 1) edits a page in namespace 0, i.e., an article (line 2), and adds a sequence (line 6) of more than 12 characters (line 3) which contains at least five consequent caps (line 4) without having removed a sequence of at least five consequent caps (line 7), and exempting templates containing caps (lines 8—9).

As of 10 June 2020, the complete list of editor’s actions which can trip a filter includes edit, move, delete, createaccount, autorecreateaccount, upload, and stashupload.

Based on the actions of an editor and the pattern defined, various filter actions, such as tag, throttle, warn and disallow, can be triggered. These actions are detailed in Table 1. Even though available, according to the logs, the actions rangeblock, block and degroup have never been used on English Wikipedia. Those severer actions were discussed controversially by the community before introducing the extension, and a lot of Wikipedians felt uncomfortable with a fully automated mechanism blocking users indefinitely or removing them from privileged groups. As far as we can tell, the functionality has been implemented but never activated (at least on the EN Wikipedia). In addition to the filter actions log, tag, warn or disallow, the actions blockautopromote and aftv5flagabuse were triggered for the last time on English Wikipedia in 2012.

Guidelines specifically call for a careful use of disallow. Only severe cases for which “substantially all good-faith editors would agree are undesirable” or specific cases for which consensus has been reached by the community should be disallowed.

### 3.2 From a Social Perspective

The Edit Filter extension brings its own extensive set of permissions which regulate read and write access to the individual filters. A special role exists on English Wikipedia, that of the edit filter manager, who has write access to all edit filters. The role is only granted to editors who have already earned the trust of the community, i.e., they are already admins and have shown that they have some technical understanding of filters. As of 10 May 2019, there are 154 edit filter managers on English Wikipedia. For comparison, as of 9 March 2019 there are 1,181 admins in English Wikipedia. The role does not exist, for example in the German, Spanish, and Russian Wikipedia. Administrators on these language versions have the abusefilter modify permission automatically.

The edit filters managers group is comparatively small and it is difficult to obtain the corresponding permissions. By comparison, there are at least 232 bot operators and 6,130 users who have the rollback permission. The edit filter managers group is not only small, it also seems to be aging: some of the 154 edit filter managers on English Wikipedia have a “not active at the moment” banner on their user page.

There are best practice guidelines which edit filter managers are encouraged to consult before introducing a new filter. These guidelines suggest a process of testing and a step-wise application of severer filter actions. Edit filter managers implement filters based on phenomena observed caught by other filters or quality control guidelines suggest a process of testing and a step-wise application of severer filter actions. Edit filter managers implement filters based on phenomena observed caught by other filters or quality control mechanisms, general experience and requests by other users. At the beginning, the filters are enabled without triggering any further actions until enough log entries are generated to evaluate whether the phenomenon is significant and frequent enough to need a filter. Filters are updated over time and it is not uncommon that the action(s) a particular filter triggers changes. When a new persistent

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>tag</td>
<td>The contribution is tagged with a specific tag which appears on log pages and allows aggregations for dashboards and similar.</td>
</tr>
<tr>
<td>throttle</td>
<td>The filter is activated upon the tripping of a rate limit. Configurable parameters are the number of actions allowed, the period of time in which these actions must occur and how those actions are grouped. (A simple example of throttling is something like “do this if page X is edited more than Y times in Z seconds”.)</td>
</tr>
<tr>
<td>warn</td>
<td>A warning is displayed that the edit may not be appreciated. The editor who tripped the filter is provided with the opportunity to revise their edit and resubmit it.</td>
</tr>
<tr>
<td>disallow</td>
<td>An error message is displayed that the edit was considered non-constructive and will not be saved. The editor is provided with the opportunity to report a false positive.</td>
</tr>
<tr>
<td>blockautopromote</td>
<td>The user whose action matched the filter’s pattern is banned from receiving extra groups from $wgAutopromote for a random period of 3 to 7 days.</td>
</tr>
</tbody>
</table>

*The stashupload action refers to uploading files to MediaWiki’s temporal storage space.*

*https://www.mediawiki.org/w/index.php?title=Extension:ArticleFeedbackv5&oldid=884572675*  
vandalism trend emerges, a filter might temporarily be set to “warn” or “disallow” and the actions are removed again as soon as the filter is no longer tripped very frequently. The Edit Filter Noticeboard\(^16\) is used to discuss such action changes, and updates to an edit filter’s pattern, a warning template, or problems with filters behaviour.

Whereas only edit filter managers have the permissions necessary to implement filters, anybody can propose new ones on the Edit Filter Requested page\(^17\). However, the latter describes criteria for meaningful filter characteristics: problems with a single page are not suitable for an edit filter, since filters are applied to all edits; no trivial errors should be caught by filters (e.g., concerning style guidelines). Furthermore, filters make editing slower, so the usefulness of every single filter and condition has to be carefully considered. More complex in-depth checks of a page should be done by a separate software, such as bots.

4 DATA AND METHOD

As noted in the introduction, the present research is exploratory — we “look [...] at data to see what it seems to say” [16, p. v]. We draw our insights from (1) existing research on Wikipedia’s quality control mechanisms, (2) documentation and discussion pages of Wikipedia or MediaWiki, and (3) database tables of the MediaWiki extension containing the patterns, actions, etc. of edit filters.

4.1 Data

We downloaded the abuse_filter table from the enwiki_p database on 6 January 2019 via Quarry\(^18\); it contained 954 rows, i.e., all filters implemented on English Wikipedia to this date. We scrutinised the following information from the table: (1) the public description (or name) of a filter (column af_public_comments), (2) comments from the edit filter managers on changes undertaken (column af_comments), (3) the pattern constituting the filter (column af_pattern) and (4) the filter action designated (column af_actions). Out of the total 954 filters, 593 were hidden from public view, i.e., the columns af_pattern and af_comments were empty for them. Thus, hidden filters were analysed based only on their names and actions invoked.

4.2 Coding

We studied this data described previously in detail and labelled it via an emergent coding scheme following a content analysis [15] in order to gain a detailed understanding of how edit filters are used in English Wikipedia. These codes emerged from the data: some of them being literal quotes of terms used in the description or comments on a filter, while others summarised the filter functionality perceived. In addition to that, some of the vandalism types elaborated by the Wikipedia community were used for vandalism related labels.

We conducted two rounds of labelling which are described in more detail next.

4.2.1 First Labelling. During the first round of coding, potential labels were created by reading the description, comments, and checking the pattern of the edit filters used. After assigning labels to all edit filters, labels that seemed redundant were merged. At the same time, codes were sorted and unified into broader categories which seemed to relate the single labels to each other. The following four general clusters of codes were identified: vandalism, good_faith, maintenance and unknown. Based on these clusters, a code book that describes them and their corresponding codes was created.\(^19\)

We encountered a number of challenges during the first round of labelling. There were cases for which no clear code could be assigned since the filter patterns and comments were hidden from public view.\(^20\) Moreover, there were also cases, not necessarily hidden, where no suitable label could be determined, since the filter pattern was ambiguous, none of the existing categories seemed to fit and/or no insightful new category emerged. These form the unclear category. It was particularly difficult to determine for a number of filters whether they were targeting vandalism or good


\(^{19}\)Quarry is a web-based service offered by the Wikimedia Foundation for running SQL queries against their public databases: https://quarry.wmflabs.org/
faith edits (labelled with vandalism? or “good_faith?”), respectively, or sometimes with both). The only thing that would have distinguished between the two would have been the contributing editor’s motivation, which we do not consider in this phase of the project.

4.2.2 Second Labelling. The whole dataset was labelled again in a second round based on the code book to address these challenges and verify the labelling. Exactly one label was assigned to every filter—the one deemed most appropriate (although alternative possibilities were often listed as notes)—without looking at the labels from the first round.

Furthermore, it was assumed that all hidden filters target a form of (graver or less grave) vandalism. The edit filter guidelines suggest that filters should not be hidden in the first place unless dealing with cases of persistent and specific vandalism where it could be expected that the vandalising editors will actively look for the filter pattern in their attempts to circumvent the filter. Hence, all hidden filters for which there were not any more specific clues (e.g., in the name of the filter) were tagged as hidden_vandalism.

The “assume good faith” guideline was followed for the ambiguous cases which received the vandalism? or good_faith? labels in the first round. Thus, we labelled as vandalism only cases where good faith was definitely out of the question. One decisive feature here was the filter action which represents the judgement of the edit filter manager(s). Since communication is crucial when assuming good faith, all ambiguous cases which have a less “grave” filter action, such as tag or warn (which seeks to give feedback and, thereby, influence a constructive contribution) have received a good_faith label. On the other hand, filters set to disallow were tagged as vandalism or a particular type thereof, since the filter action is a clear sign that at least the edit filter managers have decided that seeking a dialogue with the offending editor is no longer an option.

5 RESULTS

Based on the content analysis of the filters, we can provide an overview of the types of tasks filters take over and their usage over time.

5.1 Types of Tasks

As of January 2019, there are 954 filters on the English Wikipedia. Since deletion of a filter is marked merely by setting a flag in the database (but not removing any data), this number constitutes all filters created on the English Wikipedia up to this date. Of these, 21% (or 201) are active, 16% (or 153) are disabled, and 63% (or 600) are deleted (cf. Figure 4). In total, the majority of filters (593, 62%) are hidden.

5.1.1 Public and Hidden Filters. Historically, it was planned to hide all edit filters from the general public. The community discussions rebutted that, therefore, a guideline was drafted calling for hiding filters “only where necessary, such as in long-term abuse cases where the targeted user(s) could review a public filter and use that knowledge to circumvent it.” This guideline is, however, not always complied with, and edit filter managers end up hiding filters that target general vandalism despite the consensus that these should be public; however, these cases are usually made public eventually. Examples hereof are filters 225 “Vandalism in all caps”, 260 “Common vandal phrases” or 12 “Replacing a page with obscenities”. Often when a hidden filter is marked as “deleted”, it is made public.

Furthermore, caution in filter naming is suggested for hidden filters and editors are encouraged to give such filters only a simple description of the overall disruptive behaviour rather than naming a specific user that is causing these disruptions. Although almost 2/3 of all filters are hidden (cf. Figure 4), if we look only at the
The vast majority of edit filters on English Wikipedia could be said to target (different forms of) vandalism, i.e., maliciously intended disruptive editing (or other activity, such as account creation). Some examples thereof are filters for "silly vandalism" (i.e., inserting swear or obscene words or nonsense sequences of characters into articles), hoaking (i.e., inserting obvious or less obvious false information in articles), template vandalism (i.e., modifying a template in a disruptive way which is quite severe since templates are displayed on various pages) or spam (i.e., inserting links to promotional content, often not related to the content being edited). All codes belonging to the vandalism category together with a definition and examples can be consulted in the code book.

Some vandalism types seem to be severer than others (e.g., sock puppetry, or persistent long-term vandalism). It is mostly in these cases that the implemented filters are hidden. Labels referring to such types of vandalism form their own subcategory: **hardcore vandalism**.

**Good Faith.** The second largest category identified are filters targeting edits which are (mostly) disruptive but not necessarily made with bad intentions. The adopted name *good faith* is a term utilised by the Wikipedia community itself. Filters from this category are frequently aimed at non-constructive edits done by new editors not familiar with syntax, norms or guidelines which results in broken syntax, deleting something without running it through an "Articles for Deletion" process, copyright violations, or edits without or with improper sources.

The focus of these filters lies in the communication with the disrupting editors. A lot of the filters issue warnings intended to guide the editors towards ways of modifying their contribution to become a constructive one.

**Maintenance.** Some of the edit filters encountered on the English Wikipedia were targeting neither vandalism nor good faith edits. They had their focus more on (semi-)automated routine (clean-up) tasks. These filters form the *maintenance* category. Some of them target, for example, bugs, such as broken syntax caused by a faulty browser extension. Or there are others which simply track particular behaviours (such as mobile edits or edits made by unflagged bots) for various purposes.

The *maintenance* category differs conceptually from the *vandalism* and *good faith* ones in so far that the logic behind it is not the editors’ intention but rather "side" occurrences that have mostly gone wrong.

**Unknown.** This is an auxiliary category used to code all filters where the functionality stayed completely opaque to the observer or no better fitting label emerged although it was comprehensible what the filter was doing.

### 5.2 Development Over Time

We explored the temporal changes in the number of filters and the types of tasks filters are responsible for to get a first overview on the development of filters on Wikipedia over time.

The overall number of active edit filters has generally been stable over time. The upper limit is defined by the so-called "condition limit" that is agreed upon by the community which allows for only a certain number of conditions to be active at the same time. The reasoning behind it is that every incoming edit (or another editor’s action) is checked against all active filters and, hence, the higher the number of filters (and conditions which constitute a filter), the longer it takes until the checks are performed.

The most active filters of all times stay quite stable through the years. However, there seem to have been more active *good faith* filters in the first years after the introduction of the mechanism

---

25Sock puppetry denotes the creation and employment of several accounts for various purposes, such as pushing a point of view or circumventing bans.
(cf. Figure 7), while in later years, most filter hits are produced by vandalism-related filters. There is a certain periodicity to the filter hits, with (potentially) disrupting editors being less active in the summer months of the northern hemisphere (June to August). Furthermore, there is a sudden surge in the number of filter hits at the beginning of 2016 and a subsequently higher baseline in hit numbers. Whereas we could find no conclusive explanation for the higher baseline, the peak in 2016 seems to, at least, partially have been caused by a rise in attempted account creations: for example, out of the total 372,907 filter hits in January 2016, 71,902 are caused by the createaccount editor’s action. The majority of these were produced by the hidden filter 527 ”T34234: log/throttle possible sleeper account creations”, which is a throttle filter with no further actions enabled, so it is effectively just logging every account creation matching its criteria.

6 DISCUSSION

In the following, we discuss the findings of our descriptive investigation from different perspectives. We compare edit filters with other mechanisms and investigate the integrated use of existing mechanisms. We, furthermore, discuss possible future research directions regarding cultural differences in the different Wikipedia language versions, impact on new editors and the filters’ potential to serve as an early warning system on Wikipedia.

Comparing Edit Filters to Other Mechanisms. Together with the SpamBlackList and the TitleBlackList, edit filters disallow an edit before its publication. This commonality also applies to the page protection mechanism. However, page protection is, unsurprisingly, page-centered. By contrast, edit filters allow for a finer level of control: malicious activities of a single user or a particular type can be targeted without restricting everyone else from publishing.

Research has shown that bots, tools and humans are quite effective when fighting vandalism [6, 8]. In this sense, not only we but also the community asked themselves whether filters were needed when bots already existed. The community raised two distinct arguments for introducing edit filters.

Firstly, bots were not reverting some (obvious but pervasive) vandalism, such as mass moves of some pages to nonsensical titles fast enough, or cleaning required human intervention. One can argue that this is probably obsolete today since there are sophisticated bots such as ClueBot_NG or bots building upon the ORES infrastructure. However, operating these bots requires not only knowledge of a programming language and the Wikipedia’s API, but also the compilation of training data, and the time needed for adapting the needed machine learning model might be too long. Therefore, semi-automated tools (such as Huggle) and filters seem to be more reasonable for specific well-targeted attacks against Wikipedia because they require less technical effort to be realised: The user of semi-automated tools “only” has to learn the user interface of a software for filtering edits [3], and an edit filter manager “only” has to understand regular expressions.

Secondly, there was some discontent with bot governance (e.g., unresponsive bot operators and poorly written and tested source code). People have hoped to handle these concerns in a better manner by using edit filters. As opposed to many admin bots, the source code of the EditFilter extension is publicly available and well tested. Even though this argument still holds, the most edit filters are not public. Hidden filters seem to have higher fluctuation rates, i.e., they probably target specific phenomena for a particular period, after which, the filters are disabled and, eventually, deleted. This distribution makes sense regarding the filter policy: hidden filters are for particular cases and very determined vandals; public filters for general patterns that reflect more timeless issues.

Edit Filters Engaging with Other Mechanisms. Geiger and Ribes [8] show that quality assurance mechanisms on Wikipedia operate not only alongside each other but also collaborate in an elaborated manner within a “distributed cognition” network. The researchers trace a disrupting editor and the actions and mechanisms employed to discourage them throughout Wikipedia: the bot ClueBot_NG and several editors using the semi-automated tools Huggle and Twinkle worked together until an administrator banned the malicious user. In our exploration, we have also come across edit filters and bots mutually facilitating each other’s work. DatBot, Mr.Z-bot
and MusikBot are examples of bots conducting support tasks for filters. DatBot[26], for example, monitors filter logs and reports users tripping certain filters to administrators. There are also examples of filters supporting bot work: Filter 323 “Undoing anti-vandalism bot”, for example, tags edits reverting XLinkBot’s and ClueBot NG’s revisions. Moreover, several filters were (historically) configured to ignore particular bots: Filter 76 “Adding email address” exempting XLinkBot, is an example thereof. Some filters ignore all bots (e.g., filter 368 “Making large changes when marking the edit as minor”). Sometimes, data from the AbuseLog is also used for (semi-)protecting frequently disrupted pages.

Cultural Differences of Edit Filter Usage Between Different Wikipedia Language Versions. As already emphasized, the focus of the present work is on English Wikipedia; it would be insightful to look into the use of filters and their activity in other language versions. The difference in the governance in the different language versions might indicate existing differences in the filters employed. German Wikipedia, for example, operates 81 enabled edit filters, of which 38 are public (47%) and 43 (53%) are hidden[27]. At first glance, the distribution seems very similar, but the question is whether differences in the type and actions exist. Furthermore, no edit filter manager group exists on other major language versions, but administrators do have additional permissions. We expect that this also shapes the usage of the filtering mechanism. Additionally, not all language versions of Wikipedia use the same extensive set of automated tools. It is conceivable that some smaller Wikipedias rely more heavily on edit filters than on sophisticated machine-learning based bots.

Impact of Edit Filters on New Editors. Our work signals that edit filters focus on two significant aspects in assuring quality in Wikipedia: vandals and good-faith editors. The latter especially need support while editing Wikipedia since research has shown that the number of rules and guidelines has increased over the last few years and it has become increasingly difficult for inexperienced editors to comply with them [2]. However, we have not considered this aspect in our current study. Halfaker and colleagues [10], for example, showed a negative impact on the retention rate of new users because of the increased usage of bots and semi-automated tools (e.g., Huggle). Further analyses could help one to understand the experiences of editors who have tripped filters and whether filters are useful or counterproductive for integrating inexperienced good-faith editors. Filters have the ability to give some feedback before an edit is published, they provide editors with the possibility of improving their contribution before submitting. Further research on the repercussions of edit filters on newcomers could determine whether this mechanism is helpful to retain new editors or whether it rather discourages them from participation in one context or the other.

Edit Filters as an Early Warning System. In our study, we have taken a first step towards understanding how edit filters are used by the Wikipedia community. However, we have not yet examined the actions in more detail to see how certain situations outside (e.g., political events or Internet trends) or inside Wikipedia (e.g., certain content) trigger different filters. A brief spot-check of the pages edited most frequently during the filter hits peak at the beginning of 2016 did not uncover any interesting underlying pattern. After UserLogin (where all 71,920 attempted account creations were logged), the page with most filter hits in January 2016 was Skateboard, where a filter was triggered 660 times. More research is needed to better understand how filters react to certain attempts to change, how long it takes to implement filters for specific attacks and whether the AbuseLog entries can be used as an early warning system that a bigger vandalism wave is underway. In our study, we have not scrutinized the patterns of the hidden filters; even though we reflected on some of their characteristics (i.e., their public descriptions and appointed actions), a more detailed inspection is needed. Furthermore, an ethnographic analysis, which may answer several valuable questions, such as for which cases does the community implement a bot and for which a filter, is missing.

Edit Filters as a Contemporary Mechanism. On the one hand, it seems astonishing that edit filters, although based on a simple rule-based approaches, still enjoy vast popularity and have not yet been abolished by the Wikipedia community. One reason might be that rule-based systems are more transparent than more sophisticated machine learning-based approach, such as those used by ClueBot NG or ORES. Even users without a formal CS education can understand regular expressions with a little training. Edit filters are simpler to work with: it is easier to add yet another rule than tweak parameters in a machine learning-based software. A filter can, for example, disallow edits by specific users directly. Therefore, both methods — the machine learning-based and the rule-based approach — can be seen as complementary quality assurance mechanisms. Further research is needed to understand better when to use which mechanism and how they can benefit from each other.

On the other hand, edit filters, regarding their effect on certain content moderation aspects in the Wikipedia community – what is and what is not allowed – take on a new significance in the light of the upload filters discussed at the beginning of this article. Various concerns exist regarding the usage of edit filters in the community. These concerns might be one primary reason why the community has never agreed on using more rigid filter actions, such as range and range_block. In addition, some community members were apprehensive about filter governance, particularly how restrictive the access is to the edit filter management group. However, even though these concerns exist, edit filters provide an interesting object of study for developing filter systems in other contexts.

Even though we cannot discuss the individual aspects of either side in detail here, we want to provide impulses to think about the direction in which further research can be done regarding these two aspects.

7 CONCLUSION

This work offers an initial descriptive investigation of edit filters on English Wikipedia and their role among other quality control mechanisms. We traced why the community introduced filters, summarised how filters work and what they seem to be doing currently on the English Wikipedia. Edit filters, together with page protection and the title/spam blacklists, are the first mechanism to detect potentially malicious activity on Wikipedia. All these mechanisms
can be configured to disallow certain types of edits outright or, in the case of edit filters, also other editors’ actions. What is more, they all spring to action before an edit is even published, which makes them highly invisible to both regular Wikipedia users and researchers who mostly rely on revision history for conducting their studies. Edit filters were born partially out of discontent with bot governance and the intention to disallow blatant which was pervasive and cumbersome to remove. This also explains why most filters are hidden: in order to make it difficult for specific motivated vandals targeted by the mechanism to circumvent the measure. In addition to taking care of vandalism, we found edit filters to issue warnings with improvement suggestions for “common newbie mistakes”, such as adding a large amount of text without references or not formatted according to WikiSyntax. Whether this is an effective measure for guiding new editors towards a productive contribution instead of alienating them by reversing the changes they made is something future research should look into. With the present study, we aimed to raise the awareness of a powerful, previously disregarded, quality control mechanism on Wikipedia. Other intriguing questions that remain unanswered include: What proportion of quality control work do filters take over? What types of editors trip filters (IP editors vs. registered accounts, how old are the registered accounts in question)? Are there certain kinds of pages that are more likely to trigger (certain) edit filters? We hope that others will continue the inquiry into this previously unstudied mechanism and find the artefacts we have assembled so far useful to build upon.

ACKNOWLEDGMENTS
Without Stuart Geiger this research would have never happened. His suggestion and the fruitful discussions he offered on the topic of algorithmic quality control mechanisms on Wikipedia helped to bring this research to light. We are also grateful to our anonymous reviewers for the valuable comments which very much helped to improve the paper.

REFERENCES