Harassment on Wikipedia

T318022 / Aishwarya Vardhana, September 2022

WIKIMEDIA FOUNDATION

1. **Purpose**
2. Questions
3. Themes
4. Review

WIKIMEDIA
FOUNDATION

Purpose

This review a synthesis of harassment research by WMF from 2015-2022 that identifies **major themes** in the problem space as well as **user needs**, **challenges**, **considerations**, and **previous work**.

WIKIMEDIA
FOUNDATION

1. Purpose
2. **Questions**
3. Themes
4. Review

WIKIMEDIA
FOUNDATION

Questions about harassment on wiki

- **Who** experiences harassment and why?
- **What type** of harassment do people experience?
- **How frequent** and long is the harassment?
- **Where** does harassment happen?
- **What do people do** when they're harassed?

WIKIMEDIA
FOUNDATION

Questions about the process of reporting

- How do people **discover** reporting?
- What **percentage of people report**?
- **Why don't** people report?
- What percentage of people **don't know how** to report?
- Do people **want to report**?
- What kind of **reporting paths** exist on-wiki and how effective are they?
- How **effective** does the community perceive these current processes to be?

WIKIMEDIA
FOUNDATION

Questions that go beyond wiki

- Do people want **private reporting** and to what degree?
- What percentage of people go **off-wiki?** Why? Where do they go?

WIKIMEDIA
FOUNDATION

Questions about admins

- Do Wikipedia **admins experience harassment**? What type and where?
- How confident are they at **identifying and stopping** harassment?
- What do they think **they need** to make **Wikipedia safer**?
- What types of **behavior** do admins believe is **most disruptive**?
- How much **faith** do admins have **in WMF**?

WIKIMEDIA
FOUNDATION

Questions about online harassment broadly

- What is **online harassment**, who does it, and who is involved?
- What are current **best practices** for reporting systems?

WIKIMEDIA
FOUNDATION

What don’t we know?

- What is harassment like on **medium and small wikis**?
- How many editors **leave Wikipedia due to harassment**?
- **Quantitative data** on harassment on and off-wiki e.g. How many harassment reports are officially filed on English Wikipedia per day? How many reports are successfully resolved? Can harassment reports be categorized by complexity and length of time to resolve?

WIKIMEDIA
FOUNDATION

1. Purpose
2. Questions
3. **Themes**
4. Review

WIKIMEDIA
FOUNDATION

Overview of themes*

- **Harassment is a prevalent problem**
  - Harassment of various forms (e.g. name calling, trolling, hate speech) is a prevalent problem on English, German, Italian, and Arabic Wikipedia, with content vandalism (e.g. vandalizing someone's user talk page) and trolling/flaming being the two most frequently experienced forms of harassment. While a reporting system can help people feel safer, an ineffective system also creates distrust, deters volunteers from reporting in the future, and lead to responder burnout. As far as we know, small-to-medium wikis have no formal reporting systems.

- **Existing processes and reporting systems (formal and informal) are difficult to discover and use.**
  - A significant portion of the community is unsure how to report harassment and many choose not to report a reportable offense due to fear of backlash. IRC and email are two of the three most common channels for reporting to admins, both of which are off-wiki. The other most common channel is an Admin's user talk page. Both on and off wiki systems have high barriers to entry.

WIKIMEDIA
FOUNDATION

Overview of themes*

- **Volunteers don’t report out of fear**
  - Majority of editors who are harassed don’t report publicly out of fear of retaliation or backlash, so instead they try to ignore the harassment. People have cited defensive cliques, complex issues, and biased participants as reasons for specifically not reporting on AN/I, a noticeboard for incident reports en Wiki. Additionally, the length of time it takes to resolve a report discourages people from reporting.

- **Admins experience harassment**
  - Admins on Wikipedia are harassed, most commonly in the form of being stalked (i.e. wikihounding) and their edits reverted, and most often due to unpopular administrative decisions. While admins are highly experienced Wikipedians, more than half are unsure if they have the skills to intervene or stop harassment. Admins across different platforms feel they don’t receive enough support from WMF to handle misconduct nor are they confident in the Foundation’s ability to handle it.

WIKIMEDIA
FOUNDATION

Overview of themes*

- **People want to report**
  - While there is strong dissatisfaction and confusion regarding existing reporting systems, a significant number of people still engage with them.

- **Significant support for private reporting channel**
  - There is strong support for a private reporting system that includes varying degrees of transparency. For example, a third of folks want public access to aggregated statistics so that they can see what types of reports are being filed and resolved. Another 30-40% want access to case summaries, which would allow for more transparency without necessarily compromising anonymity.

WIKIMEDIA
FOUNDATION

Contradictory themes

- **Volume**
  - Our assumption is that if we build a reporting system into the platform, more people will report. This increase in reports will need to be bolstered by increasing resources for responding to reports. Additionally, in order to create trust in the reporting system, there will need to be good communication between the system and the reporter as well as support for the responder.

- **Privacy**
  - The community and the Foundation have a wide variety of views on how to handle privacy. There is general support for a semi-transparent (translucent) system and shared understanding that complete opacity or transparency will not work.

WIKIMEDIA
FOUNDATION

1. Purpose
2. Questions
3. Themes
4. **Review**

WIKIMEDIA
FOUNDATION

Sense of safety on Wikipedia

- **Safety survey 2015 (link)**
  - Of the 3,845 Wikimedia users who participated, 38% of the respondents could confidently recognise that they had been harassed, while 15% were unsure and 47% were confident that they had not been harassed. Similarly, 51% witnessed others being harassed, while 17% were unsure and 32% did not witness harassment.

- **Engagement survey 2017 (link)**
  - This survey was created in response to wanting more alignment between WMF and the communities it serves. The section of the survey relevant to the PIRS is the series of questions on “improving community health”.

- **Safety survey 2022 (link)**
  - The Community Safety survey was conducted to understand contributors’ sense of safety on different wikispaces, and how it changes over time. In the last 30 days, have you felt unsafe or uncomfortable contributing to Wikipedia?

WIKIMEDIA
FOUNDATION

Safety survey 2015 (Source)

- **Who is being harassed?**
  - 38% have been harassed
  - 51% witnessed harassment
  - Women experience more harassment than men
  - Cultural minorities experience more harassment than culturally represented

- **Why were they harassed?**
  - Difference in point of view
  - Administrative actions or status
  - Edits or content

- **What type of harassment is most common?**
  - Content vandalism
  - Trolling/flaming

WIKIMEDIA FOUNDATION

Safety survey 2015 (Source)

- **What do people do when they're harassed?**
  - Majority ignore the incident
  - Many discuss it with the community
  - Many explicitly ask the harasser to cease

- **How effectively do targets of harassment feel they can handle harassment?**
  - 42% felt not effective at all
  - 49% felt somewhat effective
  - 09% felt completely effective

- ~80% of the survey was male

WIKIMEDIA
FOUNDATION

Engagement survey 2017 (Source)

- **What percentage feel unsafe in a Wikimedia online or offline space?**
  - 32% feel unsafe

- **What needs to improve?**
  - 50% said tools and processes for reporting users
  - 43% said noticeboards
  - 43% said blocking tools/mechanisms
  - 48% said administrative selection and review processes

- **Who do people report to?**
  - Other volunteers more than to WMF or their local chapter or affiliates
  - WMF and local chapter or affiliates are seen as less adept at resolving issues

- Most male, western audience took survey

WIKIMEDIA
FOUNDATION

Safety survey 2022 (Source)

Felt unsafe or uncomfortable contributing

- Portuguese: **23 - 29%**
- Spanish: **21 - 28%**
- English: **14 - 18%**
- French: **8 - 15%**

WIKIMEDIA
FOUNDATION

Targets of harassment (Meta, report)

This report aims to understand attitudes towards existing harassment resolution processes as well as reasons for or against participation in these enforcement mechanisms, from the perspective of those who have experienced harassment.

- **Participants report that our existing enforcement systems are overly complicated and difficult to understand.**
  - Write-in survey responses noted the existence of loopholes, unclear redirections, the expectation that one may be asked to make reports to the very people one wished to report, and an utter lack of clear instructions on how to report.

- **The current reporting system opens reporters up for reprisal, backlash, or undue public scrutiny.**
  - Many of our write-in responses specifically named fear of reprisal as a major negative in our current system. Some of them used a specific jargon term, “boomerang”, to refer to this phenomenon, suggesting that this is so common as to warrant a special name for it.

- **A slight majority of our survey respondents have never made a report.**
  - 54% of survey respondents have never made a report. This includes 40% of respondents who have, or had, held administrator rights.

- **Six in ten respondents have purposefully chosen not to report incidents.**
  - Reasons given include a fear of backlash or reprisal, belief that the outcome would be ineffective, and the process of making reports being too confusing or difficult. Write-in answers also indicated that occasionally, the people in charge of receiving reports are the very people that are the subject of complaints.

Targets of harassment (Meta, report)

This report aims to understand attitudes towards existing harassment resolution processes as well as reasons for or against participation in these enforcement mechanisms, from the perspective of those who have experienced harassment.

- **Two-thirds of non-administrator survey respondents are unsure or do not know how to report problematic behaviour.**
  - By contrast, current or former administrator respondents were far more confident in their knowledge of how to report - 83% reported that they did know how to report such behaviour.

- **There is a general desire for a private on-wiki reporting channel.**
  - When survey respondents were asked what venues should be available for reports, the most common option chosen was “other private route”; the third most common choice was “on a separate private channel, on-wiki”.

- **It takes too long to resolve cases of harassment. Length of case discourages reporting**
  - This was expressed by users making reports as well as the administrators expected to handle them.

WIKIMEDIA
FOUNDATION

Targets of harassment (Meta, report)

- **Communities without guidelines have harassers who continue to harass and evade consequences.**
  - Without prompting, we routinely heard from respondents about certain communities with a bad reputation for being especially combative or hostile. What they had in common was a lack of guidelines around behavior or reporting and a general “blind eye” attitude towards their community members’ histories of rule-breaking behavior, especially if paired with a long history of contribution.

- **Participants were divided as to what the precise role of WMF should be in enforcement systems.**
  - While there is broad consensus within the Wikimedia community that the Foundation should be responsible for certain cases involving minors or credible threats of violence, this consensus breaks down when it comes to most other matters.
  - Survey respondents alternately decried the Foundation’s involvement while also viewing it as a needed route that bypasses local reporting systems that are being handled by the people they wish to report. Others wanted the Foundation to act as a “backup” option if there were no global administrators, oversighters or stewards available. Still others were upset that the burden of handling harassment reports, especially while organizing Wikimedia events, was shifted to volunteers rather than the Foundation.

WIKIMEDIA
FOUNDATION

Targets of harassment (Meta, report)

- **Survey respondents wanted access to aggregate statistics and case summaries, not necessarily full case details.**
  - Our current systems provide full public visibility of all cases made on-wiki. However, when asked what information the general public should see with regards to reporting on Wikimedia projects, more respondents chose aggregate statistics and summaries over full case details. This was true of both administrators and non-administrators.

- **Respondents generally still view reporting as worthwhile.**
  - While users were much more likely to view the entire enforcement process as ambiguously useful at best, survey respondents were still generally positive about local admins, the WMF, and event organizers' likelihood of addressing reports. Slightly over half of survey respondents said that it was "definitely" or "probably" worthwhile to make a report. Two of our interviewees also noted that, even though they knew (or believed) that the people they reported to were powerless to act on their reports, they still wanted to make them. This suggests that the act of reporting is itself an action that people wish to perform, regardless of outcome.

WIKIMEDIA
FOUNDATION

Community sentiment (Source)

- **Who**
  - Mostly surveyed in English, ⅓ in Spanish, 1 Hindi
  - 51% women, 39% men, 4% NB or self-described
  - Majority in Europe
  - 39% Admins, 55% current or former organizers
  - Nearly all of our respondents have spent over a year on Wikimedia projects, with about a third reporting over a decade of experience with our projects.

- **What**
  - Of those who have made reports, a little over half did so to report unwanted behavior directed at themselves (52%).

- **Where do people report**
  - Notice boards – 65%
  - Email – 55%

WIKIMEDIA FOUNDATION

Community sentiment (Source)

- **Not reporting**
  - 60% of respondents chose not to report something that could have been reported
  - 54% have never made a report

- **How to report**
  - 53% know how to report
  - 31% are unsure
  - 17% don't know how
  - 55% of those who made reports felt they understood how to do so well

- **Do you know how to report harassment, bullying, or other problematic behavior?**
  - 83% of admins said yes
  - 33% of non-admins admins said yes

- **Who is reporting**
  - Admins are more likely to report than non-admins
  - Most learn how to report by themselves or through other community members

Community sentiment (Source)

- **How well do Wikimedia projects currently handle reports?**
  - 30% reports are handled **well**
  - 43% reports are handled **moderately well**
  - 27% reports are handled **slightly or not well at all**

- **Is it worthwhile to make a report on Wikimedia right now?**
  - 55% think it's **worthwhile**
  - 23% think it **might or might not be**
  - 22% think it's **probably or definitely not**

WIKIMEDIA
FOUNDATION

Community sentiment (Source)

- **Opinions on privacy**
  - 44% believe reporters should have access to some form of private reporting channel
  - **Strong belief** that there should be private channel for reporting
  - **General support** for certain cases to be private and inaccessible to the public, namely cases involving identifying misinformation, legal issues, or serious abuse or harassment cases

- **What type of information does the community want public access to?**
  - 30% aggregated statistics
  - 30-40% case summaries
  - 10-15% full case details, both resolved and in-progress

- **What type of cases should be private?**
  - Supported privacy for cases involving personally-identifying information, harassment, and legal issues
  - Write-in responses also suggested privacy upon request for the targets of harassment
  - The need to take context into account

WIKIMEDIA
FOUNDATION

Community sentiment takeaways (Src)

- **Reporting**
  - Significant plurality of people engage in reporting systems
  - People are confused or dissatisfied with the system. There is difficulty in finding and understanding the reporting system, especially for newcomers
  - Backlash and ambivalence about effectiveness discourages reporters
  - People still believe it's worth reporting

- **WMF**
  - There is confusion over precise role of WMF in reporting and enforcement
  - Distrust of Wikimedia staff processes

- **Types of cases**
  - Difficulty of bringing reports against editors who mock a class of users such as certain minorities rather than an individual
  - Feeling that current enforcement system is only suitable for simple cases

WIKIMEDIA
FOUNDATION

Community sentiment takeaways (Src)

- **Privacy**
  - Lack of capacity for local admins to handle issues of harassment
  - One write-in answer said that they were shown how to report by example, and specifically pointed to the fact that most of these reports are completely publicly visible. This may indicate that the public visibility of most current on-wiki reporting systems acts as a mechanism to teach would-be reporters how to make reports.
  - Low levels of support for all reports to be public (100% transparency) or all reports to be private (100% privacy)

WIKIMEDIA
FOUNDATION

2017 Admin survey (Source)

- **Who**
  - Most Western male audience
  - Large # of admins taking survey were male and western
  - Majority have 10+ years on Wikipedia
- **Majority can**
  - identify harassment
  - identify sock puppetry
  - identify vandalism
  - feel they're provide enough resources to solve, mitigate, or intervene in cases of vandalism
  - identify wiki hounding
- **More than 50% are unsure if they have the skills to intervene or stop harassment**
- **Significant population (>30%) feels they're not provided enough resources to solve, mitigate, or intervene in cases of harassment**

WIKIMEDIA
FOUNDATION

2018 Harassment on Arabic Wikipedia (Source)

- **Block requests were grouped into nine broad categories:**
  - Username
  - Vandalism
  - Non-encyclopedic
  - **>> Personal attack <<**
  - Propaganda
  - Sock puppet
  - Trivial accounts
  - Use of profanity
  - Engaging in edit wars
- **What did we find?**
  - Of 782 relevant requests **73.8% were sustained** by moderators
  - The Personal attack category is the **fourth-most-common category** (67 instances).
  - **The Personal attack category** report-to-block rate was under 50%. It is treated differently by moderators than reports of other categories.
  - Threats, insults, attacks, and harassment make up a significant percentage **(8.5%)** of analyzed block requests. When the disproportionately large Username category is removed from analysis, this percentage rises to **14.3%.**

WIKIMEDIA
FOUNDATION

2018 Harassment on Arabic Wikipedia (Source)

- **Types of offense**
  - Insults
    - Content
    - Personal
  - Threads and commands
    - Claim to power
    - Threat
    - Command
  - Harassment
    - Harassment
    - Alleged harassment
    - Perceived harassment
    - Possible sexual harassment
  - Political propagandizing

WIKIMEDIA
FOUNDATION

2019 Admins on Arabic Wikipedia (Source)

- Admins are **frequently subject to various forms of harassment** on Arabic Wikipedia. The vast majority occurred on-Wiki.
- **Harassment** is viewed as the most serious threat to the Arabic wikipedia community, while **trolling** is viewed as the least serious.
- Very confident **identifying** harassment, less confident **handling** it but still feel they are effective at stopping it. Admins note particular lack of preparation by the WMF for handling misconduct (M = 2.11).
- There are no significant patterns regarding how many attackers perpetrated the harassment experienced by Arabic admins, nor with respect to how long periods of harassment tended to last. **The most frequent form of harassment reported** is that another **user followed you and reverted your edits**.

WIKIMEDIA
FOUNDATION

2019 Admins on German Wikipedia (Source)

- German admins agree with their Arabic colleagues in that they **have not received sufficient WMF support** to handle misconduct

- **Requests**
  - 74% WMF-supplied training
  - 47% best practices for handling harassment e.g. “how to stop trolls from coming back.”
  - 42% dispute resolution
  - 26% training for specific tools

- German admins have been generally effective at stopping harassment cases in which they have been involved.

WIKIMEDIA
FOUNDATION

2019 Admins on Italian Wikipedia (Source)

- Italian admins are **highly experienced, engaged, and focused** on their local projects.
- They express **strong, unmet desire for technical training**, particularly *digital security*, the most-requested and least-accessed training type.
- Harassment is infrequently experienced by Italian admins, but over half have suffered negative outcomes due to harassment.
- Italian admins' experiences with harassment tend to be short in duration, with the most common practice being that a harasser **followed you and reverted your edits.**

WIKIMEDIA
FOUNDATION

2018 Reporting channels research (Source)

- **Three most commonly named channels for reporting to known and trusted admins:**
  - IRC
  - On-wiki talk pages
  - Email

- **At Wikimania 2018, a survey of administrators showed that email was the most common channel they suggested for handling harassment.**

- **How to people use informal systems?**
  - To communicate to other administrators and see if cases are already being handled elsewhere.
  - To help reporters navigate the formal system. They showed reporters which channels were most appropriate, and which ones to avoid due to inactivity or poor fit.
  - To put reporters at ease, helping them talk through their issues and de-escalate or calm down reporters before going on to consider next steps.

WIKIMEDIA
FOUNDATION

2018 Reporting channels research (Source)

- **Who's involved**
  - Moderators
  - Reporters
  - Accused users
  - Observers

- **Informal and formal systems complement each other and meet different sets of user needs**

- **Different wikis balance formal to informal differently based on policy, available moderator labor, and project values**

- **We don't want to hurt informal systems**
  - Formalizing previously informal networks can also be costly, causing these reporting paths to lose much of the flexibility and speed that makes them advantageous in the first place. We don't want to take away responders from doing important work on informal systems.

WIKIMEDIA
FOUNDATION

2018 Reporting channels research (Source)

- **Pros of formal system**
  - Purpose designed
  - Lower the barrier of participation
  - Clear documentation
  - Allows moderators to better keep track of and deal with reports
  - Provides a way to set precedents and leave traces in cases where a pattern of misconduct emerges
  - Lend reports legitimacy by codifying and claiming socially-approved values around reporting misconduct and abuse

- **Cons of formal system**
  - Set boundaries, might not capture all dimensions of misconduct
  - Rigid
  - Slow
  - Requires lots of trust by all parties
  - Unsuitable for emergencies, acute abuse or particularly complex cases
  - More easily abused or manipulated

WIKIMEDIA
FOUNDATION

2018 Reporting channels research (Source)

- **Pros of informal system**
  - Fast
  - Flexible
  - Discreet
  - Allow for more complex cases that require speed and discretion
  - Can accommodate more edge cases and involve different methods of mediation

- **Cons of informal system**
  - Opaque
  - Difficult to access → accessible to only clearly highly knowledgeable and connected editors
  - Can make both involved groups uneasy
  - Assumed to be less legitimate and not governed by same community values
  - It can be even harder to resolve a dispute in the process when that process is informal and, definitionally, somewhat shielded from outside eyes

WIKIMEDIA
FOUNDATION

2017 AN/I survey (Source)

- **49% say the discussions on AN/I are "almost never" or "rarely" focused and neutral**

- **59% want some form of more structured reports**

- **77% dissatisfied with AN/I have hesitated or not filed a report for fear of retribution**
  - Compared to dissatisfied users, users who are neutral are filing reports and weighing in on AN/I more and are admonished less.
  - Users who are neutral and users who are satisfied are somewhat more similar in their usage and feelings towards the AN/I process.
  - There is a stronger dissatisfaction expressed about **the way that AN/I cases** are handled rather than a negative view of the AN/I type process.

- **Of those who did not report they did so because of the following breakdown**
  - Defensive cliques (19.35%)
  - Complex issues, toxicity, and boomerang effect (12.90%)
  - Biased participants (9.68%)
  - (In close fourth) Easier to ignore the problem, ineffective / inconsistent, no chance of action, avoiding drama

2017 AN/I survey (Source)

- **Ideas for how to improve AN/I**
  - More moderators and clerks
  - Policy changes
  - Structured reports
  - Other technical improvements

- **Transparency and openness is working**

WIKIMEDIA
FOUNDATION

Reporting systems rubric (Full report, summary)

- **Other platforms' reporting systems will inform people's expectations of our reporting system, and so should be considered for our own future designs**
- **4 criteria for the rubric:**
  - Accessibility
  - Ease of use
  - Communications
  - Privacy
- **Reddit**
  - Strengths include immediacy of reporting option, relative ease of reporting, and all reports are private
  - Weaknesses include lack of guides around reporting and highly variable moderation style which is fond from subreddit to the next
  - Well suited to one-off instances of unacceptable content
  - Ill suited for reporting either harassment in private messages or long histories of unacceptable behavior

Reporting systems rubric (Full report, summary)

- **Facebook groups**
  - Strengths include flexibility for moderators by offering a breadth of possible actions for sanctioning users, the system captures some useful information as well. System is constantly developing.
  - Weakness is lack of clear communication when it comes to data visibility and functionality e.g. reporting comments requires making some counterintuitive choices. Although the system captures personally identifying information, it never tells users that it does so, and neither users nor moderators are certain exactly what information is visible to Facebook, and what remains within the group. This system is less trusted by its users.
  - Well suited for reporting a specific type of incident
  - Ill suited for cases more complex than this, or for longer-term issues
- **We will look at this report more closely when in the Design & Development phase of the incident reporting system project**

WIKIMEDIA
FOUNDATION

Reporting systems rubric (Full report, summary)

- **These platforms are teaching their users what to expect of reporting systems on other platforms**
- **Facebook Groups and Reddit put the "report" link in as many places as possible to make it visible**
- **Reports are standardized**
  - A standardized report greatly speeds up and structures the report
  - Makes it easier for more people to report
  - Helps moderators understand reports
  - Drawback is these templates make it difficult to report more complex cases
- **Opaque communications can lead to distrust, as we see in Facebook Groups**
- **How do we adhere to transparency in a way that is safe—both for reporters and the moderators handling reports—and respects the privacy of reporters?**

WIKIMEDIA
FOUNDATION


The community and previous product teams have attempted to design reporting systems before. We are not the first.

Previous work

- List of proposals from the community
- Anti-Harassment Tools For Wikimedia Projects
- Community health initiative/User reporting system - Meta
- The last product team to work on a reporting system laid out this structure in 2019.

WIKIMEDIA
FOUNDATION

Previous work

- Safety survey 2015
- Engagement survey 2017
- Safety survey 2022
- Targets of harassment (Meta page, Full report)
- Community harassment
- 2017 Admin survey
- 2018 Arabic Wiki Harassment research
- 2018 Admins on Arabic, German, and Italian Wiki
- 2018 reporting channels research
- 2017 AN/I survey
- Reporting systems rubric

WIKIMEDIA
FOUNDATION