FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Assignment of bachelor's thesis

Title:	Wikipedia: Defense Against Vandalism
Student:	Martin Urbanec
Supervisor:	Ing. Josef Kokeš, Ph.D.
Study program:	Informatics
Branch / specialization:	Computer Security and Information technology
Department:	Department of Computer Systems
Validity:	until the end of summer semester 2023/2024

Instructions

1) Describe Wikipedia (https://www.wikipedia.org), its philosophy and history, and explain how these factors make it vulnerable to vandalism. List major historical incidents and attempt to estimate the volume of these issues.
2) Research the current technological and organizational defenses Wikipedia utilizes to combat vandalism. Evaluate their effectiveness.
3) Analyze possible new approaches to these issues: For each, propose a technical (or nontechnical) solution, evaluate its suitability with regards to Wikipedia philosophy, gather available historical data and verify whether the proposed solution is viable (F1 score, recall, FPR). Estimate the likelihood and timeframe for adding it to Wikipedia.
4) Discuss your results and possible extensions to them.

Electronically approved by prof. Ing. Pavel Tvrdík, CSc. on 12 January 2023 in Prague.

Bachelor's thesis

WIKIPEDIA: DEFENSE AGAINST VANDALISM

Martin Urbanec

Faculty of Information Technology
Department of Computer Systems
Supervisor: Ing. Josef Kokeš, Ph.D.
May 10, 2023

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis: Urbanec Martin. *Wikipedia: Defense Against Vandalism.* Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2023.

Contents

List of Figures

List of Tables

List of code listings

I would like to express my gratitude and appreciation to my supervisor Ing. Josef Kokeš, Ph.D. whose guidance during my work on the thesis was invaluable and crucial to the completion of the study. I would like to thank the members of the Faculty of Information Technology at the Czech Technical University for their constant support during my bachelor studies. Last but not least, I would like to thank my family, friends and fellow Wikipedia functionaries for their support during my work on this project.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended. In accordance with Section 2373(2) of Act No. 89/2012 Coll., the Civil Code, as amended, I hereby grant a non-exclusive authorization (licence) to utilize this thesis, including all computer programs that are part of it or attached to it and all documentation thereof (hereinafter collectively referred to as the "Work"), to any and all persons who wish to use the Work. Such persons are entitled to use the Work in any manner that does not diminish the value of the Work and for any purpose (including use for profit). This authorisation is unlimited in time, territory and quantity.

In Prague on May 10, 2023 .....

Abstract

The thesis focuses on improving countervandalism workflows on Wikipedia. As of writing, countervandalism relies a lot on manual efforts by regular users, administrators and other stakeholders. Relying solely on human labor is not sustainable in the long term. To fix this, the thesis researches new countervandalism techniques Wikipedia could implement. To ensure the researched techniques can be deployed to all 300+ language editions of Wikipedia, all researched techniques need to depend solely on revision metadata.

Four different countervandalism techniques were suggested: (a) requiring wikipedians to identify themselves, (b) disallowing logged-out contributors, (c) disallowing users with their IP address in an external blacklist from contributing, (d) prohibiting edits by users with several prior edits reverted. All four techniques were submitted to the first round of review where their compatibility with Wikipedia's philosophy was verified. All except the first one passed the first round of review.

The second round of review was focused on viability of the three remaining solutions based on a randomly generated sample of 100 edits, manually classified by Wikipedia's administrators. While disallowing logged-out contributors indeed managed to prevent most of the vandalism, it also prevented a lot of otherwise constructive edits. On the other hand, disallowing edits by users with a lot of reverted edits proved to have potential (with thresholds to-be-clarified in a subsequent research). IP blacklists showed no results – in the generated sample of 100 edits, there were no IP addresses present on the blacklist.

Keywords Wikipedia, vandalism, anti-abuse, site security, data analysis

Abstrakt

Bakalářská práce se zabývá zlepšením boje proti vandalismu na internetové encyklopedii Wikipedii. V současné době je boj proti vandalismu prováděn ručně běžnými uživateli, správci a dalšími funkcionáři. Spoléhání se čistě na lidskou práci není dlouhodobě udržitelné. Za účelem nápravy tohoto stavu tato práce hledá nové techniky boje proti vandalismu, které by Wikipedie mohla implementovat. Práce se zabývá pouze technikami využívající čistě metadata jednotlivých revizí, čímž je zajištěna nasaditelnost technik na všech 300+ jazykových verzí Wikipedie.

Práce se zabývá čtyřmi technikami boje proti vandalismu: (a) vyžadování identifikačních údajů od wikipedistů, (b) zákaz přispívání odhlášeným (c) zamezení přispívání osobám na externím IP blacklistu (d) znemožnění přispívání uživatelům s příliš mnoho revertovanými editacemi. Všechny čtyři techniky byly prověřeny ve dvou kolech: v prvním se práce zabývá jejich souladem s myšlenkami a ideologií Wikipedie. Všcehny navržené řešení s výjimkou (a) touto první kontrolou prošly.

Druhé kolo kontroly se zabývalo ověřením životaschopnosti technik. Toto ověření proběhlo na základě náhodně vygenerovaného vzorku 100 editací, které byly ručně klasifikovány administrátory Wikipedie. Zákaz odhlášených příspěvků sice skutečně vedl k zamezení většiny vandalismu; na druhou stranu zároveň zakázal spoustu konstruktivních editací. Největší potenciál mělo zamezení přispívání uživatelům s větším množstvím revertovaných editací (konkrétní parametry musí být určeny v dalším výzkumu). U IP blacklistů nebylo dosaženo žádných výsledků – v náhodně vygenerovaném vzorku 100 editací nebyla žádná IP adresa uvedená na blacklistu.

Klíčová slova Wikipedie, vandalismus, boj proti zneužití, bezpečnost webových stránek, datová analýza

List of abbreviations

CU	CheckUser
DMCA	Digital Millennium Copyright Act
FN	false negatives
FP	false positives
IP	Internet Protocol
NDA	non-disclosure agreement
TN	true negatives
TP	true positives
TSV	tab-separated values
VPN	virtual private network
WMF	Wikimedia Foundation, Inc.

Glossary

Community Users participating in a certain Wikimedia project (or as the context may direct, participating in *any* Wikimedia project).

MediaWiki A software package that can run a wiki, used by all Wikimedia projects. MediaWiki is mostly maintained by developers associated with the WMF, Wikipedia or both.

Vandalism A non-constructive edit that disrupts Wikipedia's content. Wikipedia traditionally only labels wilful non-constructive edits as vandalism; for the purpose of this thesis, the word is used more extensively, to mean all non-constructive edits regardless of the user's intent.

Revert "Undoing or otherwise negating the effects of one or more edits, which results in the page (or a part of it) being restored to a previous version."[1] An edit can also be reverted partially; during a partial revert, only a portion of the edit is reversed and other parts of it are left intact.[1]

Wikimedia projects Projects (such as Wikipedia) operated by the Wikimedia Foundation (WMF). Each Wikimedia project has its own core principles, but all of them follow the free content model, with the main goal being the free dissemination of knowledge.[2]

Wikipedian A user contributing to Wikipedia in any way or form.

Wikimedian A user contributing to any Wikimedia project.

Introduction

Wikipedia is one of the top 10 websites in the world, serving its users over 500 millions of pages every day. Despite its popularity and size, Wikipedia heavily relies on human labor in various areas, including areas with significant potential for automatization. The extensive usage of human workforce poses a significant issue as the amount of available person-hours is limited, especially considering Wikipedia is a volunteer-run project. One of the areas heavily relying on human labor is Wikipedia's defences against vandalism. That's the focus of this thesis.

Not only does countervandalism heavily rely on human labor, it is also a critical maintenance-type duty wikipedians need to work at. If countervandalism was no longer performed, the quality of Wikipedia's content would rapidly degrade over time, eventually reaching a point when it would no longer be useful to the readers. In other words, countervandalism is a never-ending type of work which still needs to be performed even if no new content is added to Wikipedia.

The fact that countervandalism is such a critical area, combined with the amount of manual work it currently entails, is what motivated me to select the defense against vandalism on Wikipedia as the topic of this thesis. The human workforce currently used for countervandalism purposes might be put to better use elsewhere, such as article creation, content curation or mentorship of new contributors, where human creativity would be better used.

The thesis aims to research the existing countervandalism techniques, as employed by Wikipedia in the present day. Each of the currently used defenses are briefly described, including identification of its strong and weak points. Based on the set of currently employed defense tactics, possible future defense techniques are suggested, which could be used to better detect vandalism automatically.

Each of the newly identified countervandalism techniques are reviewed in two rounds. In the first round of reviews, the suitability of each of the techniques are reviewed, as well as its compatibility with Wikipedia's philosophy. This will help eliminate solutions that are either impractical or would be rejected by Wikipedia's leadership, because implementing such solutions would be in a violation of one or more of the core Wikipedia's founding principles. Results of the first round of reviews will establish a set of techniques worth for a more thorough review.

The subject of the second round of reviews is the viability of the techniques which passed the first round. Viability evaluation will establish whether the techniques actually identify a sufficient proportion of actual vandalism without also blocking constructive edits. At this stage, the evaluation is based on a random sample of edits that were saved on Wikipedia in the past.

This sample is reviewed by the author and other expert wikimedians, using the same judgement as applied while making decisions about Wikipedia's actual incoming edits. Each of the sampled

edits is either classified as vandalism or constructive. Using those classifications, it is possible to calculate the amount of vandalism that was successfully prevented, as well as the amount of false positives. Based on the viability of each technique, it is possible to recommend which techniques would be most useful to implement.

To achieve the goals described above, the thesis first provides the readers with the necessary background. The history of Wikipedia is covered in order to explain why certain founding principles of Wikipedia exists, as well as why certain obvious antivandalism techniques are not appropriate in Wikipedia's case. Furthermore, principles related to wikipedians' work are covered to further clarify the ideology Wikipedia is built on. Last but not least, to tighten the scope of this thesis, vandalism as interpreted on Wikipedia is defined, including various types of vandalism that need to be covered.

In the following chapters, the thesis explains how antivandalism on Wikipedia currently works, both socially and technically. First, Wikipedia's administrative structure is covered in order to describe the types of users who routinely participate in Wikipedia's countervandalism system. Second, technical antivandalism solutions Wikipedia already uses are enumerated, such as blocks or automated edit filters that prevent common types of vandalism. At the end of the thesis, the evaluation of newly identified countervandalism techniques is performed.

Chapter 1

Background

This chapter provides the necessary context regarding antivandalism on Wikipedia. Information from this chapter explains why certain obvious antivandalism techniques aren't viable in Wikipedia's case – they go against the principles which made Wikipedia what it is now.

1.1 Historical context: from Nupedia to Wikipedia

Wikipedia was founded on January 15, 2001, by Jimmy Wales and Larry Sanger, and it immediately succeeded. According to Similarweb, Wikipedia is the fifth most visited site in the world (and the most-visited site managed by a non-profit organisation).

However, the path to Wikipedia's inception wasn't straightforward. Wikipedia was not Wales's first attempt to start an online encyclopedia. In 2000, Wales founded Nupedia and hired Larry Sanger to manage it as its editor-in-chief. Nupedia operated on much more traditional principles than Wikipedia – articles had to be written by subject matter experts. Before each article was published, it had to undergo an extensive peer review process. [3, pp. 46-47][4]

During the years of Nupedia's existence, the high barrier of contributor entry proved to be a significant issue for the project's growth. Nupedia grew too slowly – before it ceased operation in 2003, it produced a total of 24 approved articles (with another 74 within the review process). [5, pp. 10-11] Ben Kovitz (Sanger's friend from Internet philosophy mailing lists) met with Sanger in early January 2001, and introduced the idea of *wiki software* to him. [4]

Content on wiki sites is collaboratively written by the site's own readers. The first wiki software, *WikiWikiWeb*, was created in 1995 by Ward Cunningham. The word *wiki* comes from the Hawaiian language and it means *quick* (Hawaiian term *wiki wiki* translates to *superfast* in English[4]). [3, pp. 41-42]

Sanger recognized the wiki mechanism as a potential solution to Nupedia's growth problems. Soon, he proposed adding a wiki to Nupedia to Wales. [4] On January 10, 2001, Sanger announced the creation of Nupedia's wiki via the project's e-mail conference. [6] The idea behind Nupedia's wiki was to take work off Nupedia's expert editors, enabling them to focus on reviewing content written by laypersons, which could be later incorporated into Nupedia as the actual encyclopedia.

To differentiate the new project from Nupedia, it was given the name *Wikipedia* (a compound word consisting of *wiki*, Hawaiian for *quick*, and *encyclopedia*, to denote the site's purpose).

Wikipedia was immediately successful. On February 12, 2001 (less than a month since the site launch), it had its first thousand of articles. Several months laters, ten thousand articles were available for Wikipedia's readers. [3, p. 47]. On the other hand, Nupedia either did not manage to reach out to the academic community, or the academics weren't interested in its mission. Wikipedia was successful enough that when Nupedia's server hosting crashed in September 2003, it was never restored. [4]

Nupedia and Wikipedia during its first weeks of service were both operated by Bomis, a dot-com company created by Wales. Bomis operated sites with content geared to a male audience and became successful after focusing on X-rated media. Up until late 2002, Wikipedia was a for-profit subsidiary of Bomis and hosted on `wikipedia.com`. [7, pp. 56-58] The costs of running Wikipedia were increasing with the project's popularity, while Bomis's revenues were declining due to the dot-com crash. [8] To cover the costs, Wales and Sanger decided to switch to a new funding model for Wikipedia – a charity. In 2003, all intellectual properties related to Wikipedia were transferred to a newly-founded non-profit organization – the Wikimedia Foundation (WMF). The WMF takes care of Wikipedia's hardware needs, as well as necessary institutional-level management, to the present days. Together with the switch of control to the WMF, Wikipedia's domain was changed to `wikipedia.org`, where it operates presently. [3, pp. 47-48][7, pp. 56-58]

1.2 Wikipedia's principles

Wikipedia's operation follows several key principles which affect the decisions made, including those related to countervandalism and anti-abuse. Any new solution has to be compatible with the Wikipedia's principles, otherwise it cannot be deployed[1].

The principles can be split into several groups:

The five pillars Core editorial principles (further expanded within Wikipedia's policies and guidelines)

The wiki principle Content is directly editable by the readers and changes are immediately visible.

Technical openness To enable third-party developers[2] to conduct development and research with ease, Wikipedia releases a significant amount of open data.

Each of the identified groups is covered further below in this section, including the implication each has on countervandalism.

1.2.1 Five pillars

Wikipedia's core editorial principles can be summarized into the five pillars. Although the exact phrasing of each pillar differs among the many different language editions of Wikipedia, the choice of pillars (and the idea behind them) is shared across all language editions of Wikipedia. All other policies or guidelines are built on top of the five pillars, giving them a specific meaning in the specific context of any particular project. Below is the text of the five pillars, as defined by the English Wikipedia: [10]

[1] Wikipedia community's resilience is fairly significant when it comes to refusing changes/features/ideas that it considers inappropriate. In 2014, the WMF decided to introduce a *superprotection* (make certain pages only editable by the WMF staff). Its first and only usage only lasted several days, and the superprotection was fully undeployed few months later. [9]

[2] Third-party as in outside of the Wikimedia Foundation staff members, not necessarily outside of the Wikimedia community.

Wikipedia is an encyclopedia "Wikipedia combines many features of general and specialized encyclopedias, almanacs, and gazetteers. Wikipedia is not a soapbox, an advertising platform, a vanity press, an experiment in anarchy or democracy, an indiscriminate collection of information, nor a web directory. It is not a dictionary, a newspaper, nor a collection of source documents, although some of its fellow Wikimedia projects are."[10]

Wikipedia is written from a neutral point of view "We strive for articles in an impartial tone that document and explain major points of view, giving due weight for their prominence. We avoid advocacy, and we characterize information and issues rather than debate them. In some areas there may be just one well-recognized point of view; in others, we describe multiple points of view, presenting each accurately and in context rather than as the truth or the best view. All articles must strive for verifiable accuracy, citing reliable, authoritative sources, especially when the topic is controversial or is about a living person. Editors' personal experiences, interpretations, or opinions do not belong on Wikipedia."[10]

Wikipedia is free content that anyone can use, edit, and distribute "All editors freely license their work to the public, and no editor owns an article – any contributions can and may be mercilessly edited and redistributed. Respect copyright laws and never plagiarize from any sources. Borrowing non-free media is sometimes allowed as fair use, but strive to find free alternatives first."[10]

Wikipedia's editors should treat each other with respect and civility "Respect your fellow Wikipedians, even when you disagree. Apply Wikipedia etiquette, and do not engage in personal attacks. Seek consensus, avoid edit wars, and never disrupt Wikipedia to illustrate a point. Act in good faith, and assume good faith on the part of others. Be open and welcoming to newcomers. Should conflicts arise, discuss them calmly on the appropriate talk pages, follow dispute resolution procedures, and consider that there are 6,641,095 other articles on the English Wikipedia to improve and discuss."[10]

Wikipedia has no firm rules "Wikipedia has policies and guidelines, but they are not carved in stone; their content and interpretation can evolve over time. The principles and spirit matter more than literal wording, and sometimes improving Wikipedia requires making exceptions. Be bold, but not reckless, in updating articles. And do not agonize over making mistakes: they can be corrected easily because (almost) every past version of each article is saved."[10]

Most of the pillars have a limited impact on countervandalism (they're designed to frame contributing to Wikipedia as a whole, rather than countervandalism itself), but some pillars determine what countervandalism currently looks like at Wikipedia. Any reforms made to it need to follow especially those pillars (but also all of the others); without that, the reform wouldn't be adopted by the Wikimedia community.

One of those important pillars is *Wikipedia is free content that anyone can use, edit, and distribute*, which requires Wikipedia to be editable by anyone. This makes certain obvious countervandalism solutions impossible to deploy. For example, it wouldn't be possible to implement a countervandalism solution which makes it impossible to participate for certain members of the Wikimedia movement.

Wikipedia's account creation form also reflects the *anyone can edit* principle. The only mandatory fields included in the form are the username and the password (an email address is frequently provided, but is not required), which enables everybody to edit without having to provide any sort of personally identifying information. This reinforces the *anyone can edit* principle, as it makes it easier to edit Wikipedia from countries with a limited freedom of speech. Wikipedians who choose to do so can still identify themselves[3], while wikipedians who need or want to keep

[3]For example, the author of the thesis voluntarily contributes under his own name

a certain level of privacy can have it, even from Wikipedia's functionaries.

The *Wikipedia's editors should treat each other with respect and civility* pillar requires editors to *assume good faith* (assume everyone comes to Wikipedia in order to help build it, rather than to break it down). This applies to countervandalism efforts as well: instead of mercilessly indefinitely blocking everyone who saves an edit which needs to be reverted, it is necessary to try to communicate with the revision author first, explain why it had to be removed and to only block when they refuse to follow the explained principles, policies and/or guidelines.

1.2.2 Technical openness

As further shown in Section 2.2.1, countervandalism depends on several tools, such as Huggle (Section 2.2.1.1), SWViewer (Section 2.2.1.2) or Twinkle (Section 2.2.1.3). Those tools are created by volunteer developers, i. e. developers who are not a part of the WMF staff. To enable volunteer developers to create and maintain tools, the developers needs access to information regarding the Wikimedia projects, such as the revision metadata and content, in a machine-processable way.

To provide the information/access the developers needs, the Wikimedia Foundation maintains a group of data services, as well several API endpoints, to allow seamless integration of external tools with the Wikimedia projects. [11] In this section, selected such data services are reviewed.

In addition to benefiting the tool developers, the amount of available data also benefits researchers – it is possible to freely use the WMF-provided data services in order to research Wikipedia itself[4]. [11] As detailed in Section 4.2.1, the author makes use of the dumps to research new techniques to combat vandalism in this thesis.

1.2.2.1 Wiki replicas: Quarry

For many tools (especially tools relating to countervandalism), real time access is a must-have requirement as edits need to be patrolled as quickly as possible. One of the tools providing real time access are the Wiki replicas. Wiki replicas contain a sanitized version of Wikimedia's production databases (the sanitization happens real-time). [12][13] In Figure 1.1, there is an example query, which generates a list of wikipedians based on who they thank and who they're thanked by. This can be used if any cliques[5] of wikipedians show up after clustering the thanks data. [14]

A significant advantage of Wiki replicas over API querying is higher developer freedom: if the API does not provide an endpoint returning data exactly in the desired format (or all the desired attributes), developers may decide to run an appropriate SQL query on the Wiki replicas to retrieve exactly the information they need. On the other hand, tools using the Wiki replicas might have to be updated more often, as they'd be affected by any schema changes happening in the production databases.

Quarry, depicted in Figure 1.1, is a tool allowing anyone[6] to connect to Wiki replicas via their web browser. Web-based access is often used for publishing and sharing SQL queries, but it isn't a feasible solution for an autonomous tool. Direct access to the replica SQL servers is possible from both the Toolforge cluster[7] and the Cloud VPS cluster. [12][15]

[4] Or the non-Wikimedia world, for example identifying topics of public discourse via Wikipedia's pageviews data.

[5] A *clique* is an informal group of people within a community, which reinforces its own interests (sometimes in contrary of the interests of the whole community).

[6] To be precise, anyone with an existing Wikimedia account.

[7] Platform-as-a-service cluster for tech-savvy wikimedians where they can deploy various tools to.

Figure 1.1 An example query ran via Wikimedia's Quarry interface. By Jan Spousta (2015), CC0, `https://quarry.wmcloud.org/query/5581`.

Social network of editors on cswiki

History ★ Unstar

Fork

This query is marked as a draft by Jan Spousta.

SQL

Toggle Highlighting

```
USE cswiki_p;
SELECT log.log_title as gets_thx, log.log_user_text as gives_thx, COUNT(*) as cnt, min(log.log_id) as log_id_min
FROM logging as log,
(
  SELECT log_title as user, COUNT(*) as cnt_gets
  FROM logging
  WHERE log_type="thanks" AND log_id > 1200000
  GROUP BY log_title
) as gets, /*those who were thanked by someone*/
(
  SELECT log_user_text as user, COUNT(*) as cnt_gives
  FROM logging
  WHERE log_type="thanks" AND log_id > 1200000
  GROUP BY log_user_text
) as gives /*those who thanked to someone*/
WHERE log.log_type="thanks" AND log.log_id > 1200000
  AND gets.cnt_gets > 10 AND gets.cnt_gets + gives.cnt_gives > 20
  AND gives.user = log.log_user_text AND gets.user = log.log_title
GROUP BY log_title, log_user_text
ORDER BY COUNT(*) DESC
LIMIT 10000;
```

All SQL code is licensed under CC0 License.

1.2.2.2 Wikimedia API

Another way to access data in real time is to make use of the Wikimedia API[8]. Using the API, countervandalism tool developers can consume data, just as they can with the wiki replicas. In addition to that, the API can be also used to perform changes at Wikimedia projects (in fact, it is the only way to do so, barring the regular editing interface).

1.2.2.3 Wikimedia dumps

Unlike the wiki replicas and the API, Wikimedia dumps offer bulk access to data about various Wikimedia projects. Dumps include information about many different aspects of Wikimedia: various metadata (list of existing revisions, lists of existing categories/files, etc.), the content itself (both as-of dump generation and historical per-revision data) or user behavior-related dumps (pageviews or user-navigation dumps[9]). [11]

The dumps are in many different data formats. There are XML dumps forming the native MediaWiki-generated dumps (containing revision metadata, revision content and certain project-specific metadata), SQL dumps (containing dumps of actual MediaWiki database, comparable to Wiki replicas described above) and TSV/CSV dumps (containing many different miscellaneous data like pageviews or analytics-ready precomputed datasets like `mediawiki_history`). [11]

[8] Strictly speaking, *the* Wikimedia API is not the right way to put this. Wikimedia operates a handful of different APIs[16] that can be used to integrate with the Wikimedia projects. The distinction among the different APIs is negligible, as this thesis does not focus on open data at Wikimedia but at countervandalism.

[9] An example is the clickstream dump, which contains information informing how users navigate between articles based on the referrer header (useful to determine which hyperlinks are used and which are not).

The last mentioned dataset, `mediawiki_history`, contains various pre-computed information in 70 columns regarding users (creation, promotion and blocking), revisions and pages. Unlike the wiki replicas, it is highly denormalized and it contains information from all Wikimedia projects. [17] This makes it easy to use for data analysis, such as the evaluation of effectiveness of various countervandalism techniques in this thesis. Section 4.2.1 covers additional implementation details.

1.2.2.4 Impact of technical openness on countervandalism

Unfortunately, not all users of the resources the WMF makes available to the general public are good faith. A side effect of the amount of data available is that malicious users wishing to disrupt Wikipedia with a great impact are able to target their attack better, based on information obtained from Wikipedia itself.

For example, it is possible to make use of Quarry (see Section 1.2.2.1) to see which templates are unprotected, but transcluded in a high amount of pages, to make a high-impact template vandalism (see Definition 1.4). While access to potentially dangerous queries can be restricted[10], often, the same query would be used both by a vandal and a trusted wikimedian to e. g. identify highly used templates that should be protected but aren't.

1.3 What is vandalism?

▶ **Definition 1.1.** Vandalism on Wikipedia is "editing (or other behavior) deliberately intended to obstruct or defeat the project's purpose, which is to create a free encyclopedia, in a variety of languages, presenting the sum of all human knowledge."[18]

Vandalism, in the traditional Wikipedia's definition, is *intentional* disruption of Wikipedia's content. When bad faith cannot be ascertained based on available evidence, edits that (potentially unintentionally) disrupt Wikipedia's content are refered to as *experiments*. [18]

For the purpose of this thesis, the difference between *vandalism* and *experiments* is not applied. This decision was made for two reasons. Firstly, the difference between vandalism and experiments largely matters only while communicating with the author of the problematic edit. This is because one of the Wikipedia's principle is to *assume good faith* of all contributors, unless clear evidence exists to the contary[19].

Secondly, it is technically challenging to truthfully and meaningfully determine the user's intention using the approaches discussed further in the thesis in Chapter 3. Considering the impact of each edit on Wikipedia's content is always either positive or negative (regardless of user's intention), it was decided to simply ignore the difference rather than trying to overcome the associated technical complexity descibed in the previous paragraph.

1.3.1 Types of vandalism

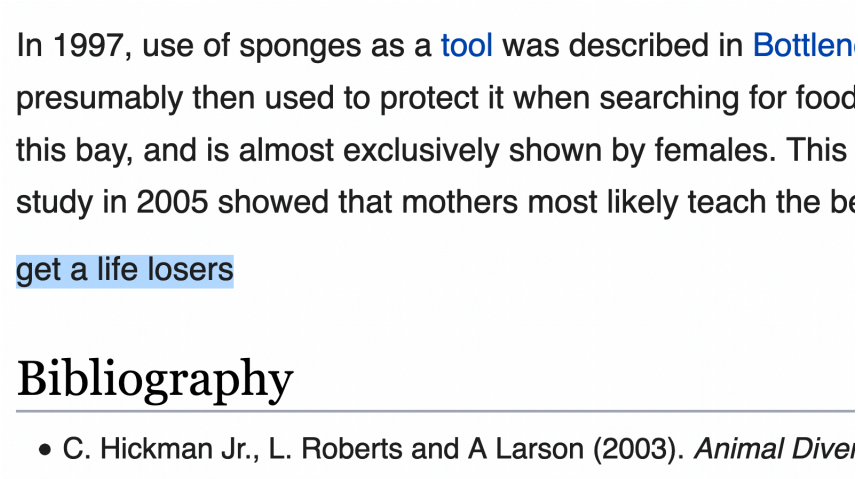
Wikipedia recognizes several types of vandalism. Some of those types have a greater harming potential than others, depending on the noticeabillity of the disruptive edit or visibility of the incident. Most important types of vandalism (including their properties) are covered in this section.

[10] In theory. All queries are potentially dangerous; when direct SQL access is allowed, it is nearly impossible to ensure no dangerous connection can be made in the published data.

1.3.1.1 Silly vandalism

▶ **Definition 1.2.** Silly vandalism is "adding profanity, graffiti, or patent nonsense to pages; creating nonsensical and obviously unencyclopedic pages, etc."[18]

Figure 1.2 An example of silly vandalism. Screenshot of a vandalized English Wikipedia article *Sponge*, CC-BY-SA 3.0.



Silly vandalism is by far the most common kind of vandalism. Many of the offenders are students, who spot the little *Edit* button in top right during their computer lessons and they decide to see for themselves whether it actually works[11]. In a way, this type of silly vandalism is useful for Wikipedia – it proves that the "anyone can edit" principle is a true statement – something that will hopefully be remembered by whoever committed the vandalism. A typical example of silly vandalism is shown in Figure 1.2.

Silly vandalism is often annoying, but by definition, it is easy to notice even for those with a limited familiarity with the topic. This makes it easier for Wikipedia patrollers to quickly notice the vandalism. As further covered in Section 2.2.1, people working in Wikipedia's first line of defense against vandalism generally aren't experts on articles they protect from vandalism. Defense techniques identified in the thesis will likely be effective against this class of vandalism.

1.3.1.2 Subtle vandalism

▶ **Definition 1.3.** Subtle vandalism is "vandalism that is harder to spot, or that otherwise circumvents detection, including adding plausible misinformation to articles (such as minor alteration of facts or additions of plausible-sounding hoaxes), hiding vandalism (such as by making two bad edits and reverting only one), simultaneously using multiple accounts or IP addresses to vandalize, abuse of maintenance and deletion templates, or reverting legitimate edits with the intent of hindering the improvement of pages."[18]

Subtle vandalism is one of the most problematic types of vandalism, as it requires a deeper knowledge of a particular topic to get spotted. A typical example of subtle vandalism is a slight change of factual information (exchanging one believable information for another one). [18] Subtle vandalism is more challenging to catch (as patrollers can't be expected to know everything).

[11] It does. The vandalising edit gets published and usually, is caught and reverted a few minutes after the fact.

Figure 1.3 An example of subtle vandalism. Screenshot of a vandalized Czech Wikipedia article *Cheb*, CC-BY-SA 3.0.

Sometimes, vandals decide to make up a full article; a 2012 example of this is discussed further in Section 1.3.3.1.

While subtle vandalism is potentially very dangerous to Wikipedia, it is remarkably difficult to reasonably detect, particularly due to the plausibility of the inserted documentation. The mechanisms discussed in the thesis will likely not be effective against subtle vandalism.

An example of subtle vandalism is depicted in Figure 1.3. In this case, a user changed the zip code for Czech city *Cheb* from (correct) `350 02` to (incorrect) `420 02`. Fortunately, incorrect zip codes are reasonably easy to notice (it is not something that changes regularly and zip codes are included in easily available databases). Other cases of subtle vandalism, especially hoax articles, often take years to notice; see more in Section 1.3.3.1.

1.3.1.3 Template vandalism

▶ **Definition 1.4.** "A template is a Wikipedia page created to be included in other pages."[20]
▶ **Definition 1.5.** Template vandalism is "modifying the wiki language or text of a template in a harmful or disruptive manner."[18]

Many Wikipedia's features are standardized across all articles. A typical example of this feature are infoboxes – on nearly every bibliographical page, it is possible to find a summary table near the top-right part of the article. This table is called an infobox and it provides the readers with a quick overview of the person described in the article. To ensure infoboxes (and other shared features) look the same on all articles, templates are employed.

Whenever an article author wants to include a bibliographical infobox, they call an appropriate template, provides it with the necessary parameters (such as the image, date and place of birth/death, occupation, awards received and similar) and the infobox is rendered in a standard way. In addition to provide a standard way of infobox inclusion, it also simplifies the procedure for making design changes (often, only the template needs to be changed; articles that use the template are updated automatically).

While templates are a critical functionality of Wikipedia, they also unlock an unique opportunity for the vandals. If they manage to find a highly-used template, they can quickly vandalize

hundreds or thousands of articles at once. While this makes template vandalism a serious problem[18], it is reasonably easy to prevent. Considering both the number of templates and number of experienced wikipedians is relatively low, page protection (see Section 2.3.3) can be effectively employed.

Template vandalism tends to be noticeable, as it doesn't add/change any of the template's features; instead, it disrupts articles the template is transcluded to. As such, defense mechanisms this thesis identifies will likely be effective against this class of vandalism.

1.3.1.4 Vandalbot

▶ **Definition 1.6.** "A vandalbot is a script which automatically performs some kind of malicious edit or similar operation to a wiki at high rate." [21]

Most of Wikipedia's vandalism is introduced manually, with limited intend to actually harm Wikipedia. As mentioned above, in many cases, pages are vandalized by schoolkids just to try whether it works. Even most of the vandalism performed by the so-called long-term abusers is done manually or semimanually.

However, there are cases when a user decides to intentionally disrupt Wikipedia, and writes a bot to insert malicious edits at a high rate, which is called using a *vandalbot*. The high rate of vandalism would make vandalbot attacks highly problematic. However, since whenever someone uses a vandalbot, they do it with a clear intend to disrupt Wikipedia, vandalbots attacks often happen at night. This further increases the impact of such attacks.

While vandalbots are realtively rare (and they require certain level of technical skills on the attacker's end), they're highly problematic, and have a high damaging potential. An example of a recent vandalbot attack is covered in Section 1.3.3.

1.3.2 Volume of vandalism

In 2007, a group of researchers from the University of Minnesota researched the impact of vandalism on Wikipedia. The researchers concluded that between that between September 2002 and October 2006, Wikipedia served 188 million damaged pageviews out of 51 billion total pageviews, i. e. 0.36 %. Furthermore, the researchers also stated that the probability that a reader encounters damaged article increased exponentially from January 2003 to June 2006. [22] This indicates that vandalism was an issue even when Wikipedia was still in its early days.

The volume of vandalism today depends on the language edition of Wikipedia that is examined. Larger editions have more readers and in turn more patrollers and more vandals, while smaller editions do not receive a significant amount of edits. According to the Statistics tool[12], Czech Wikipedia rollbackers reverted 10,494 edits between May 5, 2023 and January 25, 2023 (a total of 100 days). Considering rollback may be only used to revert vandalism[23], this roughly equals to 100 vandalism edits per day.

1.3.3 Examples of vandalism on Wikipedia

This section covers historical major instances of vandalism on English and Czech Wikipedias. Focus is given to cases that triggered certain attention in some form, such as an article in the news, were mentioned in a public lecture or with a similar public attention.

[12] Available from `https://statistics.toolforge.org/rollback?lang=cs&family=wikipedia&days=100`

1.3.3.1 2007–2012: The Bicholim conflict

In July 2007, an article about the *Bicholim conflict*, an armed conflict between the Portuguese and Indian Maratha Empire, was published on the English language Wikipedia edition. Two months after the article was created, it was labelled as a *good article*, a designation confirming the article follows a core set of editorial principles (it is well written, factually accurate, verifiable, neutral, etc.). The good article designation is assigned to less than 1 % of English Wikipedia articles. [24][25]

However, in December 2012, a Wikipedia user *ShelfSkewed* decided to give the article's sources a more thorough review. After finishing, they concluded that "After careful consideration and some research, I have come to the conclusion that this article is a hoax—a clever and elaborate hoax, but a hoax nonetheless." [26]. *ShelfSkewed* noticed that none of the books the *Bicholim conflict* article allegedly was based on appears to exist. As such, they decided to nominate the article for deletion. The nomination passed in late December 2012, resulting in the article's deletion from Wikipedia. [24][25]

The case of the *Bicholim conflict* is sometimes quoted as an example of subtle vandalism[27] (see Definition 1.3). However, it is not the longest-surviving hoax article at English language Wikipedia. According to Wikipedia's own list of identified hoaxes, this title belongs to *Method of focal objects* (a fictitious problem solving strategy), which was included in Wikipedia from April 2005 to September 2022 (ie. more than 17 years). [28]

1.3.3.2 2014: Satan's bolete

In November 2014, the Czech Wikipedia's article about *Satan's bolete* was changed. Prior to the change, it read "Satan's bolete is a *toxic* mushroom"; after the change, it read "Satan's bolete is an *edible* mushroom" (emphasis mine). The edit summary associated with the change reads "I'll change this back soon, I just want to prank my sister". [27]

This is a typical example of silly vandalism (see Definition 1.2) – nearly anyone with basic knowledge regarding mushrooms would be able to spot this edit. The damaging edit was removed a minute after it was inserted into Wikipedia. [27]

1.3.3.3 2021: Swastika covered thousands of Wikipedia articles

In August 2021, approximately 53 thousands of English Wikipedia articles were covered with swastika. A vandal noticed that a high-use template, *Wbr*, was not adequately protected, and changed it to cause a swastika to appear over the screen whenever it was used.

This is a classic example of template vandalism (see Definition 1.4). Unfortunately for Wikipedia, although the vandalism was removed a minute after its introduction, the template change managed to propagate across thousand of Wikipedia articles, catching attention of both international and Czech newspapers. [29][30][31]

1.3.3.4 2022: Polish Wikivoyage vandalbot attack

On September 02, 2022, the Polish Wikivoyage[13] experienced a major vandalism attack. At 00:21 UTC, the vandal made a few edits to gain the autoconfirmed rights (see Definition 2.2),

[13] Wikivoyage is one of the Wikimedia projects; it serves to provide a free traveller's guide.

which enabled them to move pages[14]. Several minutes after, they started to move various pages to randomly generated strings (along the lines of *ToWHk1*); this is shown in Figure 1.4. By 03:12 UTC, when the vandalbot attack finished, 1200 of pagetitles were changed to random strings (the account got blocked by 05:07 UTC by a Wikimedia Steward; see Definition 2.9).

Judging by the number of page moves performed by the user, this instance of vandalism likely was a vandalbot attack[15]. The damage caused by the attack was fully remedied by 17:15 UTC of the attack day.[16]

Figure 1.4 Record of the 2022 Polish Wikivoyage vandalbot attack, as shown via the *Special:Logs* special page.

- 01:27, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Tałas to Hif7D6 (revert) (thank)
- 01:27, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Powiat kłodzki to 6g7AsM (revert) (thank)
- 01:27, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Prespa to HCo2V4 (revert) (thank)
- 01:27, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page BF4kCt to 2nNjL5 (revert) (thank)
- 01:27, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Wołoczyska to VaS1nw (revert) (thank)
- 01:27, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Nowy Urengoj to OciMNf (revert) (thank)
- 01:26, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Jordanów to TPV5dm (revert) (thank)
- 01:26, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Moquegua (region) to PX2tyi (revert) (thank)
- 01:26, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Sinj to CdExfC (revert) (thank)
- 01:26, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Mamonowo to NILgpy (revert) (thank)
- 01:26, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Uciechów (województwo wielkopolskie) to ToWHk1 (revert) (thank)
- 01:26, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Województwo kujawsko-pomorskie to Jl8Ghu (revert) (thank)
- 01:26, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Bienszangul-Gumuz to 7lSpfk (revert) (thank)
- 01:26, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Wogezy (góry) to U6cdhF (revert) (thank)
- 01:25, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Ozurgeti to Rd9g9G (revert) (thank)
- 01:25, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Holon to BF4kCt (revert) (thank)
- 01:25, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Karaczajo-Czerkiesja to MujWtE (revert) (thank)
- 01:25, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Dunajowce to VdgCCi (revert) (thank)
- 01:25, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Šternberk to 2XmVcm (revert) (thank)
- 01:25, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Velenje to B99Q4J (revert) (thank)
- 01:25, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Enoturystyka to XwOGB2 (revert) (thank)
- 01:25, 2 September 2022 LebeDasLebenImmer (talk | contribs | block) moved page Amarusi to DcNMRX (revert) (thank)

[14] On all Wikimedia projects, changing an article title is refered to as *moving a page*.

[15] This is challenging to prove with certainty, especially without access to WMF's confidential access log records; however, certainty is hardly ever important for all practical purposes.

[16] This was determined using Special:Contributions page at the Polish Wikivoyage, using `Martin Urbanec` as the username of the Wikimedia Steward who remedied the attack.

Chapter 2

Countervandalism of today

As mentioned in the introduction of the thesis, countervandalism on Wikipedia today is mostly done via manual human labor. This is ineffective and it consumes human attention which could be better used in different focus areas that exist on Wikipedia. In this chapter, the countervandalism process, as well as related context, are covered. Information mentioned in this chapter comes from Wikipedia's policies, guidelines and help pages, which are publicly available from Wikipedia itself.[17]

2.1 Wikipedia's administrative structure

Countervandalism on Wikipedia is supported by various people, including wikipedians with no formal role and/or responsibilities, community-appointed functionaries as well as Wikimedia Foundation staff. This section defines roles fulfilled by people involved in countervandalism at Wikipedia, including the relationship of each of the roles with defense against vandalism.

The goal of this section is not to provide a complete overview of all exisisting roles within Wikipedia. This would be both impractical due to the significant amount of various groups[18], as well as irrelevant to this thesis's scope, as not all roles relate to countervandalism.

While reading this section, it is important to realize Wikipedia is not a hierarchically organized project. While certain users have higher level of access than others (and many times, users with higher privileges benefit from greater community respect), this does not grant privileged users any level of editorial content, or decision-making power outside of policy-mandated boundaries.

Wikipedia itself describes its system as follows: "Wikipedia's administrative tools are often likened to a janitor's mop, leading to adminship being described at times as being *given the mop*. Just like a real-world janitor might have keys to offices that some other workers are excluded from, admins have some role-specific abilities, but – also like a real-world janitor – they're not more important than the other editors" [32].

[17] An attentive reader might note that Wikipedia's policies and guidelines differ on each of the language versions. For the sake of clarity, unless otherwise noted, this chapter is based on English and Czech editions of Wikipedia.

[18] English Wikipedia recognizes around 30 different user groups; Czech Wikipedia around 25 user groups; this does not include global user groups that are recognized on all Wikimedia projects.

2.1.1 Ordinary users

▶ **Definition 2.1.** Ordinary user is any Wikipedia user that was not granted any advanced user permissions.

On Wikipedia, the vast majority of operations can be performed by any user without prior approval. Anyone can edit any page without asking for permission. [33] This extends to countervandalism as well, as ordinary users have access to all key parts of countervandalism on Wikipedia, as it works today (see further details later in this chapter), namely:

- ability to see recent edits via the `Special:RecentChanges` page or equivalent,
- view each edit individually, seeing exactly what was changed by which edit (by viewing the history), and
- edit pages in a way that effectively restores a prior version of the article (so-called "revert").

This enables anyone to participate in countervandalism on Wikipedia, with no prior permissions. Advanced permissions are only necessary for stopping an ongoing vandalism, for example, by blocking the offending user.

2.1.2 Autoconfirmed and extended confirmed users

▶ **Definition 2.2.** Autoconfirmed users have a certain minimum experience with editing Wikipedia and can move pages, edit semiprotected articles and are exempt from various edit filters. Users become autoconfirmed automatically when the user meets specified conditions, such as a number of edits or the length of user's tenure. Administrators may also confirm individual user accounts, manually granting them the same level of permissions as given to autoconfirmed users. A similar user group, although with stricter requirements, is called *Extended confirmed users*. [33]

Membership in the *Autoconfirmed users* and *Extended confirmed users* user groups is automatically conferred on users based on their tenure (number of days since registration) and their number of edits. It is used by several countervandalism and antiabuse mechanisms (most of which are covered further in this chapter), such as: [33]

- restricting the ability to move pages to autoconfirmed users,
- protecting certain pages so they can only be edited by autoconfirmed and/or extended confirmed users, or
- making certain edit filters to ignore autoconfirmed and/or extended confirmed users.

Precise requirements for a user to be considered autoconfirmed / extended confirmed vary from project to project. The default is at least 4 days of tenure (no edit count is required by default). At the English and Czech Wikipedias, users are required to have at least 4 days of tenure and at least 10 edits[33][34]; other Wikimedia projects may have more or less strict rules for granting autoconfirmed rights.[19]

While the requirements for both groups vary, the *Autoconfirmed users* user group exists at all sites that are powered by MediaWiki, including all Wikimedia projects and as such, Wikimedia functionaries may rely on its existence. The *Extended confirmed users* group is a custom group defined by site configuration, and is not guaranteed to exist everywhere (it only exists on Wikimedia projects where there is a use case for such a group).

[19] For example, the Chinese Wikipedia requires users to have at least 7 days of tenure and at least 50 edits to be considered autoconfirmed. [35] This is the strictest requirement of all Wikipedias.

2.1.3 Rollbackers and patrollers

▶ **Definition 2.3.** Rollbackers may "revert consecutive revisions of an editor using the rollback feature", eliminating vandalism with a single click. [33]
▶ **Definition 2.4.** Patrollers may mark pages created by others as patrolled (reviewed). On certain language editions of Wikipedia[20], patrollers may also mark individual edits as patrolled. [36]

Rollbackers are able to quickly remove vandalism that was added in one or more consecutive edits by the same user. For example, if a vandal edits the article *Prague* three times, each time removing one random word, rollbackers are able to eliminate all three edits by a single click, which speeds up vandalism removal. [23] However, as explained above, even users with no flags/advanced rights can remove vandalism with no issues. As such, rollbackers actually do not have access to a feature that's fully unavailable for ordinary users. For that reason, the role is usually thought of as a flag that's assigned to users deemed sufficiently trustworthy, rather than a formal role.

Patrollers mark pages (or on some language editions, edits) as patrolled, to record that the edit was reviewed by an experienced wikipedian. This helps wikipedians involved in countervandalism to avoid reviewing a page that was already reviewed by someone else, essentially saving their time. [36] While this ability can be only used by users who are patrollers, marking an entity as patrolled doesn't affect the project in any significant way, as whether a page is patrolled or not has no impact on its visibility. For that reason, just like rollbackers, patrollers are considered a flag confirming trustworthiness rather than a formal role.

The flags *Rollbacker* and *Patroller* are frequently (more frequently than other user rights described further in this section) granted to users at the same time. This happens because both are useful for countervandalism involved people; certain Wikimedia projects decided to merge both flags (making it technically impossible for a user to be a patroller without being a rollbacker, or vica versa). This similarity of both flags is the reason why both are described in the same section of the thesis.

2.1.4 Administrators

▶ **Definition 2.5.** Administrators are users with access to advanced technical features of the project, such as the ability to (un)delete pages, (un)block users, (un)protect pages and grant/revoke most access rights to other users. [33]

Administrators have access to site features which are not available to ordinary users, mostly for security reasons. As it would be too dangerous if everyone had the ability to delete any article (or worse, block other users), those technical abilities are reserved for administrators. [32]

Because of their advanced on-wiki access, administrators have a key role in countervandalism, as their abilities make it possible for them to enforce policy and their decisions. While rollbackers can issue a warning and ask the offending user to stop vandalism, if the offenders do not comply, an administrator has to step in and remove and/or restrict the user's editing capabilities. This also applies when a vandalism occurs in a form of a new page, rather than a change to an existing page. In those cases, an administrator has to delete the offending page from Wikipedia.

Whenever a non-administrator needs attention from the administrators, they need to request it. This can happen in several ways, including:

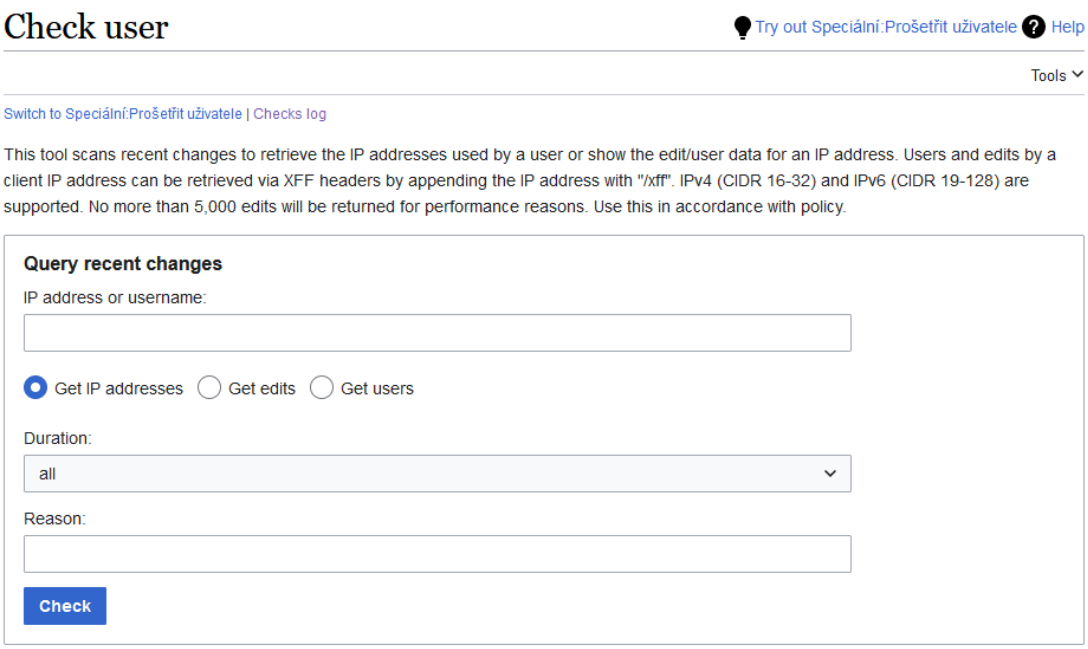
[20] Such as the Czech Wikipedia.

- a post on the Administrators' noticeboard (or a similar noticeboard-page)[21],
- inserting a special template to a page (for example, happens when nominating a page for deletion), or
- various off-wiki communication channels, such as IRC or instant messaging solutions.

For countervandalism to be effective, administrators need to attend the requests for administrative attention quickly, at any time of the day.

2.1.5 Checkusers

Figure 2.1 Screenshot of Special:CheckUser at Czech Wikipedia, April 09, 2023.



▶ **Definition 2.6.** Checkusers are a group of users with the ability to see IP addresses used by all users, including registered users, as well as the ability to see the list of contributors using a particular IP address (or contributors coming from a particular IP address range). [33]

By definition, checkusers can see private information about Wikipedia's registered users, namely, their IP address (usually, this makes it possible to determine the user's approximate location). The private information access is not unlimited though – checkusers may only access private information via a dedicated CheckUser interface (CU interface). Usage of the CU interface is logged and audited by the Arbitration Committee[22] and the Ombuds Commission[39], which aims to ensure user privacy is not violated.

[21] Usually, the main Administrators' noticeboard is called *Wikipedia:Administrator's noticeboard* and can be accessed by looking for that page using Wikipedia's search.

[22] Primarily on the English Wikipedia, where all arbitrators are checkusers and oversighters by policy and are better equipped to audit the CUs. [37] However, Arbitration Committees on all Wikipedias have the authority to remove CUs from their position. [38]

While this can be considered invading the privacy of the wikipedians, when a vandal uses a number of registered Wikipedia account, additional information is usually needed to effectively fight abuse. Checkusers assist antivandalism effort in the following two ways:

1. Determining whether two accounts are operated by one physical individual.
2. Placing blocks (further covered in a following section) based on information obtained from the CU interface.

Restrictions implemented by checkusers based on information from the CU interface may only be edited by other checkusers, to ensure proper review and decision making. Thanks to checkusers, it is possible to identify regularly abused IP ranges, and restrict editing from those ranges. [40]

To protect privacy of the users, information available to checkusers is retained for 90 days. [41]

2.1.6 Oversighters

▶ **Definition 2.7.** Oversighters are a group of users that can mark any material from Wikipedia (revisions, articles, files or similar) as hidden from all users, including administrators. [33]

Oversighters are responsible for ensuring private information is not posted on Wikipedia. Examples of private information include potentially libelious content or personally identifying information (such as birth numbers, SSNs and similar). Ocasionally, vandals decide to harass third-party users by humiliating them in public, sometimes, using Wikipedia. Since information posted by this kind of vandals is very sensitive, it needs to be hidden from the vast majority of users, including Wikipedia's administrators. To ensure oversighters do not give the information they hide away, they need to be over the age of majority and sign a NDA with the WMF. [42]

Like checkusers, compliance of the oversighters with the WMF's Privacy policy is monitored by the Ombuds Commission and Arbitration Committee. [37][39]

2.1.7 Global sysops

▶ **Definition 2.8.** Global sysops are an international[23] group of users with administrator access on the majority of Wikimedia projects[24], in order to assist with countervandalism and routine maintenance. [43]

Unlike all the other groups covered so far, global sysops are a global group, although their access is not enabled on absolutely all Wikimedia projects. This gives them the ability to act on many different projects, assisting many of the small language editions of Wikipedia[25]. Global sysops are appointed by Wikimedia Stewards (see Definition 2.9) based on community consensus. [43]

Global sysops carry out duties of the administrators (Definition 2.5) on projects that do not have an established group of administrators. [43]

[23] International as in appointed across different Wikimedia projects and their language editions.

[24] By default, global administrators have access on Wikimedia projects with less than 10 administrators or where less than 3 administrators made an action in the past two months. [43]

[25] Wikipedia exists in over 300 languages, which heavily differ in the number of articles and the community size, among other things.

2.1.8 Wikimedia Stewards

▶ **Definition 2.9.** Wikimedia Stewards are an international group of users with complete access to all Wikimedia projects, serving as local functionaries whenever a project did not appoint any, or local functionaries are inactive. [44]

Wikimedia Stewards have unlimited access to Wikipedia at all language editions of Wikipedia. Unlike all other roles mentioned above, they're elected by the global Wikimedia community on annual basis[26]. [44] They serve the Wikimedia projects in the following three main ways: [44]

1. acting on behalf of local functionaries that were not appointed,
2. performing emergency actions, when local functionares exist, but are unable or unavailable to act, and
3. making actions that impact all Wikimedia projects.

In the first case, Wikimedia Stewards temporarily act on behalf of non-existing local functionaries. Ocasionally, this also includes Wikimedia projects with a small number of local functionaries (for example, Wikimedia projects with only two administrators). Typically, this is the case of small Wikimedia projects which are not yet fully developped.[27] For certain local functionaries, the Wikimedia Stewards are authorized to delegate their substituting authority to other users, e.g. duties of the administrators are performed by both Wikimedia Stewards and global sysops on projects with insufficient amount of admins.

In their second capacity, Wikimedia Stewards protect the Wikimedia projects by carrying out an action that must be performed immediately, without waiting for attention from local functionaries. In the context of antivandalism, this can mean a major night vandalism attack, where all local administrators are currently asleep.

In their third capacity, Wikimedia Stewards act based on authority derrived from movement-wide elections. Local functionaries are appointed based on consensus of their local community (for example, the author was appointed a Czech Wikipedia administrator following a discussion among Czech Wikipedia participants; participants of the English Wikipedia did not have an opportunity to affect this) only, and as such, may only make decisions involving their own project. In the context of countervandalism, this mainly includes globally locking an account (rendering it unusable at all projects) or global blocks, further covered in Section 2.3.2.

2.1.9 Wikimedia Foundation staff

▶ **Definition 2.10.** WMF staff are paid employees of the Wikimedia Foundation, which hosts Wikipedia. As the legal owner of Wikipedia, the Wikimedia Foundation reserves a wide variety of powers, exercised by its staff members.

Following Wikipedia's principles covered in Chapter 1, the WMF staff rarely engages in matters ordinarily handled and decided by the community, including countervandalism. However, certain situations tend to require (or benefit from) a direct intervention of Wikipedia's legal operator (which is the WMF). Such instances include cases when: [46]

- community-taken actions showed to be ineffective, or
- legal considerations mandate an official action (such as DMCA compliance).

[26] Global wikimedia community includes participants from any and all Wikimedia projects.

[27] Typically, but not always. In 2023, the Wikimedia Stewards temporarily acted on behalf of German Wikipedia bureaucrats, as this position was not filled for a short while due to resignations. [45] The only action the Stewards performed in this way was the appointment of additional bureaucrats.

Actions carried out by the WMF are based on its Terms of use, where the WMF reserves the right to carry out investigations or terminate user access, among other things. [38] By carefully using its authority, the WMF and its staff help the community to address the most dangerous and problematic type of vandalism. The Wikimedia Stewards serve as the WMF's "first point of contact with the community" [44].

2.2 Countervandalism procedure on Wikipedia

Countervandalism procedure on Wikipedia can be divided into three different layers, handled by three different group of people. Each layer operates independently for each other and consists of different group of Wikipedia contributors. Namely, the three layers are formed by: [47]

1. patrollers, who focus on keeping Wikipedia free from vandalism,
2. editors, who monitor articles on topics they're interested in, and
3. readers, who sometimes notify Wikipedia contributors about statements that make no sense to them.

This section explains how the first two groups engage with countervandalism, and how they can benefit from changes suggested later in this thesis. The last group, the readers, is deliberately excluded – by definition, readers only notice vandalism casually (when they read an article that contains vandalism which was not yet spotted), and do not engage in countervandalism in a systematic way. It is only included in the above list for completeness.

2.2.1 First line of defense: Recent changes patrollers

Figure 2.2 Recent changes at Czech Wikipedia as of April 4, 2023, 09:30 UTC.

The first line of countervandalism defense is formed by the recent changes patrollers (in the remainder of this subsection, refered to as *patrollers*). Patrollers spend their Wikipedia time by monitoring *Recent changes*, which lists all changes that were saved on Wikipedia (an example screenshot of Recent changes is shown in Figure 2.2). Whenever a patroller spots a disruptive change, they revert it and reinstate the previous state of the article. Most of the patrollers hold the *Patrollers* and *Rollbackers* flags as defined in Section 2.1.3. [47]

In most cases, patrollers do not have a deep understanding on the topic they're reviewing. This is because the same group of patrollers reviews changes on all articles and no humans can have knowledge on everything. The majority of silly vandalism (Definition 1.2) cases are noticed at the patrollers layer, as those tend to be the easy to notice vandalism examples.

On larger wikis with many edits[28], patrollers use software tools like Huggle, SWViewer or Twinkle to deal with the number of edits to review. Those tools helps patrollers to patrol more effectively, by letting them navigate between revisions, user talk pages and other resources more effectively. Some countervandalism tools also lets users (if sufficiently privileged) to issue a block or to delete a page without ever leaving the tool, furthermore speeding up the countervandalism process. In the remainder of the section, the most important tools (namely Huggle, SWViewer and Twinkle) are introduced.

2.2.1.1 Huggle

Huggle (depicted in Figure 2.3) is a countervandalism tool developed by Adam Shorland, Petr Bena *et al.* Huggle loads and displays edits made to Wikipedia to its users, simplifying their review. It prioritizes edits that are potentially unconstructive; judgement is done based on user history (users considered trusted are put on a global whitelist; historically problematic users have a higher user-badness score). Final decision is left in the patroller's hands; the patroller decides whether to revert the edit or keep it intact. [48]

Whenever a user decides an edit is unconstructive and needs to be reverted, Huggle both reverts the edit and posts a standardized warning message[29] onto user talk page, to inform them about what happened. The level of the warning is picked automatically based on user's history. A default user warning message is used, but the patorller may optionally provide a detailed information about the edit; if that happens, Huggle picks a more informative warning message. [49][48]

2.2.1.2 SWViewer

SWViewer (depicted in Figure 2.4) is a countervandalism tool. Similarily to Huggle, SWViewer displays edit diffs to review for countervandalism; users decide about each edit whether it should be reverted or not. While the tool can be used to patrol a single wiki, unlike other tools, it is not restricted to a single wiki, which is a significant advantage of SWViewer. It is possible to subscribe to many different Wikimedia projects and patrol them all at once. This simplifies patrolling of small wikis which do not yet have a fully developped community by the global sysops and other global patrollers. [50]

2.2.1.3 Twinkle

Twinkle is an English Wikipedia gadget[30] used by many English Wikipedia patrollers. It aims to assist editors with common Wikipedia maintenance tasks, including tasks around countervandalism. For example, Twinkle provides an easy way to quickly: [51]

- nominate an article for speedy deletion,

[28] English Wikipedia has 5800 non-bot edits per hour, while Czech Wikipedia only has 100 non-bot edits per hour.

[29] Each language edition of Wikipedia has its own set of user warning messages. Huggle has a database of such messages, and uses them as appropriate/instructed.

[30] A gadget is a JavaScript program or a CSS snippet developed by Wikipedia editors themselves, rather than MediaWiki developers.

Figure 2.3 Screenshot of Huggle. By Petrb, CC BY-SA 3.0, `https://commons.wikimedia.org/w/index.php?curid=33194849`.

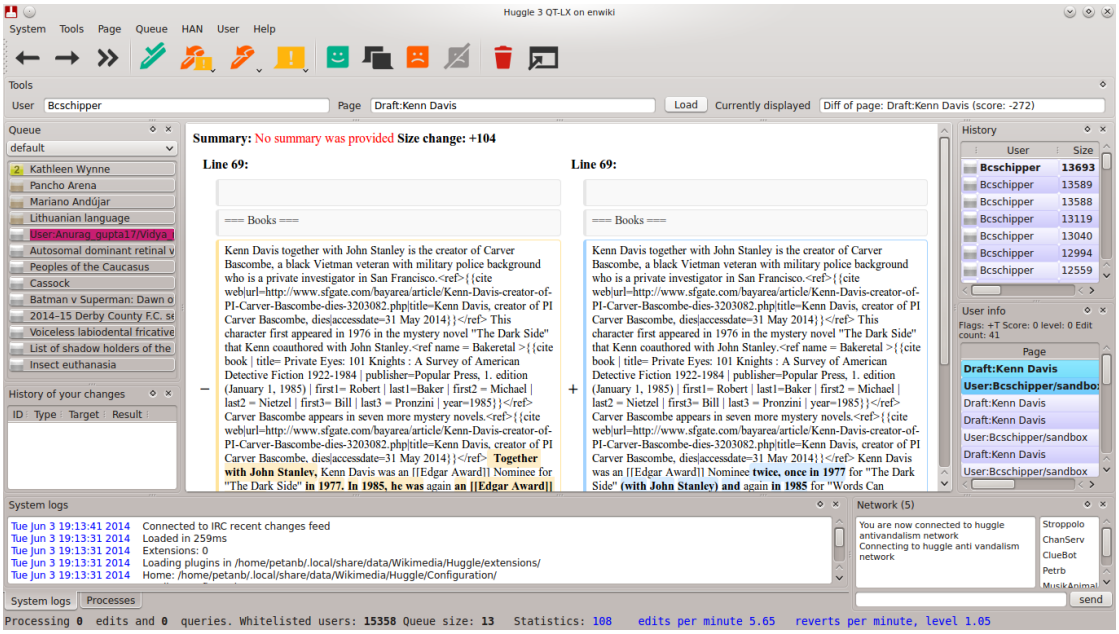
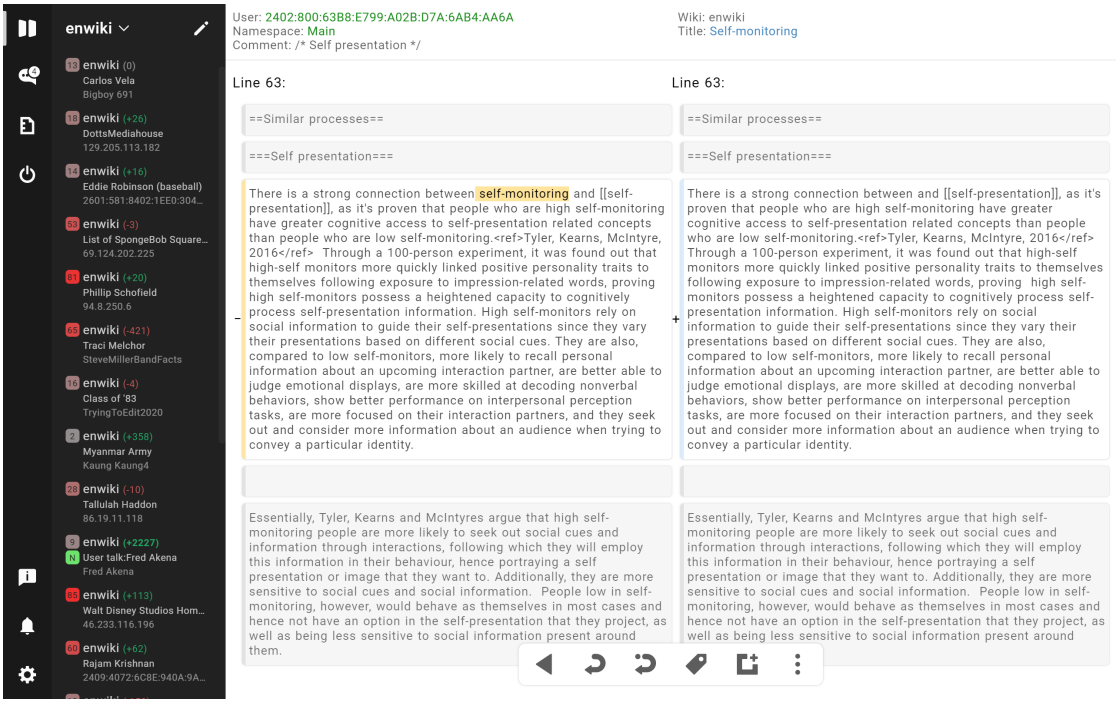


Figure 2.4 Screenshot of SWViewer. By Team SWViewer, CC BY-SA 3.0, `https://commons.wikimedia.org/w/index.php?curid=93407902`.



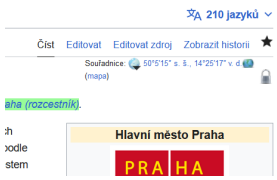
- post a warning message on a user talk page, or
- make a report to the administrators.

When patrolling, users often notice several examples of vandalism in a short period of time. However, even when that happens, each offending user needs to be notified immediately, to follow the *assume good faith* principle covered in Section 1.2.1. Twinkle makes sending a warning a matter of one or two clicks, instead of posting a warning message manually. [51]

2.2.2 Second line of defense: The watchlist

Any registered wikipedian has a watchlist, which is used to monitor changes happening on a certain subset of articles. The watchlist displays all recent edits made to the list of *watched pages*, as depicted in Figure 2.6. Wikipedians generally watch pages they created (or significantly expanded), to see how they develop over time (some wikipedians also watch articles they're interested in). Ocasionally, wikipedians spot that an incorrect edit was made to the article and decide to revert it. [52]

Figure 2.5 Wikipedia's *Watch this page* interface



This forms the second line of defense against vandalism. Vandalism edits that were not noticed by patrolling recent changes (as described in the previous section) can be noticed by wikipedians watching the affected page.[31] The watchlist line of defense is more effective at fighting against subtle vandalism (see Definition 1.3), as users having a page in their watchlist generally have at least a limited overview of the associated topic. [47]

Sometimes, patrollers who normally patrol recent changes intentionally add recently vandalised articles to their watchlist, especially if it is an article about a popular topic, where the popularity did not yet reach a level warranting a (semi)protection.

Figure 2.6 Wikipedia's watchlist

[31] Assuming the page in question is, in fact, watched by at least one active wikipedian. This is not always the case.

2.3 Technical solutions in use

To facilitate the countervandalism defense, MediaWiki and its extensions offer several technical solutions to combat with vandalism. Normally, any Wikipedia page can be edited by any user (both logged-in and logged out users); Wikipedia's countervandalism technical tools mainly focus on restricting this ability. In other words, there are tools that make it possible to prevent a particular individual from editing (blocks) or everyone from editing a certain page (page protection). Most technical tools described in this section can only be used by administrators.

2.3.1 Local blocks

Wikipedia employs several kinds of blocking mechanisms. Most of those mechanisms are *local*, which means they only affect the Wikimedia project they are used at. This section aims to describe what different types of local blocks are currently deployed. Some of the local blocks have a global equivalent (which affects *all* Wikimedia projects, regardless of where the action originates from), which are covered in Section 2.3.2.

2.3.1.1 Manual blocks

▶ **Definition 2.11.** "Blocking is the method by which administrators technically prevent users from editing Wikipedia. Blocks may be applied to user accounts, to IP addresses, and to ranges of IP addresses, for either a definite or an indefinite time, to all or a subset of pages"[40].

Manual local blocks (usually referred to only as blocks) are issued by Wikipedia administrators to technically prevent users from contributing to Wikipedia. Blocks do not exist to punish the offenders (or to retaliate), but exists merely to prevent disruptive behavior from happening. In other words, if a user vandalizes a Wikipedia article, they do not make any other edits, and the vandalism is noticed several hours after the vandalism was happened, it would not be acceptable to issue a block (as there is no disruptive behavior that would be prevented by the block). [40][53]

2.3.1.2 Autoblocks and cookie blocks

▶ **Definition 2.12.** "An autoblock is an automatic block of an IP address, done by the MediaWiki software. Autoblocks are the result of an attempt to edit the wiki from an IP address recently used by a blocked user, so that they may not make the same edits logged out or under a different username."[54]
▶ **Definition 2.13.** Cookie block is an automated block issued by the MediaWiki software based on presence of a block ID cookie, which is automatically set when a different type of local block takes effect. [54]

Unlike manual blocks, autoblocks are automatically introduced by the MediaWiki software when certain conditions are met. For example, when administrators manually block a user, MediaWiki automatically & temporarily blocks IP addresses recently used by the blocked user (unless an administrator directs otherwise). This makes it slightly more difficult to bypass blocks since merely creating a new account is not sufficient to circumvent a block. [54]

Cookie blocks are a specific type of autoblocks, to further increase the difficulty of block circumvention. When an administrator blocks a user account, MediaWiki sets a cookie within the user's browser. Upon receiving such a cookie, MediaWiki does not allow the user to edit.

Unless cookies are manually cleared (or a different browser used), this means blocked users are unable to edit even after moving to a different IP address. [54]

2.3.2 Global blocks

▶ **Definition 2.14.** "Global blocks are technical actions performed to prevent an IP address or range of IP addresses from editing all public Wikimedia wikis, for a fixed period of time or indefinitely. Global blocks disable account creation from the blocked IP by default, and can also prevent editing while logged in to an account."[55] Global blocks are issued by the Wikimedia Stewards. [55]

Similar to local blocks, global blocks are technical actions to prevent users from editing Wikimedia projects. Unlike the local blocks described in the previous section, global blocks take an effect across all public Wikimedia wikis, instead of just on one wiki. As of the time of writing, global block capabilities are significantly more limited than local blocks. Global blocks may only be applied based on user's IP address (usernames are not allowed), and they're not associated with any autoblocking (IP-based or cookie-based). [55]

Global blocks are issued by the Wikimedia Stewards mainly in cases of cross wiki abuse (when a user vandalizes several Wikimedia projects in a short period of time). Ocasionally, global blocks are also issued preventively; those blocks are described in more detail in the following paragraph. Global blocks are supposed to be only used sparingly, when no other methods (such as a combination of local blocks, page protections and similar) are effective. [55]

Preventive global blocks are issued to enforce the *No open proxies* policy, which was introduced in 2006. The policy prohibits[32] users from editing Wikipedia via "publicly available proxies (including paid proxies)"[56]. Over time, the term "open proxy" has been interpreted to include VPNs, web hosting companies and similar, as they make it possible for their user to mask their IP address. The policy was enacted because an IP address is one of the very few pieces of information Wikipedia receives about its users (as described in the preceeding paragraphs, blocks can only be targeted based on account name or IP addresses). [56]

2.3.3 Page protection

▶ **Definition 2.15.** Page protection is a MediaWiki feature allowing to protect certain pages from editing (or moving) by certain groups of editors. The protection can be applied indefinitely or for a specific period of time. Even when a page is protected from editing, anyone can view its wikitext source code. [57]

Page protection is a technical capability possessed by the administrator, which makes it possible to ensure a page is editable only by a certain group of wikipedians. Several types of page protection exists; their implications are described further in this section. Who can edit any given protected page is determined via the type of protection, and it ranges from *almost everyone* to *only administrators*. [57][58] Example of the page protection interface is shown on Figure 2.7.

Administrators protect pages for several reasons, including: [57][58]

- when one page is vandalized by many users,
- to protect highly-visible pages (such as the Main page) from getting vandalized, or

[32] Technically speaking, the policy doesn't prohibit *editing* from an open proxy; it merely authorizes blocks of any open proxies (paid or not) for any length of time. In practice, this makes editing Wikipedia via an open proxy nearly impossible

- to ensure legal responsibilities are met (concerns policy pages or license text).

Page protection shouldn't be used when an alternate solution (such as blocks) exists. [58]

2.3.3.1 Semiprotection

▶ **Definition 2.16.** Semiprotection is a type of page protection, which protects the page from editing by non-autoconfirmed users. [57]

When a page is either vandalized or at risk of getting vandalized and an administrator decides to protect it, they usually use of semiprotection. This level of protection does not prevent established wikipedians from contributing to the page, while it manages to prevent most naive attempts to vandalize it. Since the requirements for autoconfirmed accounts are generally fairly low (see Definition 2.2), it is quite easy to workaround if one wants to.

Semiprotection is automatically available in all MediaWiki installations by default. [59] As such, all Wikimedia project administrators can use semiprotection in their respective projects.

2.3.3.2 Extended confirmed protection

▶ **Definition 2.17.** Extended confirmed protection is a type of page protection which protects the page from editing by non-extended confirmed users. [57]

For cases where semiprotection becomes ineffective (as in, there are too many vandalism edits getting published despite the semiprotection), several Wikimedia projects[33] adopted the extended confirmed protection. This level of protection comes with similar benefits as semiprotection does (established wikipedians generally continue to be able to edit the page), while the level of efforts needed to get around the protection is significantly higher.

2.3.3.3 Full protection

▶ **Definition 2.18.** Full protection is a type of page protection, which protects the page from being edited by all non-administrators. [57]

When even the extended confirmed protection cannot be used (or is not available), or for pages that are extremely risky to edit (such as the Main page), administrators can decide to fully protect page. Fully protected pages can be only edited by the administrators themselves and as such, are nearly[34] guaranteed to avoid vandalism.

Often, full protection is enabled on Wikipedia's core templates, which are used on thousands of pages. Those templates are considered sensitive because of the impact template vandalism (see Definition 1.4) can have when performed on those pages. The templates are sometimes maintained by tech-savvy wikipedians who are not administrators. To preserve editing abilities for this group of people, some language editions of Wikipedia adopted template editors protection level. The idea of this protection level is similar to full protection, but it keeps editing open to the template editors (who are generally appointed by the administrators).

[33] For example, English and Czech language versions of Wikipedia both made this decision.
[34] Barring administrator account compromises, etc.

2.3.3.4 Pending changes protection

▶ **Definition 2.19.** Pending changes protection is a type of page protection which does not disallow affected groups from editing the page, but rather delays the publication of the edit until an authorized user reviews it. [57]

Pending changes protection is only enabled on a couple of Wikimedia projects – those with FlaggedRevs enabled. Since April 2017, the WMF does not allow FlaggedRevs to be deployed on any additional wikis. [60] Instead of *prohibiting* users from editing a given page, pending changes protection delays the publishing of edits by new wikipedians until an experienced user takes care of them.

2.3.4 Edit filters

▶ **Definition 2.20.** Edit filters is a tool allowing the administrators to configure that a certain action is taken automatically every time an edit matching pre-defined criteria is saved. This capability is used mainly to address common patterns of harmful editing. [61]

Edit filters are included in the *AbuseFilter* MediaWiki extension, which allows administrators[35] to configure an action to be taken every time an edit meeting an administrator-defined pattern is saved. For example, it is possible to automatically prevent unregistered editors from adding any external links, or automatically block any user who removes more than 2000 characters from any article. [62][63]

The following actions can be taken automatically by an edit filter (sorted roughly by impact): [62]

Tagging the edit When this action is used, the edit filter tags the matching edit for furhter review. Tags applied to an edit are visible in the Recent changes interface. By this action, an edit filter can be used to provide patrollers with further information about the edit.

Warning the user The user is warned that an edit filter matched their edit, including a warning message defined by the filter's author. They're given an opportunity to change the edit or to insist on the edit getting saved as prepared.

Throttling matching edits When this action is used, the edit filter only lets a certain number of matching edits to be saved. This can be used for patterns that are sometimes used when legitimate editing, but usually not in a batch (such as, significant content removal).

Disallowing the edit The edit filter does not allow the matching edit to be saved. However, the user has an opportunity to try again with a different edit.

Delaying automated promotion If the user is not autoconfirmed yet, they will not become one for a predefined period of time, despite meeting the conditions defined for autoconfirmed users are met.

Blocking the user When this action is used, the AbuseFilter automatically blocks the revision author for a predefined period of time, preventing not only the edit that matched the filter, but also all subsequent edits by the same user.

Edit filters are available at all Wikimedia projects and any administrator has the ability to edit them. However, not all administrators are sufficiently tech-savvy to be able to change the edit filters; as such, the capability is left unused on a significant amount of the Wikimedia projects. Filters can be also configured at the global level from Meta-Wiki by the Meta-Wiki's administrators, which is useful for cases of cross-wiki vandalism[63]

[35] And potentially, other sufficiently privileged users.

2.3.5 ClueBot NG

▶ **Definition 2.21.** "ClueBot NG is an anti-vandalism bot that tries to detect and revert vandalism quickly and automatically."[64] ClueBot NG only operates on the English language version of Wikipedia. [64]

ClueBot NG is a fully autonomous antivandalism bot. The bot calculates the probability an incoming edit is vandalism (based on a neural network); if the probability is higher than a predefined threshold, the bot automatically reverts the edit. [65]

2.3.6 ORES

▶ **Definition 2.22.** ORES is "a web service and API that provides machine learning as a service for Wikimedia projects"[66]. The system is designed to help automate critical wiki-work, such as countervandalism. [66]

ORES is designed to automate certain critical Wikipedia workflows, including countervandalism. It provides two types of scores: edit quality and article quality. Edit quality is detected via two models, the `reverted` model, which predicts whether an edit is going to be reverted and the `goodfaith` model, which predicts whether an edit was saved in good faith. The `reverted` model is closer to the thesis's aims. [66]

2.4 Evaluation of currently used antivandalism techniques

In this chapter, the currently employed antivandalism techniques are evaluated. Countervandalism techniques described above can be split into the following four main groups:

1. tools preventing certain *users* from editing *any* pages,
2. tools preventing (nearly) *all* users from editing a *particular page*,
3. tools (such as ORES) augmenting patroller's view, and
4. fully automated countervandalism bots.

Each of the four groups listed above are evaluated independently, focusing on its advantages and disadvantages, based on Wikipedia's workflows, practices and principles described in the preceeding chapters.

2.4.1 User-blocking techniques

User-blocking techniques include techniques focusing on disallowing a certain contributor from contributing. Ideally, user-blocking techniques should prevent a particular *physical user* from contributing. Those techniques mainly consist of local or global blocks (including blocks issued manually by the administrators, as well as blocks issued by the MediaWiki software). As described above, blocks can be either targeted based on the contributor's username or their IP address. However, neither of those possibilities does a good job at identifying a particular physical user.

If a block is configured using the contributor's username, it is trivial to bypass by merely creating a new account, assuming one ignores the consequences such blocks have, like autoblocks. When

the autoblocking mechanism kicks in, the block converts from a username-based one to an IP-address one.

Blocks based on the IP addresses are more difficult to bypass (partially due to Wikimedia's no open proxies policy described in Section 2.3.2). However, because of the rise of IP masking services (such as Apple's *iCloud Private Relay* or Cloudflare's *1.1.1.1*), which are either turned on by default, or advise the users to turn it on to increase their privacy, it is increasingly problematic to rely on IP addresses as the sole source of identification of Wikimedia users.

As the WMF traditionally does not want to implement privacy-invasive solutions such as fingerprinting, one of the core issues associated with user-blocking techniques is their inability to identify the target.

2.4.2 Page-focused techniques

As opposed to user-blocking techniques described in the previous subsection, page-focused techniques are designed to prevent any (or most) contributors from editing a particular Wikipedia page. By definition, this makes all page-focused techniques easy to bypass if the user's goal is merely to disrupt Wikipedia in any way – it is sufficient to pick a new Wikipedia page to disrupt. On the other hand, page-focused techniques are effective for highly-visible pages, such as articles about politicians or other high-profile individuals.

2.4.3 Augmenting tools

Augmenting tools improve interfaces used primarily by the recent changes patrollers using the workflows described in Section 2.2.1. Tools like ORES display additional information to the patroller, which can then be used as part of the *Is this non-constructive?*[36] judgement patrollers have to make. While augmenting tools can present valuable information, they do not make any actions on their own.

2.4.4 Fully automated countervandalism bots

Fully automated countervandalism bots such as ClueBot NG (see Definition 2.21) automatically determine whether incoming edits are vandalism, and if they determine them as such, they revert the vandalism. This happens with no direct human involvement. As those bots are fully automated, they help with protecting Wikipedia against vandalism 24/7, which improves the coverage.

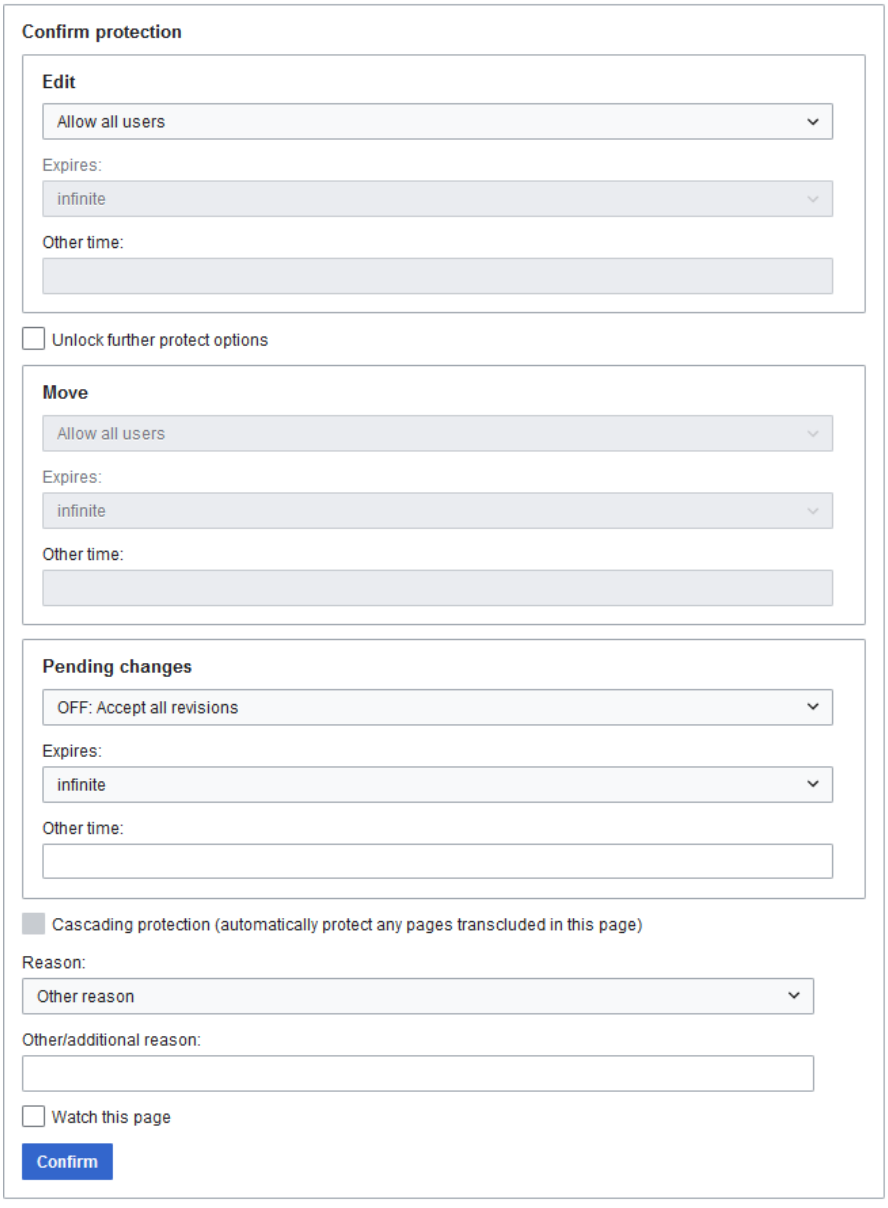
However, as noted in the ClueBot NG's FAQ, "Setting up ClueBot NG can be a complex process"[65]. The process to set ClueBot NG up requires generating a large dataset specific to the wiki in question, which includes both constructive and vandalism edits. The maintainers recommend using at least ten thousands of both types of edit, to allow ClueBot NG to learn based on those edits. [65]

Because of the large investment that'd need to happen, fully automated countervandalism bots are not adopted by many wikis. In addition to that, it would be problematic to use such bots to protect a large group of wikis, for example, to take some workload off from Wikimedia Stewards and Global Sysops (Definition 2.9 and Definition 2.8 respectively). Similar considerations apply to other fully automated countervandalism bots as well.

[36] As dictated by the *Assume good faith* principle, patrollers presume edits are constructive.

Despite the facts stated above, on wikis that adopted them, countervandalism bots are a critical feature to revert vandalism quickly. When analyzing ClueBot NG's outages (ranging in length from a couple of days to several weeks), Geiner concluded in 2013 that the overall time-to-revert edits was almost doubled when the bot was not in service. [67]

Figure 2.7 Page protection interface.



Chapter 3

Future countervandalism techniques

In this chapter, new techniques that could be used to combat vandalism are suggested. To limit the scope of the research in this thesis, only techniques based solely on edit metadata information are discussed (in other words, techniques requiring access to the edit diff are out of scope for this thesis). This decision was made to ensure the techniques are easy to review and implement.

3.1 Techniques to review

Four different countervandalims techniques are identified in this section. Each identified technique is briefly described in this section (including an indication whether a particular technique was tried in the past).

Techniques are suggested primarily based on the issues described in Section 2.4. The amount of data each technique needs to operate is taken into account as well, as each technique needs to be: (a) implementable, as it cannot make use of data Wikipedia does not have, and (b) testable (only publicly available data can be used during the review).

In contrast to other automated already-existing approaches, like ORES (Definition 2.22) or ClueBot NG (Definition 2.21), techniques tested in this thesis only make use of edit metadata, rather than the edit content itself. This is done to simplify scaling the solutions to other Wikimedia projects in the future (as the structure of edit metadata is fixed across all Wikimedia sites).

3.1.1 Requiring participants to identify themselves

One of the identified issues is that Wikipedia's blocks are based purely on IP addresses, and that their reliability decreases with time. As such, it is increasingly difficult to prevent users from committing additional vandalism, once the first one has been identified.

This issue can be resolved 100 % precision[37], by requiring Wikipedia contributors to connect their Wikipedia account with their real life identity. If the administrators decide to block the user,

[37] Nearly. Abusing multiple accounts has two main forms. One is called sockpuppetry, where one user creates a handful of Wikipedia accounts. The other form is called meatpuppetry, where a Wikipedia editor canvasses

they are unable to register a new account for the duration of their block, dramatically increasing the effectiveness of Wikipedia's blocks.

For reasons described in Section 4.1.1, this solution was never deployed on any Wikimedia projects.

3.1.2 Disallowing contributions by logged out users

Another approach to increase the effectiveness of Wikipedia's blocks is to make it more difficult for vandals to bypass the block. As described in Section 2.3.1.2, MediaWiki includes several mechanisms aiming at making bypassing blocks by creating another account more difficult. However, those approaches are only used for logged in accounts.

This leads to the second possible technique: disallowing contributions by logged out users (and making it technically impossible to edit without first creating a Wikipedia account). Assuming only registered accounts were used to edit, MediaWiki's native mechanisms to avoid block circumvention by blocked registered account holders would be automatically used for any and all blocks. Requiring registered accounts would also make it slightly more time expensive to edit (including malicious edits), as it would be necessary to first create an account.

Disallowing unregistered contributors was tried at Portugalese (in 2020)[68] and Farsi Wikipedias (in 2021)[69]. Partial restrictions were imposed in 2011 at the English Wikipedia, where unregistered contributors may edit, but they may not create new articles[70].

3.1.3 Presence of user IP addresses in blacklists

Instead of improving the effectiveness of the currently existing (manually introduced) blocks, one can decide to focus on adding new sources of blocks into the system. In other words, rather than relying on administrators to manually block offensive users and IP address, Wikipedia's software can be instructed to automatically block certain IP addresses meeting certain conditions.

For example, it is possible to automatically block IP addresses that are present on an IP blacklist. Those blacklists are already maintained for other purposes, such as blocking delivery of e-mail spam. It is possible that some of the vandals already send spam (and vica versa). If that is the case, incorporating existing IP blacklists might decrease the number of manual blocks that need to be issued. For the purpose of the thesis, StopForumSpam blacklist is used, which records IP addresses of spammers attacking forums, blogs or other wikis[71].

To ensure there are available data to evlauate this technique, it will only be evaluated on contributions by logged out users, where the IP address of the user is publicly available[38].

As of April 2023, this solution is being tried at several pilot Wikimedia projects via the *StopForumSpam* extension. [72]

3.1.4 Number of reverted edits

From time to time, one user saves more than one malicious edits to Wikipedia, in a short timeframe. This is especially true for vandal bots (see Definition 1.6), but to a limited degree,

their friends to create an account on their own, and to take a certain action on the wiki. Requiring participants to ID would help with sockpuppetry, but not meatpuppetry.

[38] This is about to change with the IP masking project.

it happens with other kinds of vandalism as well. This leads to another type of automated countervandalism approach: prevent users from editing if *too many* of their previous edits were reverted. In other words, if a user made four edits within an hour and all of them were reverted by a human editor, prevent them from saving subsequent edits.

The effectiveness of the solution depends on how *too many* is defined. The definition of *too many* consists of two thresholds: (a) maximum tolerated number of previously reverted edits and (b) length of the period in which the number of previously reverted edits is calculated. Both thresholds can be determined independently on each other; their combination determines the effectiveness of this technique.

More edits are prevented if the value of the first theshold is *decreased*. In a hypothetical case when it is set to `0`, all edits will be prevented (regardless of the second threshold value), as all editors have *at least zero* previous edits reverted. The lowest reasonable value of the threshold is one, in which case any recent reverted edit will block the user from successfully submitting an edit. This is not ideal, as wikipedians can make mistakes and ocasional reverted edits are not an issue. On the other hand, wikipedians should not have a high number of reverted edits either. A reasonable value that is not too low or too high is five, which is used for the experiment conducted in this thesis.

The second threshold has an opposite effect: More edits are prevented if the value of the second threshold is *increased*. To ensure the technique does not take years old events into account[39], the value cannot be set to a value that's *too high*. As stated in Section 2.1.5, the WMF keeps private information about its users for 90 days, which seems like a reasonable threshold for this countervandalism technique as well.

To summarize: *Too many* is defined as at least five reverted edits in the last 90 days.

3.2 Methodology used to review techniques

Techniques identified in Section 3.1 are reviewed in two rounds of review, to ensure each idea is both appropriate with regards to the Wikipedia's principles and norms, and that it is indeed effective in preventing a sufficient percentage of vandalism.

3.2.1 First round of review: Suitability and compatibility

The first round of review is based on suitability and compatibility. Each of the suggested techniques is evaluated with regards to Wikipedia's philosophy formulated by principles covered in Chapter 1, focusing on:

- the technique's effect on the openness of the Wikipedia community,
- Wikipedia's norms with regards to collecting/using personally identifying information, and
- effect on Wikipedia's compliance with its five pillars (especially the *anyone can edit* principle).

The first review is conducted via discussing the points mentioned above in context with each of the suggested techniques.

[39] Humans change and this is reflected even on Wikipedia. In the past, users who were indefinitely blocked later became administrators on their own.

3.2.2 Second round of review: Viability

Techniques passing the first round of review are subjected to the second round, where their viability is evaluated. This is done to ensure that techniques recommended by the thesis prevent a sufficient amount of vandalism, while not preventing a significant amount of constructive edits. This is important, as Wikipedia would only implement solutions that prevent vandalism and only vandalism. A small false-positive rate is nearly-impossible to avoid, but it should be kept within reasonable bounds.

Viability evaluation is based on a random sample of 100 actual Czech Wikipedia edits. The sample was generated from all Czech Wikipedia edits made in 2022, except edits that:

1. were made by a bot,
2. were made outside of the article/main namespaces (were *article edits*)[40], or
3. were not made by an autoconfirmed user (see Section 2.1.2 for definition).

Bots were excluded, because Wikipedia bots have to be approved by a bureaucrat in order to be recognized. As such, it is highly unlikely edits made by an authorized bot are vandalism. Theoretically it could happen in the unlikely[41] case of a Wikipedia bot operator going rogue; however, considering this thesis's focus, those edge cases fall out of scope.

Edits made outside of the article namespaces were excluded, because edits in the article namespace and in other namespaces generally have an entriely different purpose. Article edits are intended to add/modify content, while edits in other namespaces often have different purposes (such as, discussion around changes in Wikipedia's norms, communication among users and similar). Vandalism-related patterns will likely be different in both namespaces.

Only edits made by non-autoconfirmed users (see Definition 2.2) are included. This is done, because most easy-to-detect vandalism is submitted by new contributors (either logged out, or having a new user account). As discussed in Section 1.3.1, accounts existing for a while are more likely to commit other types of vandalism (which are harder to detect), like subtle vandalism (Section 1.3).

For the purpose of evaluating *Presence of user IP addresses in blacklists*, a separate random set of edits is generated, where the third condition is replaced with *were made by a logged in user*. This is done because IP addresses are only available for contributions made by logged out users.

Viability evaluation is done manually. First, the sample is reviewed by the author and other experienced wikipedians. Each edit is classified by exactly one expert wikimedian as either vandalism or constructive (the classification is binary; it is not possible for an edit to be both or neither; additionally, it is not possible to skip an edit during the evaluation, to ensure the sample size stays at 100 edits). Then, techniques suggested in Section 3.1 are ran on the sample.

3.2.2.1 Calculation of F1 score, recall and FPR

Based on acquired data and the manual vandalism classification, the F1 score, precision, recall and false-positive rate are calculated based on the math formulas shown below: [73]

[40] Wikipedia divides its content into many different namespaces. For example, articles are included in the article namespace (ID 0), while discussions with Wikipedia's contributors is included in the User talk namespace (ID 3).

[41] Unlikely, but not impossible. In August 2017, an administrator of Wikimedia Commons suddenly started deleting random files. This resulted in their prompt desysopping; shortly afterwards, they were globally banned by the community and subsequently by the WMF.

$$Precision = \frac{\#\ of\ True\ positives}{\#\ of\ True\ positives + \#\ of\ False\ positives}$$

$$Recall = \frac{\#\ of\ True\ positives}{\#\ of\ True\ positives + \#\ of\ False\ negatives}$$

$$F1 = 2\frac{Precision * Recall}{Precision + Recall}$$

The precision, recall and F1 metrics are used instead of accuracy to account for class imbalance. In the sample generated in the previous section, there are fewer vandalism edits than constructive edits.

Chapter 4

Results and discussion

This chapter describes the results acquired via the review described in the previous chapter, including the evaluation of the results. It also recommends one solution to be implemented and estimates how difficult it would be to implement it, especially with regards to the necessary engineering effort and possible side effects.

4.1 First round of review: Suitability and compatibility

In this section, the first round of review (as defined in Section 3.2.1) is conducted. All four techniques are discussed and a conclusion is made whether implementation would be possible without violating any of Wikipedia's principles or policies.

4.1.1 Requiring participants to identify themselves

To implement this idea, the WMF would have to store (and verify) personally identifying information about the wikipedians, such as the scans of their official ID cards. Effectively, this would allow the WMF to tie each wikipedian to their real-life identities.

While this might sound like an acceptable tradeoff, it is important to note that Wikipedia does not operate only in countries where the freedom of speech is guaranteed. Wikipedia also operates in countries like Russia, Belarus, Saudi Arabia or China, where the freedom of expression is either limited or nonexistent. Since Wikipedia keeps the same principles even in those countries, wikipedians who edit from those countries risk getting arrested, imprisoned or executed for editing Wikipedia. This is not a theoretical risk; arrests/imprisonments/executions of Wikipedia contributors happened in the past:

- Mark Bernstein (one of top 50 wikipedians participating in Russian Wikipedia) was arrested in 2022 by Belarusian authorities. *The Verge*, while being unable to determine what exactly was Bernstein charged with, stated that Bernstein was "accused of editing Wikipedia articles about Russia's invasion of Ukraine"[74].
- Wikipedians Ziyad al-Sofiani and Osama Khalid were arrested in 2020 by Saudi Arabia's authorities and later sentenced to 8 and 32 years in prison respectively. According to *Ars Tech-*

nica, their charges included "swaying public opinion and violating public morals by posting content deemed to be critical about the persecution of political activists in the country". [75]

Wikipedians editing in such countries need to be able to participate anonymously to decrease a risk of their identification leaking from those with access (hypothetically, this could be the WMF staff and possibly CheckUsers/Stewards, who currently possess access to IP addresses of all participants[42]).

Considering the facts described above, requiring wikipedians to identify to the WMF[43] would break the third pillar – *Wikipedia is free content that anyone can use, edit, and distribute* – which introduces the *anyone can edit* principle. As such, this solution **fails** the first round of review, and will not be considered further.

4.1.2 Disallowing contributions by logged out users

In order to be able to disallow contributions by logged out users, one only needs to know the logged-in status at the edit time. This is already the case, as edits are either attributed to username (when logged in) or to user's IP address (when logged out). This means no new information need to be collected from the users (or their devices) for implementation to be possible.

By its nature, disallowing contributions by logged out users makes it more challenging for the users to edit – instead of merely clicking the *Edit* button, one has to create an account. In accordance with the *anyone can edit* principle embodied by the Wikipedia's five pillars, accounts can be created without providing any identification information (only the username and password are mandatory, where the username is only a string showing how the wikipedian wishes to be called). Because of that, requiring all wikipedians to fill the account creation form is not an undue burden in terms of the amount of information required (similar to arguments shown in the previous section). Due to the low amount of information the sign up form requires, filling only takes a short amount of time. This means mandatory account creation cannot be interpreted as an excessive burden even in terms of the amount of time it takes.

As such, in the absence of arguments to the contrary, one has to conclude that this solution **passes** the first round of review, and are considered further in the thesis.

4.1.3 Presence of user IP addresses in blacklists

Implementation of this technique consists mostly of periodically comparing IP addresses of editors with a predefined set of blacklists. This can be accomplished in two ways:

1. online verification: send IP addresses to the blacklist operators on every edit, letting them to score/flag the edits in real time as edits are made,
2. offline verification: periodically download a dump of blacklisted IP addresses and compare the IP addresses using the offline dump.

First solution benefits from the quickest possible data update (assuming the APIs of the blacklist operator use the freshest possible data), but it endangers user privacy. Since Wikimedia is inherently open on contributions-level data, it is possible to see when was each edit made with

[42] IP addresses of registered Wikipedia users are considered as protected by the WMF's privacy policy. Should this technique be implemented, the status of identification information would likely be comparable with other already privileged information, hence the comparsion.

[43] Or other bodies.

a very high precision. Considering that verification would happen online whenever an edit is made, blacklists operators would have a timestamp for each IP, which would precisely correspond to the edit timestamp. Effectively, they'd be able to associate IP addresses with wikipedians, creating an undue risk for user privacy. This issue is resolved in the second approach, where verification happens offline, without ever transferrng user data outside of the WMF's servers.

In theory, shifting the responsibility for Wikipedia-used blocks to third parties might cause issues of users being wrongfully rejected from participation without any possibility to request review. Since Wikipedia's functionaries generally aren't able to modify third party blacklists, this means Wikipedia's functionaries have to retain the authority to override the blacklist-provided blocks on case-by-case basis (for instance, by granting account-specific exemptions).

Concerns mentioned above might cause the technique to be unsuitable for Wikipedia. However, all those concerns are resolvable by implementation means and are not fundamental by nature. As such, this solution **passes** the first round of review.

4.1.4 Number of reverted edits

This technique fundamentaly consists of making automated judgements about accepting user's edits based on their previous Wikipedia contributions. Considering the WMF releases contributions data to the world, such analysis can always be performed, either by the WMF itself or by third party researchers. This aspect of Wikipedia contribution is denoted in the WMF's Privacy policy, which explicitly states that Wikipedia contributions may be analyzed by anyone, even if this leads to (otherwise private) conclusions about Wikipedia's users. [76]

Because anyone is able to conduct an analysis based on Wikipedia's contributions data, there is no reason why the WMF shouldn't have this ability. As such, this solution **passes** the first round of review.

To summarize the conclusions drawn above, all proposed techniques except *Requiring participants to identify themselves* pass the first round of review. While some of the other techniques come with important concerns, such concerns are not fundamental in nature and can be resolved during the implementation phase. As such, all those solutions are reviewed for viability in the following section.

4.2 Second round of review: Viability

The viability review consists of four stages:

1. generating the data to evaluate from the Wikimedia Dumps,
2. manually classifying all 200 edits[44] as constructive or vandalism,
3. simulating each proposed technique using the data acquired, and
4. evaluation.

Each of those steps is described as part of this section.

[44] As noted in Section 3.2.2, there are two samples, each having 100 Czech Wikipedia edits (and only one is used for evaluating each technique). Hence, each technique is evaluated based on 100 random edits, but it is necessary to classify all 200 edits.

4.2.1 Preparation: Generating the data

All techniques suggested by the thesis rely solely on edit metadata, namely:

- revision author (for *Disallowing contributions by logged out users* and *Presence of user IP address in blacklists*), and
- prior edits made by the revision author (for *Number of reverted edits*).

Both the revision author and the user's prior edits are available in the `mediawiki_history` dataset, which can be downloaded from Wikimedia Dumps. The dataset is published as one or more TSV datafiles (depending on the project's size). New snapshot of `mediawiki_history` is generated every month, and each snapshot contains information about all revisions that were made up to the snapshot generation date. Data for the Czech Wikipedia (which is the main subject of this thesis) are released as one TSV datafile per year.

As discussed in Section 3.2.2, the viability review is based on a random sample of 100 edits meeting certain conditions, which need to be taken care of. The random sample is generated using the `mediawiki_history` dataset described above, namely, the 2022 data taken from the `2023-03` snapshot.

The `mediawiki_history` dataset is processed via *Apache Spark*, an open source engine designed for processing big data[77]. The inspiration on how to process the data files was taken from [78]. *Apache Spark* allows users to process a set of TSV files via a SQL query, which is more convenient than writing code to process large TSV files.

Data to analyze are generated by first creating a list of all revisions meeting the criteria from Section 3.2.2 (irrespective of logged in status of the editor). This is done using a SQL query shown in Code listing 4.1. This list is then postprocessed via Python's *pandas* library, to generate both sampled lists defined in Section 3.2.2. The postprocessing is shown in Code listing D.1.

Complete generated list (after sampling) is available in Appendix A for the all-revisions list and Appendix B for the sample of revisions by logged out.

4.2.2 Baseline evaluation

Before evaluating the individual techniques, the sample edits were classified by experienced Wikipedia administrators and patrollers as either vandalism or constructive. The classification (along the list of sampled data) can be seen in Appendix A.

Summary of the observations is available as Table 4.1 (for the sample of all edits) and Table 4.2.

Table 4.1 Summary of baseline vandalism classification – all edits

Variable	Value
Total edits	100
Vandalism edits	42
Reverted edits	44
Reverted and not vandalism	5

Code listing 4.1 SQL query getting all edits meeting the defined criteria

```
SELECT
  event_timestamp,
  revision_id,
  revision_parent_id,
  page_title,
  event_user_text,
  event_user_revision_count,
  event_user_registration_timestamp,
  revision_is_identity_reverted,
  event_user_seconds_since_previous_revision,
  event_comment
FROM wmf.mediawiki_history
WHERE
  -- select the~right snapshot
      snapshot = '2023-03'
  AND wiki_db = 'cswiki'

  -- mediawiki_history also contains information unrelated
  -- to revision creation, such as promoting users or deleting pages.
  AND event_entity = 'revision'
  AND event_type = 'create'

  -- filter down to article edits submitted in~2022
  AND event_timestamp LIKE '2022-%'
  AND page_namespace_historical = 0

  -- ensure user is not autoconfirmed
  -- NOTE: autoconfirmed is an~implicit group, and as such,
  -- it is not included in~the~event_user_groups field. Instead,
  -- this requirement has to be checked using the~group definition.
  AND (
      -- user is logged out
          event_user_revision_count IS NULL

      -- group definition taken on 2023-04-19
      -- from https://cs.wikipedia.org/w/index.php?curid=332038
      OR (
              event_user_revision_count < 10
          AND (
            UNIX_TIMESTAMP(event_timestamp) -
            UNIX_TIMESTAMP(event_user_registration_timestamp)
          ) < 4 * 24 * 3600
      )
  )
```

Table 4.2 Summary of baseline vandalism classification – logged out edits

Variable	Value
Total edits	100
Vandalism edits	38
Reverted edits	39
Reverted and not vandalism	7

4.2.3 Review: Processing the data

Each technique described in Section 3.1 is simulated through a Python snippet using the *pandas* dataprocessing library for Python. The snippet determines the outcome of each technique and adds it as a new boolean column to the dataframe. The snippets may only make use of information the solutions would have on runtime (metadata of the evaluated edit and potentially, metadata of any historical edits[45]).

The following snippets are used for simulating each technique (all snippets are available in Appendix E):

Disallowing contributions by logged out users Code listing E.2

Presence of user IP addresses in blacklists Code listing E.3

Number of reverted edits Code listing E.4

Using the `get_scores` function defined in Code listing E.1 and the columns generated using the snippets described above, the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) are generated. Math formulas described in Section 3.2.2.1 are then used to calculate the F1 score, recall and FPR as instructed in the thesis assignment. Since the F1 score is calculated based on the recall and precision values, precision is calculated as well. Knowing the precision is benefitial to determine whether F1 score is low because of low recall or low precision (or both), which can be important to decide which solution performed the best.

4.2.3.1 Disallowing contributions by logged out users

Code listing E.2 was used to calculate the TP, FP, TN and FN (results are shown in Table 4.3). The code makes use of `event_user_revision_count`, which reports number of other edits made by the same user prior to saving this one[79]. The field is `NULL` if the user is logged out, because it is impossible to determine how many other edits a logged out user saved.

4.2.3.2 Presence of user IP addresses in blacklists

Code listing E.3 was used to calculate the TP, FP, TN and FN (results are shown in Table 4.4). The code to determine whether an IP is included in a list of networks is taken from [80], the author's prior work on reviewing the impact of Apple's *iCloud Private Relay* on countervandalism. Built-in capability of the `iptables` library to determine whether an IP is in a network is intentionally not used for performance reasons (as it avoids parsing the network CIDR specification more than once).

[45] Fetching historical data can be challenging in terms of implementation; this is neglected at this stage and accounted for in Section 4.4.

Table 4.3 Data acquired from running the *Disallowing contributions by logged out users* model

Variable	Value
Total edits	100
True positives	36
False positives	53
True negatives	5
False negatives	6
F1 score	54.9618 %
Precision	40.4494 %
Recall	85.7142 %
FPR	53 %

Table 4.4 Data acquired from running the *Presence of user IP addresses in blacklists* model

Variable	Value
Total edits	100
True positives	0
False positives	0
True negatives	62
False negatives	38
F1 score	0 %
Precision	N/A
Recall	0 %
FPR	0 %

4.2.3.3 Number of reverted edits

Code listing E.4 was used to calculate the TP, FP, TN and FN (results are shown in Table 4.5). Number of edits the user has reverted in the last 90 days is not directly available in the `mediawiki_history` dataset. However, the dataset contains edit's save timestamp (`event_timestamp`), the author's username (`event_user_text`) and a flag showing whether the edit was reverted (`revision_is_identity_reverted`). The *is reverted* flag only shows whether the edit was reverted in full (partial reverts are not detected), but full reverts are a reasonable approximation for the purpose of this solution. Using those data, number of recent reverted edits can be calculated.

Table 4.5 Data acquired from running the *Number of reverted edits* model

Variable	Value
Total edits	100
True positives	10
False positives	3
True negatives	55
False negatives	32
F1 score	36.3636 %
Precision	76.9231 %
Recall	23.8095 %
FPR	3 %

4.3 Evaluation

Looking at gathered data about individual models (Table 4.3, Table 4.4 and Table 4.5), the following can be concluded:

4.3.1 Disallowing contributions by logged out users

This solution has a high recall value of 85 %, meaning it can detect a large portion of vandalism edits. However, this comes at the cost of low precision (40 %), meaning the model often labels a constructive edit as vandalism. As the goal is to detect vandalism without affecting constructive editors, Consequently, while the model would help patrollers to prevent a lot of vandalism, it would make contributing harder for several constructively-editing users.

4.3.2 Presence of user IP address in blacklist

In the reviewed sample, none of the IP addresses was present in StopForumSpam-maintained blacklist, resulting in zero precision. This means blacklists appear to be ineffective against regular vandalism, which is not likely to come from botnets.

4.3.3 Number of reverted edits

This solution has a fairly high precision (77 %), while maintaining a reasonably low recall value (24 %). While this is not ideal (and means a significant portion of actual vandalism is left undetected), it is more acceptable than high recall/low precision scenario, as it means low number of constructive edits are prevented.

From the three evaluated solutions, *Number of reverted edits* performs the best.

4.4 Discussion

The two solutions that caught a certain portion of the vandalism edits (*Disallowing contributions by logged out users* and *Number of reverted edits*) are different from each other: *Disallowing contributions by logged out users* has a high recall value (at the cost of low precision), while the other one has a high precision value (at the cost of low recall). This means a decision needs to be made whether it is more important to catch most vandalism edits (at the cost of being wrong many times) or to catch a smaller amount of vandalism (but being right most of the time).

Since the focus of the thesis was to help patrollers by detecting vandalism without human involvement, recommendation made by the thesis needs to be capable of running independently on humans. In other words, the outcome should not be an augmenting tool (defined in Section 2.4.3). To avoid having constructive edits blocked, a high precision value is preferred when a tradeoff needs to be made.

4.4.1 Disallowing contributions by logged out users

The thesis concluded that *Disallowing contributions by logged out users* should not be used, as while it is successful in detecting a significant amount of vandalism (recall value of 85.7 %), it mistakenly labels a lot of constructive edits by logged out users as vandalism. This is in line with earlier findings by Javanmardi *et al.*, who evaluated contributions by registered and unregistered editors in 2009 using the *Wiki Trust Model*. In [81], they stated that while registered editors on average contribute higher quality content than logged out editors, a significant number of unregistered users also contribute high-quality content.

Those conclusions are supported by Anthony *et al.* [82], who states that while registered contributors (motivated by their reputation and committment to the Wikipedia community) make many contributions with high reliability, the highest reliability comes from the vast members of anonymous *Good Samaritans*, who contribute only once. In addition to what was mentioned before, Yochai Benkler credited Wikipedia for its low transaction costs[46] and cited *editing without the need to create an account* as one of the factors contributing to Wikipedia's success[83].

Despite several researchers agreed that edits by logged out contributors are, at the very least, beneficial to Wikipedia, the community of Portugalese Wikipedia decided in 2020 to ban all logged out editors[68]. Before that ban, all research of logged out Wikipedia editing was purely theoretical, as there was no treatment group that could be used to confirm/deny the value of contributions by unregistered users. The WMF researched the impact of that decision and conducted that "we found no significant negative impact in the analysis conducted thus far"[84], while "The number of reverts, page protections and blocks have declined considerably, indicating a decrease in the amount of vandalism on the project."[84].

This is contradictory to the frequently-cited value of unregistered contributors. The discrepancy can be caused (among other factors) by change of human behavior, based on whether Wikipedia allows unregistered users to directly edit. In other words, a certain subset of *soon-to-be* unregistered contributors could decide to sign up instead, while another subset could decide to not edit. Depending on which group of such editors is more motivated (constructive editors or vandals), the amount of vandalism can increase or decrease. Despite what was previously said, the WMF's initial research is only based on three quarters which followed Portugalese Wikipedia's decision to disallow logged out editors and as such, the long-term impact of the decision might not be visible yet.

4.4.2 Presence of user IP addresses in blacklist

Table 4.4 shows there were no edits in the random sample of 100 logged out edits with their author present on *StopForumSpam*-managed blacklists. By itself, this information cannot be used to draw any conclusions about the effectiveness of blacklists in countervandalism. Observed outcome can have many causes: for example, if users are rarely blacklisted, but if they are, their edits are nearly always vandalism, making use of blacklists could still be an useful technique to implement, even if no such case made its way into the random sample of 100 edits. On the other hand, if Wikipedia editors almost never make their way to the blacklist, this solution could be indeed very ineffective and not worth implementing.

To determine which of the two described cases happened here, an alternate evaluation method was prepared. Instead of generating a random sample of 100 edits, classifying them by humans and then using the spam blacklist, it is possible to get a list of 20 live Wikipedia edits from 2022, with their author being on the *StopForumSpam* blacklist. Those edits can be then manually

[46] Difficultness of making a particular transaction, such as saving a Wikipedia edit.

Code listing 4.2 Getting a list of IP addresses in the SFS blacklist

```
import pandas as pd

import ipaddress
from collections import defaultdict

sfs_nets = {
  '4': defaultdict(set),
  '6': defaultdict(set)
}

for ip in~sfsDf.IP:
  net = ipaddress.ip_network(ip)
  net_v = str(net.version)
  sfs_nets[net_v][net.netmask].add(int(net.network_address))

def is_ip_sfs_listed(ip_raw):
  try:
      ip = ipaddress.ip_address(ip_raw)
      bin_ip = int(ip)

      for netmask, range_set in~sfs_nets[str(ip.version)].items():
          bin_netmask = int(netmask)
          if (bin_ip & bin_netmask) in~range_set:
              return True
  except ValueError:
      pass

  return False

  df = pd.read_csv('bachelor-thesis-revisions.tsv', sep='\t')

  dfAnon = df.loc[df.event_user_revision_count.isnull()]
  dfAnon['sfs_listed'] = dfAnon.event_user_text.apply(
    lambda x: is_ip_sfs_listed(x)
  )

  print(
    'Total blacklisted: %d' %
    dfAnon.loc[dfAnon.sfs_listed].event_timestamp.count()
  )
  print(
    dfAnon.loc[dfAnon.sfs_listed == True]\
    .sample(20)[['revision_id', 'event_user_text', 'page_title']]
  )
```

classified and used to better gauge the viability of the blacklist solution proposed in the thesis. This idea is implemented as Code listing 4.2.

After running, the snippet found a total of 59 revisions with their author's IP address present on the *StopForumSpam* blacklist. Similar to the original approach, only edits by logged out authors were considered (IP address for logged in users is not available publicly for privacy reasons).

Table 4.6 Number of vandalism edits made by blacklisted editors in 2022 (sampled)

Label	Edits	%
Vandalism	9	45%
Constructive	11	55%

The generated sample was manually classified similar to the process described in Section 4.2.2; the sample and the classifications are included in Appendix C. After evaluating the newly acquired data, results depicted in Table 4.6 can be observed. While a singificant amount of the edits were classified as vandalism, the majority of edits were constructive, which indicates usage of *StopForumSpam* might not be a right approach.

There are a couple of possible explanations of the observed outcome. First of all, the chosen blacklist (*StopForumSpam*) might not be the right source of information for the purpose of countervandalism. A different blacklist would likely bring different results.

In addition to that, the Czech Wikipedia (where the experiment was carried out) is a medium-size Wikipedia with a established group of patrollers, administrators and other functionaries. Vandals might consider community size in their considerations (especially organized vandals), as edits posted on a very small wikis generally have a higher chance of survival. Running the same experiment on a smaller wiki might give different results.

4.4.3 Number of reverted edits

While the *Number of reverted edits* technique appears to perform the best, its success is directly tied to the value of its two thresholds: maximum tolerated amount of reverted edits and considered period of time. The experiment results could be more convincing if the approach was tested with multiple values as the thresholds, potentially increasing the solution's effectivity.

Another approach that might be tried in addition to what was described in the thesis is using *relative* thresholds instead of *absolute* thresholds. In other words, instead of preventing all users with more than 5 reverted edits in the last 90 days, prevent all users with i. e. more than 5 % reverted edits in the last 90 days. That way, the technique would adapt to the user's editing behavior. Users who make a significant amount of edits naturally have a higher absolute number of reverted edits, while they might have a lower relative number of reverted edits.

4.4.4 Recommendations

Out of the three evaluated solutions, the *Number of reverted edits* appears to perform the best. To make it possible to see its benefits in the real world, it should be deployed as a trial and evaluated. This can be done through adding a new variable to the *AbuseFilter* extension that can be used by filter authors, which would contain the number of past reverted edits within a given amount of days. When that happens, the administrators are able to create edit filters based on this technique and evaluate their performance based on logs made by the edit filters.

Chapter 5

Conclusion

First and foremost, the thesis aimed to decrease the amount of time humans need to invest into countervandalism efforts by (at least partially) automating countervandalism. The author suggested several countervandalism techniques and evaluated each of them, with the ultimate goal of suggesting one countervandalism technique for implementation and/or further research.

Other authors decided to approach automating countervandalism slightly differently: their solutions (like ORES or ClueBot NG) take revision content into account. While those solutions (especially ClueBot NG) are helpful in some projects, they are difficult to scale across Wikimedia projects. Scaling is a significant problem because Wikipedia covers 300+ languages (more than e. g. Google Translate).

For those reasons, a decision was made to suggest countervandalism techniques that only depend on revision metadata, which is equally available across all Wikimedia projects, resulting in easier scaling. In addition to easing the scaling process, working with revision metadata is also significantly easier than working with revision content, which means the solutions suggested by the thesis are significantly easier to implement engineering-wise.

Namely, the thesis goes through four different countervandalism techniques: requiring wikipedians to identify to the WMF, prohibiting logged out contributors from participating, checking IP address blacklists prior to allowing an edit and disallowing contributions from authors with significant amount of reverted edits. The first out of those solutions was deemed unsuitable as being contary to Wikipedia's principles, while the others were submitted to the viability evaluation.

The viability review showed that prohibing logged out contributions prevents many constructive edits due to its high false-positive ratio. However this conclusion is not definitive, as user behavior may change in the case this solution is implemented. Since account creation does not come with significant costs, users currently editing as logged out might submit their contributions regardless, using their account. Similar user behavior changes are impossible to either prove or disprove via the metodology used by this thesis.

Furthermore, the viability reviewed proved that taking user's past history into account has some potential. Further research needs to be done to determine the most impactful thresholds. A recommendation was made to include number of user's recent reverts in the AbuseFilter extension and to pilot this solution at a willing Wikimedia project.

Thanks to its limited scope, the thesis did not review any of the suggested techniques in greater depth. However, its overall goal of serving as an initial exploration of metadata-only antivandalism solution was successful. Future research should be done on each of the proposed techniques, in order to identify spaces for improvement.

Appendix A

Sampled revisions: All

Below is a list of sampled revisions (including the manual classification) generated by means described in Section 4.2.1.

Meaning of columns is the following:

vandalism? Denotes whether the edit was classified as vandalism by a human Wikipedia administrator

event_timestamp Edit save timestamp

revision_id Internal revision ID, which can be used to see the revision by going to `https://cs.wikipedia.org/wiki/Special:Diff/revision_id`, replacing `revision_id` by the actual revision ID from the table below.

page_title Title of the article, which was affected by this edit.

vandalism?	event_timestamp	revision_id	page_title
TRUE	2022-03-17 9:48:46	21043705	Výr velký
FALSE	2022-09-01 0:39:20	21638102	Seznam fotbalistů s 500 a více vstřelenými brankami
FALSE	2022-07-30 20:24:36	21532766	Stanislav Lukeš
FALSE	2022-04-13 11:04:25	21158400	HC Hlinsko
FALSE	2022-07-16 9:33:18	21472942	Edvardovo jezero
TRUE	2022-11-28 16:34:53	22070378	Vánoční koleda (Dickens)
TRUE	2022-09-16 7:28:23	21682033	Manfred Kokot
TRUE	2022-04-28 15:58:27	21202270	Horní Bludovice
FALSE	2022-08-31 10:59:29	21636240	Reduta Jazz Club
TRUE	2022-02-17 19:54:47	20952021	Antiopresivní přístup v sociální práci
FALSE	2022-12-04 20:17:40	22178029	Litobratřice
TRUE	2022-02-24 19:01:19	20974506	Hermann Bahr
TRUE	2022-09-23 16:36:47	21705053	Psychopatie
FALSE	2022-04-19 8:23:29	21175007	Prvočíslo
FALSE	2022-03-12 17:05:32	21028415	Jiří Rajmund Tretera
FALSE	2022-08-20 20:23:22	21606350	Krevní paraziti ptáků
FALSE	2022-09-27 15:29:29	21716280	Miloslava Hrdličková-Šrámková

FALSE	2022-08-01 12:26:13	21536911	Palladium (multifunkční komplex)
TRUE	2022-01-05 19:59:48	20792995	Filip Grznár
TRUE	2022-02-01 8:00:58	20893631	30. září
FALSE	2022-01-31 8:35:11	20890960	Steven F. Udvar-Hazy Center
FALSE	2022-07-10 11:17:18	21458300	Uhlíkové clo
TRUE	2022-08-18 8:51:55	21598565	Lumen
TRUE	2022-12-15 7:22:51	22223507	Odysseia
TRUE	2022-07-09 8:26:40	21451724	Útok žraloka bílého v Opavě
TRUE	2022-03-10 14:20:14	21021067	Husitské války
FALSE	2022-05-30 9:44:41	21341359	PZL P.11
FALSE	2022-12-30 0:41:17	22278020	Olomouc
FALSE	2022-03-28 18:16:38	21084619	Estrelský pastevecký pes
FALSE	2022-05-14 9:55:30	21269806	Volby do České národní rady 1990
FALSE	2022-10-03 19:24:36	21734431	Richard Genzer
FALSE	2022-10-03 17:50:25	21734081	Ladislav Zbořil
FALSE	2022-02-13 12:11:37	20934693	Know Nothing
TRUE	2022-10-13 19:17:44	21763856	Jiří Vyvadil
FALSE	2022-12-09 9:13:44	22195815	Jeanna Gieseová
FALSE	2022-08-19 19:22:56	21603703	Polikarpov I-17
FALSE	2022-04-14 20:43:38	21161709	Zlín Trenér
TRUE	2022-10-31 12:16:18	21817252	TikTok
FALSE	2022-11-08 15:15:35	21858256	Zdeněk Julina
FALSE	2022-03-20 12:52:26	21054979	Myers-Briggs Type Indicator
TRUE	2022-12-19 13:19:56	22238941	Hypersexualita
TRUE	2022-03-03 13:18:34	20994976	Volodymyr Zelenskyj
FALSE	2022-07-19 9:22:40	21479164	Martin Tešovič
TRUE	2022-06-13 10:52:12	21381500	Horkýže Slíže
FALSE	2022-09-21 11:36:34	21697629	Mark Regev
TRUE	2022-01-14 6:25:31	20829425	Nové Město na Moravě
TRUE	2022-01-07 21:24:37	20801454	Katharine Hepburnová
FALSE	2022-03-10 10:15:43	21019901	CZ 805 BREN
FALSE	2022-05-23 15:31:41	21318352	Lady Gaga
FALSE	2022-05-24 12:17:11	21320890	Mistrovství světa v ledním hokeji 2022
FALSE	2022-06-10 15:37:41	21374533	Breguet 521
FALSE	2022-11-18 12:32:32	21905507	Dějiny Nizozemska
TRUE	2022-11-08 10:34:24	21856707	Počítač
TRUE	2022-11-27 12:45:15	22028271	Terej modronohý
FALSE	2022-05-08 22:06:50	21249512	Formule 1 v roce 2022
TRUE	2022-05-10 15:59:14	21256013	Křižák obecný
FALSE	2022-10-02 23:51:42	21731795	Mongolský vpád na Rus
FALSE	2022-06-06 6:11:02	21360460	Volby do Poslanecké sněmovny Parlamentu České republiky 2002
TRUE	2022-01-14 8:23:20	20829700	Sparta
FALSE	2022-12-28 20:16:29	22272611	Smečno
FALSE	2022-05-23 17:31:13	21318679	24. květen
TRUE	2022-11-23 11:35:32	21935421	Kutná Hora

TRUE	2022-12-13 18:14:07	22219620	JRB Rock
FALSE	2022-08-31 13:36:33	21636728	SteamOS
FALSE	2022-07-10 18:15:12	21459659	Wimbledon 2022 – ženská čtyřhra
FALSE	2022-02-18 11:08:24	20953737	Biatlon na Zimních olympijských hrách 2022
FALSE	2022-12-16 15:32:34	22229538	Jára Kohout
FALSE	2022-03-25 7:58:53	21074338	Marvel Cinematic Universe
FALSE	2022-06-16 18:08:10	21393178	Fakulta veřejných politik v Opavě Slezské univerzity v Opavě
FALSE	2022-10-06 9:27:59	21742567	Montrealské muzeum umění
TRUE	2022-02-01 10:29:41	20894112	Mozeček
FALSE	2022-03-13 8:39:46	21029673	Skládání rychlostí
TRUE	2022-11-21 22:04:57	21921522	Trutnov Open Air Festival
TRUE	2022-03-24 7:03:48	21071306	Isaac Zida
FALSE	2022-01-15 9:21:10	20833501	Boeing X-50 Dragonfly
TRUE	2022-11-09 9:33:45	21860400	Bělbog
FALSE	2022-09-27 10:04:39	21715457	Neratov (Bartošovice v Orlických horách)
FALSE	2022-08-12 12:40:11	21579816	Opioid
FALSE	2022-11-19 11:21:11	21908564	Ladislav Vrabel
TRUE	2022-06-02 12:04:59	21351026	Bory
TRUE	2022-09-16 8:21:55	21682119	Ivan Vyskočil
FALSE	2022-12-24 17:06:34	22259294	Třída Leahy
FALSE	2022-04-10 14:38:04	21150494	Ron Wyatt
TRUE	2022-04-15 19:40:37	21163842	Drahomíra Jůzová
TRUE	2022-03-28 6:14:50	21082306	Harry Maguire
FALSE	2022-01-26 10:58:39	20874127	Vortex Media
TRUE	2022-10-16 14:57:58	21772294	Jan Nepomucký
FALSE	2022-10-26 12:49:22	21803316	Michal Bureš
FALSE	2022-06-18 21:07:50	21399156	P/\st
FALSE	2022-03-07 21:36:50	21011591	Římskokatolická farnost Sudoměřice
TRUE	2022-10-10 12:05:51	21752961	Ozonová vrstva
FALSE	2022-01-10 7:48:52	20814502	Italská socialistická strana
TRUE	2022-11-09 13:23:04	21861381	Seznam dinosaurů
TRUE	2022-06-01 19:58:05	21349263	Like House (2. řada)
TRUE	2022-08-16 12:11:20	21592316	16. únor
FALSE	2022-05-17 12:16:20	21278162	Tělesné cvičení
TRUE	2022-10-17 15:15:43	21775110	Den vzniku samostatného československého státu
TRUE	2022-04-25 16:33:32	21193751	Kouření
TRUE	2022-06-07 10:08:41	21364458	Chňapal červený
FALSE	2022-10-19 16:42:33	21782108	American Motors

Appendix B

Sampled revisions: Anonymous only

Meaning of columns is the same as in Appendix A. This appendix contains the sample of anonymous-only edits.

vandalism?	event_timestamp	revision_id	page_title
FALSE	2022-05-06 5:39:32	21224817	Muzeum miniaturního profesionálního umění Henryk Jan Dominiak
TRUE	2022-04-11 11:24:42	21152744	Jan Svatopluk Presl
FALSE	2022-01-26 19:25:02	20875820	Jan Machytka
FALSE	2022-06-28 7:38:50	21421668	Dominik Hašek
FALSE	2022-08-13 11:59:22	21583136	Uri
FALSE	2022-06-22 10:16:54	21407000	Lovci přízraků (animovaný seriál)
FALSE	2022-04-04 20:14:59	21111177	Nikotin
FALSE	2022-03-08 17:55:02	21014540	Jan Bechyna
TRUE	2022-07-25 12:07:51	21516993	Martin Rajnis
FALSE	2022-11-06 12:28:55	21842539	Josef Šilhavý
FALSE	2022-01-21 15:27:50	20858964	Tomáš
FALSE	2022-12-30 19:10:53	22280778	Eva Burešová
FALSE	2022-05-27 8:29:58	21328926	PZL.23 Karaś
FALSE	2022-06-29 17:54:13	21425865	Žraloci
FALSE	2022-09-06 6:57:24	21650915	Luboš Dörfl
FALSE	2022-07-05 23:54:48	21442544	Hudson a Rex
FALSE	2022-07-09 13:53:53	21453908	Karviná
FALSE	2022-03-26 23:22:42	21079463	Občanství
FALSE	2022-06-23 6:51:05	21409579	Nuselská
FALSE	2022-09-05 15:12:22	21649238	Blohm & Voss BV 138
FALSE	2022-07-06 16:15:49	21444006	Seznam největších měst v Evropě
TRUE	2022-11-20 13:23:07	21915586	Harry Potter a Tajemná komnata
TRUE	2022-04-07 13:23:40	21123648	Aztécká říše

FALSE	2022-08-30 20:11:07	21634917	Hedvábnička
FALSE	2022-01-20 20:40:35	20856898	Ernestinum (Příbram)
TRUE	2022-05-07 17:10:42	21237660	Yamaha
TRUE	2022-09-20 12:55:41	21694992	Chřipka
FALSE	2022-02-26 15:39:15	20979795	Eta Carinae
FALSE	2022-09-05 2:52:13	21647544	Lasse Virén
FALSE	2022-06-17 6:31:20	21394636	Seznam řek na Slovensku
FALSE	2022-10-24 4:57:12	21795414	Charles University Innovations Prague
TRUE	2022-12-11 16:13:11	22205155	Aquapalace Praha
TRUE	2022-02-11 15:39:42	20928187	Twice
TRUE	2022-03-18 17:07:31	21048687	Sudak
TRUE	2022-05-01 16:47:22	21210590	Bombový útok na Staroměstském náměstí 1990
TRUE	2022-12-30 12:37:46	22279626	Bigfoot
FALSE	2022-11-14 14:21:25	21882252	Chrastava
FALSE	2022-06-02 5:42:27	21349986	Volby do Poslanecké sněmovny Parlamentu České republiky 2002
FALSE	2022-09-07 18:52:00	21656463	Arcibiskupské gymnázium v Praze
TRUE	2022-02-16 7:29:43	20947132	Ester Ledecká
TRUE	2022-08-08 18:29:36	21568056	Jan Cemper
FALSE	2022-11-12 20:44:48	21876543	Česko Slovensko má talent (10. řada)
TRUE	2022-04-14 12:30:58	21160775	Daskabát
FALSE	2022-11-23 7:31:01	21928091	Kurtizány z 25. avenue
TRUE	2022-11-08 8:36:37	21854496	Brazílie
FALSE	2022-07-16 9:00:37	21472895	Václav Prchlík
FALSE	2022-10-03 5:12:09	21731964	Dani Filth
FALSE	2022-08-14 20:22:53	21586408	Seznam německých názvů obcí a osad v Česku, S
TRUE	2022-02-11 19:59:31	20929020	Filip Pešán
TRUE	2022-11-03 13:38:53	21830607	ČT art
FALSE	2022-11-09 13:17:47	21861366	Michael Möllenbeck
FALSE	2022-10-21 13:19:19	21787698	Kouzelná Beruška a Černý kocour (5. řada)
FALSE	2022-01-23 8:48:52	20863850	Luděk Nekuda
FALSE	2022-03-27 10:18:13	21080148	Gonggong (planetka)
TRUE	2022-11-26 17:33:52	22025560	Igor Matovič
FALSE	2022-01-20 13:01:44	20855175	Modelová železnice
TRUE	2022-01-28 16:04:18	20882435	Harry Styles
FALSE	2022-03-06 8:55:08	21004387	Varuna (planetka)
FALSE	2022-04-04 15:09:50	21107433	CoBrA
TRUE	2022-05-29 16:39:24	21339325	Patrik Le Giang
TRUE	2022-02-03 10:08:44	20900069	Marek Židlický
FALSE	2022-06-27 10:52:27	21419720	Marek Hrbas
FALSE	2022-08-14 14:44:27	21585668	Jaroslav Špillar
TRUE	2022-09-25 18:37:51	21711524	Ján Jesenský
TRUE	2022-02-18 9:27:50	20953352	Rorýs obecný

TRUE	2022-08-01 15:55:33	21537690	Fiat 500 (2007)
FALSE	2022-04-08 10:43:35	21126053	Staroměstská radnice
TRUE	2022-01-11 13:17:17	20819383	Sergei Barracuda
TRUE	2022-11-04 12:40:12	21836632	Khaled
TRUE	2022-05-11 6:53:04	21257174	Albert Einstein
TRUE	2022-05-19 13:52:52	21286881	Užovka obojková
FALSE	2022-06-16 10:42:26	21391543	Archa úmluvy
FALSE	2022-09-14 12:17:04	21676023	CEVRO Institut
TRUE	2022-08-29 10:53:27	21629709	Jiří Míšenský
FALSE	2022-01-31 9:13:20	20891029	Sovy
FALSE	2022-04-23 22:19:54	21189065	Pavol Bajza (1991)
FALSE	2022-07-12 21:32:04	21465869	Pokémon
FALSE	2022-10-30 10:05:14	21813604	Kluky (okres Mladá Boleslav)
TRUE	2022-03-25 12:00:23	21075084	Zlatá horečka
FALSE	2022-06-23 12:17:04	21410323	Karel Řehka
FALSE	2022-04-12 7:49:09	21155034	Světové skautské jamboree
FALSE	2022-03-31 21:10:52	21096229	Discovery (společnost)
FALSE	2022-11-06 14:13:32	21844822	Takeoff (rapper)
TRUE	2022-12-05 9:43:51	22180041	Karel Hynek Mácha
TRUE	2022-03-21 9:57:39	21058359	Želvy
FALSE	2022-06-20 10:50:26	21402315	Mák vlčí
FALSE	2022-12-30 20:57:31	22281140	Tramvajová doprava v Ostravě
FALSE	2022-10-23 11:09:25	21792789	Parmezán
TRUE	2022-03-03 17:04:19	20995624	Sapfó
TRUE	2022-03-06 12:32:01	21005061	Nektarinka
FALSE	2022-04-22 8:36:33	21182859	Rašismus
FALSE	2022-09-05 15:11:48	21649237	Blohm & Voss BV 138
FALSE	2022-01-03 13:39:05	20785267	Velká turecká válka
FALSE	2022-11-13 16:34:04	21879098	Věra Jordánová
FALSE	2022-05-25 10:53:46	21323657	Zara Phillips
TRUE	2022-01-03 17:44:19	20786020	Číňané
TRUE	2022-11-20 17:54:24	21916334	Ekvádor na Letních olympijských hrách 2008
FALSE	2022-12-31 9:47:13	22282219	Benedikt XVI.
TRUE	2022-01-24 8:20:11	20866984	Začít spolu
TRUE	2022-05-31 17:14:00	21345827	Hacker

Appendix C

Sampled revisions: Users with their IP on a blacklist

Meaning of columns is the same as in Appendix A; the additional column, `event_user_text`, contains the IP address of the revision author. This appendix contains the sample of edits made by logged-out editors with their IP address present on the *StopForumSpam* blacklist.

vandalism?	revision_id	event_user_text	page_title
FALSE	22201815	109.105.39.6	Antinatalismus
TRUE	21901427	109.105.39.6	Skotsko
FALSE	21633403	185.21.222.174	Gymnázium U Libeňského zámku
FALSE	21875764	109.105.39.6	Česko Slovensko má talent (10. řada)
FALSE	21839171	109.105.39.6	Max Verstappen
TRUE	21694790	109.105.39.6	Dolní Kounice
TRUE	21744770	212.79.110.139	Kyrgyzstán
TRUE	20953737	46.229.124.13	Biatlon na Zimních olympijských hrách 2022
TRUE	22194354	109.105.39.6	Okurka setá
FALSE	21904165	144.48.38.35	Seznam dílů seriálu Griffinovi
TRUE	21054385	188.75.191.142	Spotřebitel
FALSE	22222366	109.105.39.6	George Ritzer
TRUE	20887421	37.29.88.72	Tsunami
FALSE	21711280	109.105.39.6	Jan Žaloudík
FALSE	20987341	212.79.110.139	Dýšina
TRUE	20833356	110.74.220.100	Maxime Dethomas
FALSE	21037296	82.99.189.81	Nový Bydžov
FALSE	21461706	31.173.87.197	Pendolino
FALSE	21711281	109.105.39.6	Jan Žaloudík
TRUE	21477975	149.34.244.172	Matěj Kubíček

Appendix D

Code listings: Processing data

The appendix contains snippets used during processing the data as described in Section 4.2.3.

Code listing D.1 Sampling generated list of revisions

```
import pandas as pd

# in this snippet, df includes the list generated by the SQL query,
# loaded as a Pandas dataframe

# store all edits into a TSV file
df.to_csv('bachelor-thesis-revisions.tsv', sep='\t', index=False)

# store a sampled list of all edits
df.sample(100).to_csv(
  'bachelor-thesis-revisions-SAMPLE.tsv',
  sep='\t',
  index=False
)


# store a sampled list of logged-out edits
# logged in status is determined via event_user_revision_count,
# which is null for logged out users.
df.loc[pd.isnull(df.event_user_revision_count)]\
  .sample(100)\
  .to_csv(
    'bachelor-thesis-revisions-SAMPLE-anon-only.tsv',
    sep='\t',
    index=False
  )
```

Appendix E

Code listings: Simulating countervandalism techniques

The appendix contains snippets used during simulating individual countervandalism techniques as described in Section 4.2.3.

Code listing E.1 Python function `get_scores` generating TP, TN, FP and FN to analyze each technique

```
import pandas as pd

# in this snippet, dfSampled is the sampled list (generated
# as described above), loaded as a Pandas dataframe

def get_scores(model_column, anon_only=False):
  if not anon_only:
    df = dfSampled
  else:
    df = dfSampledAnon
  print('TP: {tp}\nFP: {fp}\nTN: {tn}\nFN: {fn}'.format(
      tp=df.apply(
        lambda x: (x['vandalism?'] and x[model_column]),
        axis=1
      ).sum(),
      tn=df.apply(
        lambda x: ((not x['vandalism?']) and (not x[model_column])),
        axis=1
      ).sum(),
      fp=df.apply(
        lambda x: ((not x['vandalism?']) and x[model_column]),
        axis=1
      ).sum(),
      fn=df.apply(
        lambda x: (x['vandalism?'] and (not x[model_column])),
        axis=1
      ).sum()
  ))
```

Code listing E.2 Running the *Disallowing logged out contributors* model

```
import pandas as pd

# in this snippet, dfSampled is the sampled list (generated
# as described above), loaded as a Pandas dataframe

# `get_scores(model_column)` is defined in Code listing D.1.

# run the model: determine when users were not logged in
dfSampled['logged_out'] = dfSampled\
    .event_user_revision_count\
    .apply(lambda x: pd.isnull(x))

get_scores('logged_out')
```

Code listing E.3 Running the *Presence of user IP addresses in blacklists* model

```
    import pandas as pd

    import ipaddress
    from collections import defaultdict

    # in this snippet, dfSampledAnon is the sampled list (generated
    # as described above), loaded as a Pandas dataframe

    # `get_scores(model_column)` is defined in Code listing D.1.

    sfs_nets = {
        '4': defaultdict(set),
        '6': defaultdict(set)
    }

    for ip in sfsDf.IP:
        net = ipaddress.ip_network(ip)
        net_v = str(net.version)
        sfs_nets[net_v][net.netmask].add(int(net.network_address))

    def is_ip_sfs_listed(ip_raw):
        try:
            ip = ipaddress.ip_address(ip_raw)
            bin_ip = int(ip)

            for netmask, range_set in sfs_nets[str(ip.version)].items():
                bin_netmask = int(netmask)
                if (bin_ip & bin_netmask) in range_set:
                    return True
        except ValueError:
            pass

        return False

    dfSampledAnon['sfs_listed'] = dfSampledAnon.event_user_text
      \.apply(lambda x: is_ip_sfs_listed(x))

    get_scores('sfs_listed', True)
```

Code listing E.4 Running the *Number of reverted edits* model

```
import pandas as pd

# in this snippet, dfSampled is the sampled list (generated
# as described above) and df includes data about all revisions
# from 2022 (the studied year) or 2021; both are loaded as
# Pandas dataframes.
# `get_scores(model_column)` is defined in Code listing D.1.

DAYS_TO_CONSIDER = 90
MIN_REVERTS_TO_PASS = 5

dfReverted = df.loc[df.revision_is_identity_reverted]

def get_reverted_edits_before(row):
  return dfReverted.loc[
      (dfReverted.event_user_text == row.event_user_text) &
      (
        dfReverted.event_timestamp.apply(
          lambda x: (row.event_timestamp - x).days
        ) < DAYS_TO_CONSIDER
      )
  ].revision_id.count()

dfSampled['has_many_reverted_edits'] = dfSampled.apply(
  lambda x: get_reverted_edits_before(x) > MIN_REVERTS_TO_PASS,
  axis=1
)
```

Bibliography

1. WIKIPEDIA CONTRIBUTORS. *Help:Reverting — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Help:Reverting&oldid=1137545917. [Online; accessed 9-April-2023].
2. META-WIKI CONTRIBUTORS. *Wikimedia projects — Meta, discussion about Wikimedia projects*. 2022. Available also from: https://meta.wikimedia.org/w/index.php?title=Wikimedia_projects&oldid=24068610. [Online; accessed 9-April-2023].
3. AYERS, Phoebe; MATTHEWS, Charles; YATES, Ben. *How Wikipedia works: and how you can be a part of it*. San Francisco: No Starch Press, 2008. Available also from: https://ebookcentral.proquest.com/lib/natl-ebooks/detail.action?docID=1137543.
4. REAGLE, Joseph. Wikipedia: The Happy Accident. *Interactions*. 2009, vol. 16, no. 3, pp. 42–45. ISSN 1072-5520. Available from DOI: 10.1145/1516016.1516026.
5. JEMIELNIAK, Dariusz. *Common knowledge?: an ethnography of Wikipedia*. Stanford, California: Stanford University Press, 2014. Available also from: http://ebookcentral.proquest.com/lib/natl-ebooks/detail.action?docID=1680678.
6. SANGER, Larry. *Let's Make a Wiki* [Mailing list]. 2001. Available also from: http://web.archive.org/web/20030414014355/http://www.nupedia.com/pipermail/nupedia-l/2001-January/000676.html. [Online; accessed 25-Mar-2023].
7. MEYER, Susan. *Jimmy Wales and Wikipedia*. Rosen Publishing Group, 2012. Internet Biographies (Rosen).
8. NEATE, Rupert. *Wikipedia founder Jimmy Wales goes bananas* [https://web.archive.org/web/20081110041546/http://www.telegraph.co.uk/finance/newsbysector/mediatechnologyandtelecoms/3399843/Wikipedia-founder-Jimmy-Wales-goes-bananas.html]. 2008. [Online; accessed 28-Apr-2023].
9. META-WIKI CONTRIBUTORS. *Superprotect — Meta, discussion about Wikimedia projects*. 2022. Available also from: https://meta.wikimedia.org/w/index.php?title=Superprotect&oldid=23889295. [Online; accessed 8-April-2023].
10. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Five pillars — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Five_pillars&oldid=1142829530. [Online; accessed 9-April-2023].
11. META-WIKI CONTRIBUTORS. *Research:Data — Meta, discussion about Wikimedia projects*. 2023. Available also from: https://meta.wikimedia.org/w/index.php?title=Research:Data&oldid=24654466. [Online; accessed 9-April-2023].
12. META-WIKI CONTRIBUTORS. *Research:Quarry — Meta, discussion about Wikimedia projects*. 2022. Available also from: https://meta.wikimedia.org/w/index.php?title=Research:Quarry&oldid=23911488. [Online; accessed 9-April-2023].

13. WIKIMEDIA FOUNDATION. *Wiki Replicas — Wikitech,* 2022. Available also from: https://wikitech.wikimedia.org/w/index.php?title=Wiki_Replicas&oldid=2013204. [Online; accessed 9-April-2023].
14. WIKIPEDIA CONTRIBUTORS. *Wikipedie:Pod lípou/Archiv 2015/05 — Wikipedie: Otevřená encyklopedie*. 2019. Available also from: https://cs.wikipedia.org/w/index.php?title=Wikipedie:Pod_l%C3%ADpou/Archiv_2015/05&oldid=16825985. [Online; accessed 9-April-2023].
15. WIKIMEDIA FOUNDATION. *Help:Toolforge/Database — Wikitech,* 2023. Available also from: https://wikitech.wikimedia.org/w/index.php?title=Help:Toolforge/Database&oldid=2067398. [Online; accessed 9-April-2023].
16. MEDIAWIKI CONTRIBUTORS. *API — MediaWiki,* 2023. Available also from: https://www.mediawiki.org/w/index.php?title=API&oldid=5858783. [Online; accessed 10-May-2023].
17. WIKIMEDIA FOUNDATION. *Analytics Datasets: MediaWiki History*. 2023. Available also from: https://dumps.wikimedia.org/other/mediawiki_history/readme.html. [Online; accessed 24-April-2023].
18. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Vandalism — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Vandalism&oldid=1139736888. [Online; accessed 13-March-2023].
19. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Assume good faith — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Assume_good_faith&oldid=1143744519. [Online; accessed 14-March-2023].
20. WIKIPEDIA CONTRIBUTORS. *Help:Template — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Help:Template&oldid=1137160383. [Online; accessed 13-March-2023].
21. META-WIKI CONTRIBUTORS. *Vandalbot — Meta, discussion about Wikimedia projects*. 2022. Available also from: https://meta.wikimedia.org/w/index.php?title=Vandalbot&oldid=23351344. [Online; accessed 2-April-2023].
22. PRIEDHORSKY, Reid; CHEN, Jilin; LAM, Shyong K.; PANCIERA, Katherine; TERVEEN, Loren; RIEDL, John. Creating, Destroying, and Restoring Value in Wikipedia. In: *GROUP '07: Proc. of the 2007 ACM Conference on Supporting Group Work*. 2007, pp. 259–268.
23. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Rollback — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Rollback&oldid=1153805061. [Online; accessed 10-May-2023].
24. MORRIS, Kevin. *After a half-decade, massive Wikipedia hoax finally exposed* [https://www.dailydot.com/unclick/wikipedia-bicholim-conflict-hoax-deleted/]. 2013. [Online; accessed 25-Mar-2023].
25. NEWMAN, Jared. *Fake Wikipedia entry on Bicholim Conflict finally deleted after five years* [https://www.pcworld.com/article/456243/fake-wikipedia-entry-on-bicholim-conflict-finally-deleted-after-five-years.html]. 2013. [Online; accessed 25-Mar-2023].
26. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Articles for deletion/Bicholim conflict — Wikipedia, The Free Encyclopedia*. 2022. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Articles_for_deletion/Bicholim_conflict&oldid=1086371491. [Online; accessed 25-March-2023].
27. BLAHUŠ, Marek. *Stinné stránky Wikipedie* [https://slideslive.com/38892569/stinne-stranky-wikipedie]. 2014. [Online; accessed 25-Mar-2023].
28. WIKIPEDIA CONTRIBUTORS. *Wikipedia:List of hoaxes on Wikipedia — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:List_of_hoaxes_on_Wikipedia&oldid=1145750543. [Online; accessed 25-March-2023].

29. COLLINS, Katie. *Wikipedia temporarily defaced by vandal who edited template to show Nazi flag — cnet.com* [https://www.cnet.com/culture/internet/wikipedia-temporarily-defaced-by-vandal-who-edited-template-to-show-nazi-flag/]. 2021. [Online; accessed 23-Mar-2023].
30. WODINSKY, Shoshana. *Thousands of Wikipedia Pages Vandalized With Giant Swastikas* [https://gizmodo.com/thousands-of-wikipedia-pages-vandalized-with-giant-swas-1847494288]. 2021. [Online; accessed 25-Mar-2023].
31. FIŠER, Miloslav. *Tisíce stránek na Wikipedii zakryla svastika - Novinky — novinky.cz* [https://www.novinky.cz/clanek/internet-a-pc-tisice-stranek-na-wikipedii-zakryla-svastika-40369385]. 2021. [Online; accessed 23-Mar-2023].
32. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Administrators — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Administrators&oldid=1142465330. [Online; accessed 3-April-2023].
33. WIKIPEDIA CONTRIBUTORS. *Wikipedia:User access levels — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:User_access_levels&oldid=1142719115. [Online; accessed 25-March-2023].
34. WIKIPEDIA CONTRIBUTORS. *Wikipedie:Schválení uživatelé — Wikipedie: Otevřená encyklopedie*. 2022. Available also from: https://cs.wikipedia.org/w/index.php?title=Wikipedie:Schv%C3%A1len%C3%AD_u%C5%BEivatel%C3%A9&oldid=22204468. [Online; accessed 10-May-2023].
35. "ALEXSH". *Change autoconfirmed days limit in Chinese Wikipedia*. 2008. Available also from: https://phabricator.wikimedia.org/T16624. [Online; accessed 10-May-2023].
36. WIKIPEDIA CONTRIBUTORS. *Wikipedie:Patroláři a revertéři — Wikipedie: Otevřená encyklopedie*. 2022. Available also from: https://cs.wikipedia.org/w/index.php?title=Wikipedie:Patrol%C3%A1%C5%99i_a_revert%C3%A9%C5%99i&oldid=21627728. [Online; accessed 10-May-2023].
37. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Arbitration Committee — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Arbitration_Committee&oldid=1140776610. [Online; accessed 10-May-2023].
38. WIKIMEDIA FOUNDATION. *Policy:Terms of Use — Wikimedia Foundation Governance Wiki,* 2023. Available also from: https://foundation.wikimedia.org/w/index.php?title=Policy:Terms_of_Use&oldid=189797. [Online; accessed 3-April-2023].
39. META-WIKI CONTRIBUTORS. *Ombuds commission — Meta, discussion about Wikimedia projects*. 2023. Available also from: https://meta.wikimedia.org/w/index.php?title=Ombuds_commission&oldid=24909944. [Online; accessed 10-May-2023].
40. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Blocking policy — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Blocking_policy&oldid=1146282262. [Online; accessed 9-April-2023].
41. META-WIKI CONTRIBUTORS. *Data retention guidelines — Meta, discussion about Wikimedia projects*. 2023. Available also from: https://meta.wikimedia.org/w/index.php?title=Data_retention_guidelines&oldid=24724274. [Online; accessed 4-May-2023].
42. META-WIKI CONTRIBUTORS. *Oversight policy — Meta, discussion about Wikimedia projects*. 2022. Available also from: https://meta.wikimedia.org/w/index.php?title=Oversight_policy&oldid=24154485. [Online; accessed 3-April-2023].
43. META-WIKI CONTRIBUTORS. *Global sysops — Meta, discussion about Wikimedia projects*. 2023. Available also from: https://meta.wikimedia.org/w/index.php?title=Global_sysops&oldid=24446863. [Online; accessed 29-March-2023].
44. META-WIKI CONTRIBUTORS. *Stewards — Meta, discussion about Wikimedia projects*. 2023. Available also from: https://meta.wikimedia.org/w/index.php?title=Stewards&oldid=24501925. [Online; accessed 29-March-2023].

45. META-WIKI CONTRIBUTORS. *Steward requests/Permissions/2023-02 — Meta, discussion about Wikimedia projects*. 2023. Available also from: https://meta.wikimedia.org/w/index.php?title=Steward_requests/Permissions/2023-02&oldid=24929476. [Online; accessed 10-May-2023].
46. WIKIMEDIA FOUNDATION. *Policy:Office actions — Wikimedia Foundation Governance Wiki*, 2023. Available also from: https://foundation.wikimedia.org/w/index.php?title=Policy:Office_actions&oldid=189045. [Online; accessed 3-April-2023].
47. PŮTOVÁ, Barbora; ČERNÝ, Michal; PETIŠKA, Eduard; URBANEC, Martin; ČESKÁ TELEVIZE. *Wikipedie, dvacetiletý vševěd - Historie.cs | Česká televize — ceskatelevize.cz* [https://www.ceskatelevize.cz/porady/10150778447-historie-cs/223411058220003/]. 2023. [Accessed 10-May-2023].
48. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Huggle — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Huggle&oldid=1141614988. [Online; accessed 9-April-2023].
49. MEDIAWIKI CONTRIBUTORS. *Manual:Huggle — MediaWiki*. 2022. Available also from: https://www.mediawiki.org/w/index.php?title=Manual:Huggle&oldid=5408712. [Online; accessed 9-April-2023].
50. META-WIKI CONTRIBUTORS. *SWViewer — Meta, discussion about Wikimedia projects*. 2022. Available also from: https://meta.wikimedia.org/w/index.php?title=SWViewer&oldid=24022557. [Online; accessed 14-April-2023].
51. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Twinkle — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Twinkle&oldid=1148023323. [Online; accessed 14-April-2023].
52. WIKIPEDIA CONTRIBUTORS. *Help:Watchlist — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Help:Watchlist&oldid=1148008547. [Online; accessed 14-April-2023].
53. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Sanctions against editors should not be punitive — Wikipedia, The Free Encyclopedia*. 2022. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Sanctions_against_editors_should_not_be_punitive&oldid=1120079270. [Online; accessed 18-April-2023].
54. MEDIAWIKI CONTRIBUTORS. *Autoblock — MediaWiki*. 2021. Available also from: https://www.mediawiki.org/w/index.php?title=Autoblock&oldid=4567709. [Online; accessed 9-April-2023].
55. META-WIKI CONTRIBUTORS. *Global blocks — Meta, discussion about Wikimedia projects*. 2023. Available also from: https://meta.wikimedia.org/w/index.php?title=Global_blocks&oldid=24692146. [Online; accessed 9-April-2023].
56. META-WIKI CONTRIBUTORS. *No open proxies — Meta, discussion about Wikimedia projects*. 2023. Available also from: https://meta.wikimedia.org/w/index.php?title=No_open_proxies&oldid=24615717. [Online; accessed 18-April-2023].
57. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Protection policy — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Protection_policy&oldid=1141300568. [Online; accessed 23-March-2023].
58. META-WIKI CONTRIBUTORS. *Page protections considered harmful — Meta, discussion about Wikimedia projects*. 2023. Available also from: https://meta.wikimedia.org/w/index.php?title=Page_protections_considered_harmful&oldid=24822683. [Online; accessed 21-April-2023].
59. MEDIAWIKI CONTRIBUTORS. *Manual:Administrators — MediaWiki*, 2022. Available also from: https://www.mediawiki.org/w/index.php?title=Manual:Administrators&oldid=5619515. [Online; accessed 10-May-2023].
60. META-WIKI CONTRIBUTORS. *Flagged Revisions — Meta, discussion about Wikimedia projects*. 2023. Available also from: https://meta.wikimedia.org/w/index.php?title=Flagged_Revisions&oldid=24817345. [Online; accessed 7-May-2023].

61. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Edit filter — Wikipedia, The Free Encyclopedia*. 2023. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Edit_filter&oldid=1142063563. [Online; accessed 9-April-2023].
62. MEDIAWIKI CONTRIBUTORS. *Extension:AbuseFilter — MediaWiki,* 2023. Available also from: https://www.mediawiki.org/w/index.php?title=Extension:AbuseFilter&oldid=5894472. [Online; accessed 28-April-2023].
63. META-WIKI CONTRIBUTORS. *AbuseFilter — Meta, discussion about Wikimedia projects*. 2023. Available also from: https://meta.wikimedia.org/w/index.php?title=AbuseFilter&oldid=24348193. [Online; accessed 28-April-2023].
64. WIKIPEDIA CONTRIBUTORS. *User:ClueBot NG — Wikipedia, The Free Encyclopedia*. 2010. Available also from: https://en.wikipedia.org/w/index.php?title=User:ClueBot_NG&oldid=391868393. [Online; accessed 9-April-2023].
65. WIKIPEDIA CONTRIBUTORS. *User:ClueBot NG/FAQ — Wikipedia, The Free Encyclopedia*. 2019. Available also from: https://en.wikipedia.org/w/index.php?title=User:ClueBot_NG/FAQ&oldid=881154243. [Online; accessed 26-April-2023].
66. MEDIAWIKI CONTRIBUTORS. *ORES — MediaWiki*. 2022. Available also from: https://www.mediawiki.org/w/index.php?title=ORES&oldid=5262879. [Online; accessed 9-April-2023].
67. GEIGER, R. Stuart; HALFAKER, Aaron. When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes? In: *Proceedings of the 9th International Symposium on Open Collaboration*. New York, NY, USA: Association for Computing Machinery, 2013. WikiSym '13. ISBN 9781450318525. Available from DOI: 10.1145/2491055.2491061.
68. WIKIPÉDIA CONTRIBUTORS. *Wikipédia:Esplanada/propostas/Banimento de IPs (23ago2020) — Wikipédia, a enciclopédia livre*. 2022. Available also from: https://pt.wikipedia.org/w/index.php?title=Wikip%C3%A9dia:Esplanada/propostas/Banimento_de_IPs_(23ago2020)&oldid=63538016. [Online; accessed 7-May-2022].
69. HUJI. *Temporarily disable article editing by anonymous users on fawiki*. 2021. Available also from: https://phabricator.wikimedia.org/T291018. [Online; accessed 24-April-2023].
70. WIKIPEDIA CONTRIBUTORS. *Wikipedia:Autoconfirmed article creation trial — Wikipedia, The Free Encyclopedia*. 2022. Available also from: https://en.wikipedia.org/w/index.php?title=Wikipedia:Autoconfirmed_article_creation_trial&oldid=1102729748. [Online; accessed 23-April-2023].
71. STOPFORUMSPAM. *StopForumSpam downloads* [https://www.stopforumspam.com]. 2023. [Online; accessed 26-Apr-2023].
72. BASSETT, Scott. *Deploy StopForumSpam extension to production*. 2021. Available also from: https://phabricator.wikimedia.org/T273220. [Online; accessed 24-April-2023].
73. KORSTANJE, Joos. *The F1 score: all you need to know about the F1 score in machine learning* [https://towardsdatascience.com/the-f1-score-bec2bbc38aa6]. 2021. [Online; accessed 28-Apr-2023].
74. SONG, Victoria. *A Top Wikipedia editor has been arrested in Belarus* [https://www.theverge.com/2022/3/11/22973293/wikipedia-editor-russia-belarus-ukraine]. The Verge, 2022. [Online; accessed 25-Mar-2023].
75. BELANGER, Ashley. *Wikipedia admin jailed for 32 years after alleged Saudi spy infiltration* [https://arstechnica.com/tech-policy/2023/01/wikipedia-admin-jailed-for-32-years-after-alleged-saudi-spy-infiltration/]. Ars Technica, 2023. [Online; accessed 25-Mar-2023].
76. WIKIMEDIA FOUNDATION. *Policy:Privacy policy — Wikimedia Foundation Governance Wiki,* 2023. Available also from: https://foundation.wikimedia.org/w/index.php?title=Policy:Privacy_policy&oldid=188010. [Online; accessed 22-April-2023].
77. APACHE FOUNDATION. *Apache Spark™; – Unified Engine for large-scale data analytics* [https://spark.apache.org/]. 2023. [Online; accessed 22-Apr-2023].

78. WIKIMEDIA FOUNDATION. *Analytics/Data Lake/Edits/Mediawiki history dumps/Python spark examples — Wikitech,* 2020. Available also from: https://wikitech.wikimedia.org/w/index.php?title=Analytics/Data_Lake/Edits/Mediawiki_history_dumps/Python_spark_examples&oldid=1851853. [Online; accessed 21-April-2023].
79. WIKIMEDIA FOUNDATION. *Analytics/Data Lake/Edits/MediaWiki history — Wikitech,* 2022. Available also from: https://wikitech.wikimedia.org/w/index.php?title=Analytics/Data_Lake/Edits/MediaWiki_history&oldid=2013821. [Online; accessed 9-April-2023].
80. URBANEC, Martin. *iCloud Private Relay usage analysis*. 2021. Version 1.0.0. Available also from: https://github.com/urbanecm/2021-icloud-private-relay-usage. [Online; accessed 26-April-2023].
81. JAVANMARDI, Sara; GANJISAFFAR, Yasser; LOPES, Cristina; BALDI, Pierre. User contribution and trust in Wikipedia. In: *Proceedings of the 5th International ICST Conference on Collaborative Computing: Networking, Applications, Worksharing*. IEEE, 2009. Available from DOI: 10.4108/icst.collaboratecom2009.8376.
82. ANTHONY, Denise; SMITH, Sean W.; WILLIAMSON, Timothy. Reputation and Reliability in Collective Goods. *Rationality and Society*. 2009, vol. 21, no. 3, pp. 283–306. Available from DOI: 10.1177/1043463109336804.
83. BENKLER, Yochai. *The wealth of networks: how social production transforms markets and freedom*. New Haven: Yale University Press, 2006. ISBN 0300125771.
84. WIKIMEDIA FOUNDATION. *IP Editing: Privacy Enhancement and Abuse Mitigation/IP Editing Restriction Study/Portuguese Wikipedia — Meta, discussion about Wikimedia projects*. 2023. Available also from: https://meta.wikimedia.org/w/index.php?title=IP_Editing:_Privacy_Enhancement_and_Abuse_Mitigation/IP_Editing_Restriction_Study/Portuguese_Wikipedia&oldid=24597895. [Online; accessed 1-May-2023].

Contents of the attached files

- `thesis` ..... source code of the thexix in LaTeX
- `jupyter` ..... Jupyter notebooks used for the thesis
  - `sampler.ipynb` ..... Generates a sampled set of revisions
  - `analyze.ipynb` ..... Analyzes generated sampled set of revisions
  - `blacklist-02nd-attempt.ipynb` ..... Second attempt for the blacklist solution (see Section 4.4.2)
- `data` ..... Data about examined Czech Wikipedia revisions
  - `all` ..... Sample of all Czech Wikipedia edits (see Appendix A)
    - `revisions.tsv` ..... Information about the revisions
    - `classifications.tsv` ..... Vandalism classifications of the revisions
  - `logged-out` ..... Sample of logged-out Czech Wikipedia edits (see Appendix B)
    - `revisions.tsv` ..... Information about the revisions
    - `classifications.tsv` ..... Vandalism classifications of the revisions
  - `all-revisions.tsv` ..... Unsampled list of all Czech Wikipedia revisions (result of Code listing 4.1)