WMF Design Strategy 2022-03

Content Moderation in Medium-Sized Wikimedia Projects

This project aims to fill knowledge gaps in our understanding of how editors curate and moderate content on Wikimedia projects outside the largest and most well-researched communities. We analyzed data, researched policies and processes, and interviewed editors from small and medium-sized projects. In doing so we gained new insights into editors' needs and pain points when engaging with content moderation workflows, including writing policy, reviewing recent edits, and using administrator tools. This report documents an array of findings, which are grouped into three sections: challenging our assumptions, accessibility of moderation, and the (in)visibility of moderation. Recommendations are provided for product improvements to address the needs of content moderators on medium-sized Wikimedia projects.

- Claudia Lo and Sam Walton

Research questions Main findings Recommendations

We understand very little about how moderation needs on Wikimedia projects vary at different project sizes.

The value and reliability of Wikimedia projects comes not just from the openness of contributing new content, but also from the ability of editors to build on, curate, and moderate the contributions of others. For Wikimedia communities to grow and mature successfully, they need the tools to support both sides of this equation. Editors need to be able to effectively enact and update policies, review new contributions from others, revert and delete content which contradicts their project's rules, and build processes which facilitate these workflows.

While much research has been carried out to understand the pain points for these processes on the largest Wikimedia projects, particularly English Wikipedia, relatively little is known about the needs of contributors on smaller projects. These communities have fewer editors, especially those with advanced rights, but also have less edits and pages to moderate. Fewer technical contributors also means that these communities don't create as many of their own technical solutions for content moderation as in larger communities. How do these factors influence their ability to effectively moderate content?

For the purposes of this project we define 'content moderation' on Wikimedia projects as acts that contribute to the ongoing governance and maintenance of those projects. Broadly, we could understand "moderation" as contributions to Wikimedia projects that do not fall neatly into the category of either content or technical contributions. Such work would encompass actions like reviewing new changes, reverting bad edits, article categorization, template creation, conflict resolution, policy writing, and more.

These are types of work that are critical in order to produce healthy, well-supported communities, especially as projects grow in areas such as content and editor count. For this project this definition does not include user moderation tasks such as blocking users or using the CheckUser tool, though some insight into these processes may arise naturally.

'Content moderators' are editors who engage in content moderation activities. This definition is inclusive of administrators and users with advanced user rights, but also includes any editor who participates in these processes.

In this research we focus on 'medium-sized' projects - those which are neither the largest projects, with substantially sized communities, nor the smallest, with very few local active editors. We expect that this will highlight the problems faced by smaller communities which have grown enough to have a stronger need for good content moderation processes and tools, but may lack the full array of options available in the largest communities.

Primary research questions

As Wikimedia projects grow, how do their moderation needs change?

While we have access to data on the current number of administrators, editors, and edit numbers over a given span of time, it is harder to track content moderator growth. Specifically, we have not previously studied how administrative needs may change as the communities they serve grow or shrink. Our mental models for moderation tend to categorize projects as “large” or “small” based on their article counts, editor bases or edits per month, and we have until now assumed moderation is a function of one or some of these metrics. In short, we need to update our understanding of how the size and rate of growth of a project will impact its administrative and moderation needs.

Do Wikimedia projects have different moderation priorities corresponding to their size and growth?

Related to the above, we have long observed that projects at different sizes may have radically different moderation set-ups and procedures. This can range from the simple, such as the number of administrators, to the complex, such as the number of specialized administrators, the availability and maintenance of third-party moderation tools, and the creation, adoption, and maintenance of specialized procedures governing how all the aforementioned tools are used. Where our understanding is less complete is in how projects go from a “blank slate” with no administrators, to a complex moderation system underpinned by years of convention, documented policy, and technical tools representing thousands of hours of volunteer labour.

Is growth in moderation capacity observable and measurable?

Generally speaking our metrics around moderation capacity are relatively rudimentary. We have access to straightforward metrics such as the number of users in a given user group (such as all the administrators of a project), or the number of monthly active administrators, defined as the “number of users who make at least one action changing user blocks, page protection levels, page deletion status, or the rights of other users in an average month”.[1]

However, that does not include other forms of moderation work (such as template management, AbuseFilter management) nor full counts of administrators, including currently inactive ones. We also do not have easy understandings of rates of administration

[1] Definition comes from the Product Analytics Wiki Comparison Dataset.

action, nor how they impact editing behaviors on-wiki.

Are certain moderation activities universally necessary on Wikimedia projects?

Because previous content moderation research and development has focused almost exclusively on the largest Wikimedia projects, we don't have a good sense for the processes and activities being used on smaller projects. We want to better understand if these broadly reflect the same processes as on larger projects, or if there are certain activities which are only necessary in projects of a certain size.

To what degree do barriers to moderation work arise from technological, design, accessibility, or capacity constraints?

Although many of our communities weigh heavily towards a virtual presence, editors nonetheless must deal with material and environmental constraints on their ability to participate. Additionally, editors may be barred from conducting moderation work due to accessibility issues.

On a broader scale, most wikis have very few administrators. The December 2021 Wiki Comparison provides some illuminating figures. Of the 741 Wikimedia projects, the median number of active monthly administrators is less than two. 694 (or 94%) of these projects have ten or fewer monthly active administrators. Does this severely constrain a project's ability to carry out moderation tasks?

Methodology

Identifying partner wikis

For this research we identified two Wikimedia projects as partner communities to focus our learnings on. Starting with the framing that we would focus on 'medium-sized' projects, we defined two criteria to narrow our focus. We consider Wikipedia projects which:

- Have opted-out of global administrator support, implying they have local capacity to perform core content moderation functions;
- Have between 10 and 50 monthly active administrators.

This results in a list of 27 language Wikipedias based on data from December 2020. By considering a number of factors including prior and ongoing engagement with the Wikimedia Foundation, geographic diversity, and unique circumstances which might affect this research[2], we selected the Ukrainian and Tamil Wikipedias as our target projects.

Ukrainian Wikipedia

The Ukrainian Wikipedia is the 17th largest Wikipedia project by number of articles. As of December 2021 it had 32 monthly active administrators, and 1,093 monthly active editors. The community is growing at a modest

[2] For example, Persian Wikipedia began trials to require user registration to edit, which directly impacts administrator activities.

rate, with a 5-year growth in active editors (2016-2021) of 18%.

On the Ukrainian Wikipedia the community has established a range of social and technical processes to facilitate content moderation. The project has an extensive range of user rights, including Patroller (патрульний), rollbacker (відкочувач), Interface Administrator (Адміністратори інтерфейсу), Bureaucrat (Бюрократи), CheckUser (Чек'юзери), and Oversighter (Приховувачі), in addition to an active Arbitration Committee. It also has a comprehensive range of policies and guidelines governing expected behavior on the project, most of which have been substantially edited and updated since their creation.

Administrators on the Ukrainian Wikipedia are elected by the community, and a process exists for requesting the removal of a user's administrator rights. Administrators are an active sub-community within the project - in 2021 they gathered for a 'forum of Wiki project administrators', at which project administration approaches and responsibilities were discussed.

Tamil Wikipedia

The Tamil Wikipedia sits at the smaller end of the range of Wikipedia projects considered for this project, being the 38th largest Wikipedia project by number of articles. As of December 2021, Tamil Wikipedia had 9 monthly active administrators[3] and 95 monthly active editors. While the Tamil Wikipedia is growing at a slower rate than the Ukrainian Wikipedia, at just 7% active editor growth over 4 years (2017-2021), it has a significantly higher potential for growth, with more than 75 million native speakers and strong growth in the number of speakers gaining access to reliable internet connections.

While fundamental Wikipedia policies on the Tamil Wikipedia are documented, they are fewer in number and shorter in length than on the Ukrainian Wikipedia. Many have been only sparsely edited since their creation, which was typically through a translation from the English Wikipedia circa 2005-2007. As a result, some pages (such as Wikipedia:Administrators) contain a large number of 'red links', which lead to unwritten guidance or policy.

The Tamil Wikipedia has three active content moderation user rights: Patroller (சுற்றுக்காவல்), Rollbacker (முன்னிலையாக்கர்), and Autopatrolled (தற்காவல்). The community has one interface administrator and a few bureaucrats, but has no CheckUsers or Oversighters. On-wiki administration processes are generally low activity venues. The 'articles for deletion' process, for example, has been abandoned since 2014 in favor of article talk page discussions, and while the AbuseFilter

[3] Project selection (requiring 10-50 active administrators) was made based on December 2020 data when this value was 11.

extension is enabled, it has seen no activity from local users since 2013.

Interviews

We decided to conduct this research primarily through targeted user interviews with our selected communities. Aside from research done to identify and select the Tamil and Ukrainian Wikipedias, we did further research to understand the shape of their moderation landscapes. Our aim was to ensure that our interviews could focus on exploring our research questions, rather than taking up our time (and our participants' time) by asking basic questions about their policies and processes.

We reviewed existing research, explored quantitative data, investigated on-wiki processes and contributions, and spoke to Wikimedia Foundation staff. This helped to shape our investigations and determine our research process and lines of questioning.

For the interview portion of this study, we spoke to more than 25 editors from 16 Wikimedia projects. Editors were primarily content moderators, administrators, and cross-wiki patrollers, and we also spoke to some Stewards and tool developers. Most editors primarily contributed to small or medium-sized projects, including the Afrikaans, Bengali, Czech, Punjabi, and Turkish Wikipedias. Some contributed to non-Wikipedia projects, though these were in the minority. These interviews included three editors each from our two target projects, the Tamil and Ukrainian Wikipedias.

Synchronous interviews were conducted over the Google Meet video-conferencing software. Upon request, we provided consecutive interpretation for interviewees so that participants who were uncomfortable being interviewed in English could speak in a language of their choice.

We also spoke to editors asynchronously, via email, live chat, and on-wiki. These discussions were briefer, and primarily related to individual lines of questioning that arose during the desk research portion of this project. These interviews were too brief and sporadic to draw specific conclusions about other Wikimedia projects, but helped to inform our overall direction nonetheless.

We faced some challenges in recruiting editors to interview for this project. We originally attempted to recruit using translated messages in the target projects, posted to Village Pumps or equivalent community gathering noticeboards, and via direct emails. Our response rates remained very low until we managed to reach out to bilingual community members to affirm our commitment to the project as well as encourage participation more generally.

On the Tamil Wikipedia our response rate was particularly low, though once we made contact with one active editor, they were able to collate input for us from other editors. This community was also more comfortable with a group call, in which one editor interpreted for us, rather than individual interviews.

On the Ukrainian Wikipedia our interviewing process was cut short due to the Russian invasion of Ukraine. Prior to this, we found the most success by connecting to bilingual members of their chapter, Wikimedia Ukraine, and inviting them to both participate in interviews and recommend editors whose perspectives we should include in our work.

Main findings

During this project, we discovered a few key themes that shed light on the obstacles faced by moderators of medium-sized Wikimedia projects.

We categorize our findings into roughly 3 groups: **Challenging our assumptions**, **accessibility barriers**, and **visibility of moderation work**.

Many of our findings challenge or overturn assumptions we had about the relationship between wiki size and moderation needs. Another common theme was the ways in which access to moderation was convoluted, or limited to very specific set-ups that constrained moderator flexibility. Lastly, the visibility - or lack thereof - for moderator work expresses itself as an inability to accurately assess impact, or to find new or existing tools.

Our findings include both social and technical observations. As one interviewee noted, it is unlikely that there are technical solutions to some of the problems faced by content moderators. Some workflows, for example, simply require an editor to read and summarize a long discussion or discuss a dispute with fellow editors.

Challenging our assumptions

Content moderator needs vary with project size

One of the assumptions we held at the outset of this project was that small Wikimedia communities, due to their small (and sometimes nonexistent) administrator pools, might require the greatest amount of help from the Foundation when it came to developing new moderation tools. In other words, Wikimedia communities would require more assistance the smaller they were. Over the course of our research, however, this proved to be more complicated than we had assumed.

In short, **a small admin pool does not automatically mean that a community is understaffed**. Small Wikimedia projects have slower activity rates, so most routine moderation actions can be handled by their few moderators. More complex cases may be handled by global administrators and stewards, though some interviewees noted that it wasn't always clear how to request assistance from global contributors.

Patrolling new edits is a common activity for content moderators and has been a focus of Wikimedia Foundation research and product development in the past, being a core content moderation workflow all projects engage in.

On large Wikimedia projects patrolling is challenging due to the volume of edits being made in quick succession. The problem for patrollers on these projects (and also for cross-wiki patrollers) is to understand *which* new edits require their attention. As such, there are a wide variety of patrolling tools which have been developed to tackle this problem. These include Huggle, SWViewer, and WikiLoop DoubleCheck. The primary purpose of these tools is to enable editors to take rapid actions against automatically-prioritized edits, while skipping edits which are unlikely to require attention, such as those from experienced contributors.

On smaller Wikimedia projects, however, such edit filtering tools are less necessary. In many cases, editors can review all recent edits on a daily or even weekly basis and feel confident that they have reviewed every edit which requires oversight. **As projects grow towards a 'medium' size, however, they may find the need for such tools arising**. On weekends or following significant world events, for example, the volume of editing on a smaller project may reach a level where additional filtering tools are valuable, while on a typical weekday the volume is manageable.

In the smallest projects, the RecentChanges interface was often cited as more than sufficient, even without advanced filters such as those provided by ORES.

Administrators on smaller Wikimedia projects can also often take actions without the overhead of lengthy processes or policies. On projects with just a few administrators, processes like page deletions are more likely to be enacted without discussion or strict adherence to policy. This means that **administrator actions are often easier to make on small projects than on larger ones**, owing to less bureaucratic processes.

Categorizing wikis by administrative capacity is complex

Given that our original assumption, that small administrator counts signaled wikis in need of aid, was unreliable, we needed a new way to identify productive areas of intervention.

Our second hypothesis was that a rapidly-growing wiki might outpace the administrative abilities of its moderators, and that expanding the effective reach of each human administrator under such an environment would be directly beneficial.

Loosely, we would define administrative capacity as a function of their number of monthly active administrators over the number of monthly new editors. We used the data provided by the Product Analytics team as a basis for our analysis.

Monthly active administrators are defined as the "number of users who make at least one action changing user blocks, page protection levels, page deletion status, or the rights of other users in an average month”. Monthly new editors are the “number of editors who register and make 5 or more content edits in an average month”.

We chose “monthly active administrators” as a measure of the actual number of administrators working on a project, rather than the total number of users with the administrator user right. While some wikis have rules around removing administrator permissions from accounts that have been inactive for a given duration, this is not the case universally and there are examples where inactive or minimally-active users have retained administrator status for a long period of time.

This is especially true of small wikis, since one criteria for being excluded from global moderation is that they must have at least two administrators. Our interviews with a Punjabi Wikipedia admin revealed that, in order to retain project autonomy, **some smaller communities will allow inactive administrators to keep their status, especially if they are fluent or native speakers of the language in question**. They may also have an unspoken arrangement whereby this inactive user makes one or two cursory edits a year in order to technically remain “active” and prevent global editors from retaking responsibility.

Comparing the number of active administrators against the number of monthly new editors gives us a better sense of how the rate of growth of the wiki compares to the size of its administrative group.

Communities import content moderation policies and processes from more established projects

Content moderation on Wikimedia projects takes place through a variety of local and global processes. Each project may have its own specific way of handling similar moderation issues - for example, how to delete articles that should not remain on the project - but how this is done varies.

In practice, **most wikis are too small to take on the administrative burden of creating such processes from scratch**. Therefore, they rely on copying and adapting existing policies from larger wikis, typically one with some language adjacency. English Wikipedia was an often-cited example as a “source” for such policy adaptations. In some cases further changes are made over time by the community, but **in the majority of smaller projects these pages are very slow to update and change**, even when their origin page has been further expanded or clarified.

The smallest Wikimedia projects tend to have few well developed policies and processes relating to content moderation. On projects with no more than a handful of administrators, editors make decisions based on their personal views about what content is or isn’t permissible on their project. This enables administrators on these projects to move quickly and work directly with other editors, but in extreme cases it can lead to a form of project capture, such as in the case of the Croatian Wikipedia.

Because policies and processes tend to be replicated from one Wikimedia project to the next, with relatively few nuances, **there are a number of content moderation processes which are extremely similar from one project to the next**. These processes are common enough that we could practically consider them universal needs for Wikimedia projects. Examples of these content moderation processes include:

- **Speedy deletion** - an editor can denote a page as not adhering to a set of specific documented rules. Administrators patrol tagged articles and will delete the article immediately after confirming it meets the criteria. Examples include copyright violations and obvious spam.
- **Deletion discussions** - an editor can nominate a page for deletion, initiating a discussion. Other editors participate, and after a set period of time an administrator closes the discussion and enacts the consensus. On some projects there are limits on who can participate in the discussion.
- **Maintenance tags** - an editor can add a notice to an article denoting that it may require additional work to meet some standard or policy, both as a warning to readers and also as a way of categorizing the article for editors.

Large Wikimedia projects have an outsized influence in defining these procedures, as their policies are then used as the standard for smaller wikis looking to set up similar processes. However, **the presence of these policy and process translations does not automatically mean that the community *follows* this process**, only that someone at some point saw fit to copy over those procedures. We found many examples on small projects where these pages appear to have been translated and then largely ignored.

No moderation task is isolated

With the exception of extremely cut-and-dried cases of speedy deletions or reversions, moderation tasks consist of more elements than a simple technical removal, reversion or block. Even a relatively straightforward block of a vandal involves applying messages (templated or otherwise) and closing out discussions on reporting venues. Therefore, it may help us to understand moderation tasks as **a bundle of related technical, communicative and documentary actions** as we explore future product recommendations.

Registration requirements will change content moderation priorities

Historically, the vast majority of pages on all Wikimedia projects have been editable by any reader at any time, without needing to register an account. While this has been credited as one of the reasons for Wikipedia's success, it is also a substantial avenue for vandalism.

In 2020, the Portuguese Wikipedia voted to restrict editing to users who had registered an account, preventing ‘anonymous’ editing entirely. The discussion cited concerns around the high percentage of vandalism which came from unregistered users, the lack of anonymity for these users, and the difficulty of communicating with them.

Research showed that **the registration requirement had a substantial impact on content moderation on the Portuguese Wikipedia**. The revert rate of new edits decreased by approximately 46%, the number of page protections by 66%, and the number of blocks by 82%. Qualitatively, editors reported that they felt less stressed, and could focus on other kinds of contributions. The overall effect was to reduce the workload on administrators as pertaining to anti-vandalism workflows. Early results from a similar experiment on the Persian Wikipedia are revealing similar trends.

When comparing the most common administrative actions (protecting pages, deleting pages, and blocking users), there was notably no change in the number of page deletions after registration requirements were put in place for Portuguese Wikipedia. This is likely a reflection of their pre-existing restriction on unregistered or IP editors creating new pages. However, this is not universally true. Therefore, if a broader rollout of this change were to occur, we would expect whether or not unregistered users were already capable of creating new pages to be an indicator of how moderator workloads shift in response.

Accessibility of moderation

Moderation on mobile is so poor as to be practically unusable

25% of Wikimedia contributors primarily edit from a mobile device. Despite this, **none of the moderators we interviewed reported using the mobile web interface to perform content moderation actions on a regular basis**. Editors tended to only use the mobile interface for basic content moderation actions or in emergencies, and otherwise wait until they are able to use the full desktop editing experience before performing even slightly complex edits, such as posting a message after deleting a page.

> ***“reverting [an] edit is the sort of quick editing that should work well on mobile. Unfortunately, the mobile interface makes this all but impossible”*** – *User:Dvorapa* [1]

By investigating the mobile web experience we found that **many basic functions are missing or unoptimised when initiated from mobile devices**. Undoing edits, for example, is only available on certain pages, administrator tools are difficult to use because they do not have optimized interfaces, and **basic security features like changing one’s password is not possible from mobile web**. Additionally, many of the features which are present only become available to a user when they have turned on the ‘Advanced’ editing mode in their mobile settings.

One moderator noted that they use the third-party patrolling tool SWViewer on mobile, rather than RecentChanges, because it has a functional mobile user interface. On the mobile web version of the RecentChanges page, users are unable to undo edits, meaning they are not able to effectively patrol new edits without switching to the desktop interface.

On Tamil Wikipedia just 5% of mobile edits are made outside of direct article contributions, compared to 27% on desktop interfaces, implying that **mobile editors aren’t participating in ‘behind the scenes’ content curation processes**.

Some moderators expressed a desire to be able to do more content moderation tasks on mobile, but only if they were simple, discrete, and didn’t involve much writing. Others were content to use the desktop editing experience and weren’t interested in further developments to mobile capabilities. This is likely to be a reflection of the 'survivor bias' in the contributors who have taken on roles of content moderation. In other words, **mobile users generally don’t become content moderators**.

Common content moderation processes are complex

The commonly imported and re-created content moderation processes documented above are both hard for new editors to engage with, and often require multiple steps even for experienced contributors.

For all three of the processes mentioned above - speedy deletion, deletion discussions, and maintenance tags **- editors are usually required to know about specific templates and project pages** which aren’t signposted from a specific article. To request speedy deletion of an article, for example, editors usually need to know about either a generic template, such as {{delete}}, or a specific template which flags the article against a specific criterion. New editors have no way of knowing these templates exist without being told about them by another editor or finding the relevant help or policy pages.

Deletion discussion processes are particularly hostile to both new and experienced editors. This workflow is typically a multi-step process involving the addition of templates and creation of pages.

Because processes like reporting bad content for discussion typically require multiple steps of precise template work, technical volunteers on some Wikimedia projects have created gadgets and user scripts to make the process easier. Chief among these is Twinkle, but other implementations exist on some larger projects, including NominateToDel (uk.wiki), FastButtons (pt.wiki), and Tagger (numerous projects). These tools have strongly overlapping components, **duplicating developer efforts as they re-implement the same features**. There are numerous outstanding or in-progress efforts to localize these tools for communities without them, particularly for Twinkle,[4] but doing so is challenging because of the nuances and template mapping required on each project. **Many efforts to localize Twinkle have stalled or failed** as editors discover individual components of the tool which are hard to precisely map to their local processes.

Even if a community is able to set up one of these tools to make engaging with content moderation processes easier, editors still need to discover that such tools are available to them. Gadgets and user scripts are often challenging to discover and install, so their use is generally limited to experienced editors. Each installation is also locally maintained, and therefore prone to issues.

Another negative side effect of these similar, but re-implemented, processes is that cross-wiki patrollers find these systems difficult to engage with. While a member of the Small Wiki Monitoring Team might know that a community is likely to have a speedy deletion process, for example, they are unlikely to know precisely which template or format to use to engage with that process. This means they cannot respond to issues as quickly, and must first investigate the correct steps to take.

Moderation tools are a deterrent for new editors

In addition to considering the difficulty of using moderation tools for experienced editors, it is also worth evaluating their impact on newer contributors. Because content curation and moderation is so common, it is highly likely that new editors will find themselves on the receiving end of some of these tools and processes.

Early results from the ‘Understanding Editor Drop-off’ research project have found that **new editor retention decreases substantially when newcomers' edits are reverted**, particularly when no follow-up communication is provided.

These effects are especially pronounced for editors from underrepresented demographics. In Sue Gardner's “Nine Reasons Women Don't Edit Wikipedia” blog post, editors note that **content moderation processes are a significant factor in the retention of female editors**. In particular, female editors report a sense of discouragement when their contributions are reverted or deleted and they receive templated or hostile messages. Another study found that women reported more negative responses to critical feedback than male editors.

> ***“[Contributing] to more prominent articles makes one paranoid, anyone can come along and undo your work and leave nasty messages and you get very little oversight.”*** – *Joyce* [2]

[4] It's worth noting that a recent Rapid Grant-funded project made this process easier.

80% of all first warning messages are at least partially automated, coming from bots or power tools. **New editors' first interactions with other editors are predominantly defined by these warnings and system messages** regarding edits which don't conform to project policy. This has a large impact on their desire to continue editing.

When considering the growth of smaller Wikimedia projects, this impact likely has an outsized effect due to the overall smaller pool of both new and experienced editors.

The Flagged Revisions extension requires further analysis

The Flagged Revisions extension is deployed on approximately 50 Wikimedia projects, including 24 Wikipedias. The extension enables communities to prevent edits from unregistered users from displaying to readers until reviewed and approved by an experienced editor. This significantly changes the edit review workflow for content moderators, in addition to altering the experience for unregistered users. On some projects, like the German Wikipedia, Flagged Revisions has become a core anti-vandalism workflow.

Research on the impact of this extension, both for unregistered users and content moderators, is lacking. Studies from 2008, 2010, and 2019 were largely inconclusive, without a clear consensus on whether the extension is more helpful or harmful to community health. More recent research from 2022 suggests that the extension may have more upsides than downsides.

Concerns have been raised that some projects, including Arabic and Indonesian, have a prohibitively long backlog of unreviewed changes, where articles may appear to readers to be rarely updated, due to the hidden edits from unregistered users.

Despite concerns, there have been multiple requests for Flagged Revisions (or the lighter version implemented on projects like the English Wikipedia, called Pending Changes) to be enabled on other Wikimedia projects. As of 2014, however, the Wikimedia Foundation has stated that these requests will not be approved, stating that the extension results in a substantial negative impact on community health.

From a technical perspective, the Flagged Revisions extension has not been officially maintained for more than a decade, receiving primarily volunteer support. It has a multitude of configuration options, generates a substantial database footprint, and has a large amount of technical debt.

(In)visibility of moderation

Discoverability of moderator tools is very poor

In our interviews with Tamil Wikipedia admins and patrollers, we found that some of the new tools editors are requesting are already provided by tools such as Title Blacklist or AbuseFilter. However, **they did not seem to be aware of these available tools**, or if they were they were not mentioned, suggesting that these administrators do not know how to fully employ all the features of these tools. This is unsurprising, given the tools generally have poor documentation that is further compounded by their complex nature, majority-English documentation, and the fear of causing catastrophic accidental misapplication. **There are very few ways to contain or mitigate accidental applications of tools like AbuseFilter**. Any methods for undoing or mitigating such accidents require as much, if not more, technical understanding as it takes to merely operate the tool. Given the difficulty of learning to use the tool plus the risks of misapplying it, this may discourage adoption by administrators.

Administrators feel overworked and understaffed

A common theme in our interviews was a feeling that there weren't enough administrators on a project. **Administrators almost universally stated that they felt there was too much work to do for the number of active administrators**, and that having more administrators would be a welcome change.

Interviewees noted the strict requirements for gaining adminship on most Wikimedia projects. On Ukrainian Wikipedia there is a 75% support requirement for a new administrator to be elected, but one interviewee highlighted that it was also difficult to remove administrator rights from users who were misusing the tools, trapping the project in a kind of status quo where administrator numbers don't change substantially over time. A Tamil Wikipedia administrator noted pride that in their community it is substantially easier to gain administrator rights than in most. This seems to fit into a trend whereby **administrator rights are given more freely the smaller the community**.

> ***"[on the Ukrainian Wikipedia] we cannot elect any new admins and we cannot remove old admins"***

Interviewees in the Ukrainian Wikipedia also highlighted the difficulty of onboarding new administrators. Much attention has been given to the onboarding experience for editors who are new to Wikimedia projects entirely, but relatively little work has been done to improve onboarding for new administrators. **Administrators gain a large range of impactful tools, but little guidance on how best to leverage them**. This leads to hesitancy both for potential administrator candidates and for the wider community voting on new administrators.

Additionally, some candidates described administrator tasks as less rewarding, due to a lack of visibility or end result of the work, and highlighted how resolving interpersonal disputes and summarizing discussions can take a particularly long time. Even basic statistics on routine administrative tasks, such as “number of pages deleted” or “number of blocks handed out in a given time period”, are difficult to find and collect. None of the administrators we talked to mentioned any sort of tracking ability or metric dashboard that would allow them to assess the health of their project, or the impact of their moderation actions.

The number of administrators on a project can also be misleading. While Tamil Wikipedia - for example - has 33 editors with the administrator user right, only 10 of those are active editors, and the vast majority of administrator actions are taken by just 5. Electing just one new active administrator can have an outsized impact in these smaller Wikimedia communities.

Product recommendations

Our project turned up several areas for improvement where product solutions could have a large impact on moderator burdens. We present them here in rough order of priority.

Improve content moderation from mobile

As we detailed in one of our main findings, content moderation is considerably harder on mobile than via the desktop interface. This raises barriers to new editors who wish to engage in contributions beyond direct article edits. Moderation functions such as undoing an edit, viewing a detailed history page, and moving pages are hidden behind an ‘advanced mode’ which users must opt-in to on each Wikimedia project they edit. These and other moderation tasks also tend to be cumbersome to navigate on a small screen, with many of the interfaces unoptimised for small devices. Some workflows, such as patrolling recent edits, are practically impossible on mobile devices without switching to the desktop skin.

One in four contributors to Wikimedia projects edits primarily from a mobile device, a figure which increases to 40-60% in some emerging markets. Worldwide engagement with the internet has shifted predominantly towards mobile, with more than half of all internet traffic now coming from mobile devices. Investment in mobile contributions therefore must be a priority to ensure that communities continue to maintain active contributor bases and have opportunities to grow, especially in emerging markets.

Additionally, recent global disruptions such as the outbreak of war in Ukraine point towards the fact that we cannot assume people will always have the luxury of easy access to a stable internet connection on a desktop or laptop device. As we expect disruption and displacement to continue to occur over the coming decades, improving content moderation on mobile devices will be key to ensuring that all communities have access to the tools necessary to create safe and healthy communities that allow equitable access to global knowledge.

> ***"I would love the mobile tools to be improved so I can curate on mobile more easily."*** *– User:Deryck Chan [3]*

Content moderation functionality in the mobile skin needs to be brought up to parity with the desktop editing experience, to ensure that all editors can contribute to content curation, regardless of their editing device. By building on the Advanced Mobile Contributions project, further improvements could be made to bring more features up to a quality where they can be part of the default experience, rather than requiring editors to opt-in to an advanced editing mode.

Improvements could also be made to further structure content moderation tasks, enabling processes like reviewing speedy deletion candidates to be streamlined to a few button presses.

Working within the mobile web interface could also be a beneficial starting point for work in this space. Issues of accessibility and discoverability are accentuated here, making it a good avenue for developing hypotheses to solve other challenges, such as the discoverability and complexity of moderator tools.

Make moderation tools and processes more discoverable

To improve editors' ability to discover and navigate content moderation processes, more direct functionality should be added to the user interface for editors to enact common content moderation activities. Most prominent among these includes adding a maintenance tag to an article, flagging an article for speedy deletion, or initiating a deletion discussion.

This would benefit new editors, who are unlikely to know about these processes or how to navigate them. It would benefit experienced editors who are currently required to install a gadget or navigate multiple steps and know where processes and templates are located. It would also benefit cross-wiki patrollers, who could be confident that they will encounter an understandable process on a new wiki.

By centralizing these features in the user interface, rather than in gadgets only available on some Wikimedia projects, the output of content reporting systems could be structured, enabling other improvements. These might include better APIs for community tools, further improvements to reviewing interfaces for reported content (i.e. moving away from categories and bot-curated pages), and surfacing reporting and reviewing processes as structured tasks, including on mobile devices.

Since this functionality is currently replicated across many different gadgets and user scripts, it would also reduce the burden on the volunteer technical community to build and maintain these tools.

Improve communication from content moderation tools

To improve the retention of editors who come into contact with content moderation tools, the messaging of those tools should be improved.

Improvements could be made to system messages, such as the notification received when edits are reverted. Better templated messages could be provided for use in user-maintained tools. In some cases, no default messaging is provided at all, such as for page deletions, and adding these messages could be impactful to improve new editor understanding of what is happening to their contributions.

Content moderation could even be explored as a means to onboard newer editors into more ‘behind-the-scenes’ processes. One interviewee reported that they only started working on content moderation processes after one of their articles was nominated for deletion. This communication can reveal new processes to users, and could therefore be leveraged to provide onboarding, indicating to the user that they can participate in turn.

While these changes may have unclear benefits for content moderators, they could have significant impacts on new user retention, and the topic of welcoming communication should be strongly considered when developing new tools.

Implement other common gadget functionality into MediaWiki

In addition to content reporting, tools like Twinkle include other content moderation functionality which is commonly requested on projects which don’t yet have these gadgets setup. Many of these tools implement complex tasks in a relatively simple way, which would help with content moderator onboarding and tool discoverability.

For Twinkle, the most commonly requested features are the ‘Restore and rollback’ features, which enable editors to revert to a specific earlier version of a page, and to use a rollback-like feature which additionally sends the rolled back user a templated message. Restoring a page to an earlier version in particular is a feature which could be built into MediaWiki directly and therefore made available to editors on all Wikimedia projects. This is currently only possible by manually navigating to an old version of a page, editing it, and saving the page.

Improve guidance for new editors

To reduce the content moderation workload for experienced editors, more work could be done to improve guidance for new editors. If more good faith editors avoid common pitfalls when creating content, the overall quantity of actions which need to be taken will decrease, reducing the burden on content moderators.

Many speedy deletion criteria on Wikimedia projects are easily quantifiable - an article needing to have more than a set number of words, for example. If an editor attempts to create an article but is in violation of a speedy deletion criteria, information could be provided to them to avoid creating that article, which would likely be quickly deleted.

This could improve both the onboarding experience for new editors in addition to reducing the workload for content moderators.

Carry out user research with the Flagged Revisions extension

The Flagged Revisions extension, despite being deployed on dozens of Wikimedia projects, is unmaintained and poorly understood. Some communities have become accustomed to using it and it is a core component of their content moderation workflows. However, the Wikimedia Foundation has declined to deploy the extension to new projects, and it is not currently maintained.

This falls in line with our main finding about the general invisibility of moderation work. Despite this being deployed on many wikis, this tool remains unmaintained. Additionally, we are unsure if the wikis on which Flagged Revisions is deployed continue to use it, or if they have found (or would prefer) alternate tools to perform a similar function. This is a useful example of our standing bias towards assuming the tools used by a handful of large wikis for moderation must be the tools used by *all* wikis for moderation.

Recent research, which suggested that the extension may have a positive overall impact, found it has a number of deficiencies which could be remedied, including poor feedback to affected users and an unintuitive interface for content moderators.

User research should be carried out to understand how the Flagged Revisions extension could be improved for both affected new users and content moderators. Alongside this a technical analysis could identify technical areas for improvement which would make the extension more stable and able to be deployed to more projects.

Notes

[1] User:Dvorapa, Community Wishlist Survey 2019/Mobile and apps/Add an undo/revert button to diff view - Meta

[2] Nine Reasons Women Don't Edit Wikipedia (in their own words) | Sue Gardner's Blog

[3] User:Deryck Chan, Moderator Tools/Wikimania - MediaWiki