# Wikimail analysis

**Wikimail is the system used by volunteers on WMF projects to send e-mail to one another by means of an online interface.**

Users are required to add a verified e-mail address in order to use the Wikimail system. Trust and Safety has identified three key issues for analysis:

- ➔ Disclosure of personal e-mail addresses in messages
- ➔ Measuring the extent of Wikimail harassment
- ➔ Utilization of the e-mail mute feature

This report focuses on the community perspective regarding these issues. There are no other pressing community concerns for this tool at the moment.

## Harassment on Wikipedia

The Pew Research Center 2021 survey on the State of Online Harassment found that roughly four-in-ten Americans have experienced online harassment, with half of this group citing politics as the reason they think they were targeted. Growing shares face more severe online abuse such as sexual harassment or stalking.

Similar surveys have shown that harassment is widespread on Wikimedia projects and that users are broadly dissatisfied with responses to reported incidents. Users can imagine how harassers might circumvent attempts to block their abuse or how new features might unexpectedly lead to harassment. Bad actors on WMF projects can be characterized as clever, diligent and relentless in their dedication to their task. In general this type of abuse seems to be normalized as an unavoidable byproduct of the culture at Wikipedia.

#### Extent of the problem

In 2015 the Support and Safety team published a community survey on Harassment across WMF projects. 3,845 Wikimedians participated in the study and 38% reported that they had been harassed. This is broadly in line with

numbers from the most recent Pew Research survey on the state of online harassment.

Many people were critical of the 2015 survey because they believed it disproportionately sampled users who had a personal interest in the topic:

*People who are interested in commenting about online harassment on Wikimedia projects will self-select into the survey sample, while people who are not interested in commenting about online harassment on Wikimedia projects will self-select out of the survey sample.*

The more recent Community Insights Surveys from 2017 and 2018 provide a better model for this type of analysis, integrating questions about harassment into a more general survey designed to attract a broad spectrum of participants. Support and Safety and the Anti-Harassment Tools team proposed questions related to this topic.

Those surveys found that between 20% and 30% of all participants had felt uncomfortable or unsafe in Wikimedia spaces online or offline. Among those who felt unsafe, 71% reported being bullied or harassed on Wikipedia in the last 12 months.

In 2017, the Wikimedia Foundation commissioned a study on anti-harassment by the Harvard Negotiation and Mediation Clinical Program. They found:

*There is general agreement among users that current systems for addressing user issues are deficient. Users have expressed that responses to behavioral issues are frequently inadequate. An analysis of English Wikipedia's incidents noticeboard found that of 3,093 reported cases in the past 12 months, only 1,745 had been resolved.*

The 2018 Community Insights survey reported that 55% did not know where to turn for help when they were being attacked on Wikipedia. 84% requested better reporting tools, 77% requested better noticeboards and 75% requested better wiki policies.

Users have expressed broad support for improving existing systems, particularly through better reporting and evaluation tools. The ongoing User Reporting System project is designed to address this community need.

#### Global / local problem

It's important not to generalize abuse or harassment across all Wikimedia projects or across all language communities on Wikipedia. Each wiki has its own processes for dealing with such complaints and some projects seem to have less harassment than others.

The Harvard study included the following disclaimer:

*We focused primarily on English-language communities as well as larger Spanish-language and Portuguese-language communities. Language was a barrier to participation. The research was limited to stakeholders who were comfortable communicating in English. Therefore it was unable to engage with the full breadth of the Wikimedia community.*

Much of the Wikipedia research suffers from a similar language bias but notable exceptions exist. The 2019 report on Harassment in Arabic Wikipedia offers a model for future research on harassment that might target a wider spectrum of languages. Deploying more project- or language-specific surveys would be an important next step.

Anecdotally it seems that some language communities are worse than others when it comes to harassment and the norms that govern the administrative response:

**Jetam2**: Some Wiki projects are more global and international (English Wikipedia, French Wikipedia, maybe Arabic Wikipedia?), others are less so (Slovak Wikipedia etc.) In my experience, there is sometimes a homeland vs diaspora attitude that can become a cause of harassment or at least cause feelings of being unwelcome.

**Nattes**: Public reporting of harassment is not possible on French Wikipedia because it creates backlash and is seen as abusive itself.

Analytics could uncover segmentation differences between wikis in their response to harassment. The Anti-Harassment Tools team completed some initial analysis on this topic for mute usage across English, French, German, Spanish, Russian, Italian, Dutch, Japanese, Chinese, and Portuguese wikis in 2019.

Finally, the 2017 report on Defining Conflict and Harassment on Wikipedia offers a distinction between conflict, harassment and abuse. It would be interesting to apply a similar frame across language wikis to better understand the problem.

## Wikimail harassment

The 2015 Anti-Harassment survey reported that 9% of respondents had been harassed in an off-Wiki location. This could be interpreted to include e-mail but Wikimail harassment has never actually been addressed by name in a survey. This makes it difficult to judge the extent of the problem or how widespread the problem is across language communities.

Anecdotal evidence for this report comes from replies to anti-harassment community

proposals and comments in Phabricator threads on the topic.

## Personal e-mail address exposed

Users have repeatedly complained that by disclosing their private e-mail address for use in Wikimail they open themselves to the risk of abuse by harassers sending unsolicited messages or hackers targeting their e-mail provider. This core problem has inspired a range of anti-harassment proposals over the past five years.

**Gradzeichen**: The project suffers strongly from people not asking questions, because they do not want to expose their e-mail in wikimail and people not answering questions sent by wikimail, because they do not want to expose their e-mail address.

**TheDJ**: Another solution however would be to simply NOT expose the email address in Wikimail. This has also been suggested in several places now, for various reasons (DMARC and privacy protection).

**Stryn**: If you want to send a private message (email), there is no reason why a receiver should get your email address, as it's private information.

**TBolliger_(WMF)**: As product manager for the WMF's Anti-Harassment Tools team I have created a project concept page to track this proposal.

Not everyone agrees that this needs to be addressed through new features. A few technically adept users have developed their own DIY workarounds.

#### Use a throwaway e-mail for everything

Some users minimize risk by using a disposable e-mail address that doesn't matter. They simply change the registered address whenever there's a problem. A few change their throwaway address regularly as a preventative measure.

**SMcCandlish**: Just register with an alternative or munged address.

**Platonides**: What I see many people doing here is to create an email address specific to wiki matters (eg. hotmail or gmail, it doesn't need to be a temporary one). This not only works for Special:EmailUser, but also allows safely interacting with other members of the community in mailing lists without revealing the other email address.

**Tgr**: Just set up a mailinator address or something. I can see how someone might want to send emails without exposing their real

address, but not receiving replies does not make any sense.

**Herostratus**: Every time I want to send a wikipedia email (except to a few people I trust), I have to go to a temporary disposable email site, get a temporary email, change my wikipedia email to that, send the email, then set my email back. It's a hassle and I seldom send emails because of that.

#### Automatically forward to a throwaway account

Other users handle obfuscation themselves by using their mail client. They use a secure account as a hub to forward to a throwaway account. Alternatively, they may forward notifications from a throwaway account to their main. This strategy requires vigilance to only reply from the throwaway account.

This is a more complex and sophisticated solution because it allows the registered address to remain secure and undisclosed while allowing the user to appear blasé about the disclosure of the (throwaway) address.

**Insertcleverphrasehere**: The obvious solution is just to have a second throw-away email address that you link to wikipedia that auto-redirects to your main email account that you check regularly. My email connected to wikipedia basically is just a redirect to my main email account, so when I get sent or send emails from wikipedia it goes through that account (and people see that account), and when I get replies they also arrive at my regularly checked address.

**Platonides**: Flexibility for where to send the different notifications can already be handled by using filters at your email account. Which is probably the right place. You only need to filter which kind of notification it is and depending on that forward to another email address.

**Tgr**: A simple workaround is to set up a mail filter to forward user mail to your secondary email account. Educating people about that seems like an easier path.

**A Den Jentyl Ettien Avel Dysklyver**: I already reply via a different email account to the one which receives emails.

#### Create e-mail filters for spammers

A few users didn't bother with obfuscation at all and advocated filtering unwanted messages with existing e-mail tools as if the messages were spam.

**CFCF**: It is very simple to set up a filter through your e-mail provider so that spam from certain addresses or containing certain phrases (e.g. "sent by User:Spammer") automatically goes to the spam folder. Getting around this filter is pretty easy if a user

registers new e-mail addresses to spam you, but you can have the exact same problem with users registering new WP-accounts to send you harassment.

**NickK**: If you are a harassment victim you can already set your mailbox to reject emails from certain people or at least discard it immediately. On the other hand, it is very easy to game this feature by simply setting a new account.

#### This isn't actually a problem

A few users denied the problem or argued that it was simply the nature of e-mail.

**Jeblad**: Don't use email unless you are willing to expose both the address and its content. Trying to obfuscate email addresses are simply stupid. If you want a secure communication channel use a really secure channel.

**Tetizeraz**: I don't see a second e-mail as a problem, but I consider it a convenience feature, not a security one. 2FA authentication, which has limited implementation, helps a lot more.

## Mute / e-mail blocking

The mute feature was proposed in 2016 to address the problem of Wikimail harassment. It was partially implemented over the following two years and remains the only anti-harassment proposal in this report to have been developed and launched.

The original proposal included a complex mix of permitted and blocked user lists involving group levels. The final implementation included only a simple blocking list, no list of permitted users and initially no group-level blocking. One group level prohibition was eventually added (new users) to guard against autoconfirmed sockpuppets.

**BethNaught**: The aim of this proposal is to make it so that editors can allow legitimate users to email them using the wiki email system while preventing abusive users from doing so.

**Tsoukali**: Allowing a user to choose who they want to receive emails from (or not) is a basic feature that any mailing service provides these days and it's time we caught up, as this would at least help users gain more control over what hits their inbox.

**Seraphimblade**: This is a great suggestion, and I think it would help to curb abuse perpetrated through Emailuser. Let's do this.

**Darkfrog24**: This idea is simple and practical and looks like it would be very effective.

**TBolliger**: Per feedback on wiki and discussions with the rest of the Anti-Harassment Tools team, we decided to change the default for new accounts to be to accept email from non-autoconfirmed email. So, no change in the current experience/functionality.

#### Dealing with sockpuppets

The mute feature was originally proposed to include access levels. This would have allowed users to restrict messages from certain groups. The first (and only) group to be prohibited was the undefined “brand new users.”

**MER-C**: Autoconfirmed is a trivial barrier for a dedicated sockpuppeteering harasser.

**Johnuniq**: A small number of editors are subject to long-term harassment. I know of cases where the just-released mute features would be insufficient since a harasser can create dozens of throw-away accounts and use them to annoy their target. How about adding an option to allow only notifications from users with a specified user access level.

**SPoore (WMF)**: This is one of the additional options being considered, especially for EmailUser mute. It is my favorite improvement that we’ve discussed so far because I think that it could significantly decrease the amount of throwaway accounts doing harassment by email.

**Huldra**: As one who has received hundreds of abusive mails through the wikipedia mail system, I would strongly encourage the implementation of this. It takes my harassers about 5 seconds to make a new user name.....banning any specific user name will slow them down...about 5 seconds.

**Huldra**: The above suggestion, of only allowing emails from, say users who have WP:EXTENDEDCONFIRMED (30 days/500 edits), is something I have wanted for years.

**Funcrunch**: It would be great for users to have an option to prevent anons from e-mailing them (requiring at least autoconfirmed level). Probably easier than implementing a blocklist of specific users too.

**KylieTastic**: I like the idea of this as I like to leave email on for private info that can’t/shouldn’t be added on wiki but it would be useful to be able to set to "autoconfirmed users only".

**LT910001**: Better than the binary all/none system we currently have.

#### Adding groups levels

More robust group level settings were planned, prototyped and even had a tentative launch date in 2017. The current system does not include this option (for unknown reasons).

**BethNaught**: Allow each individual user to choose which user access levels another person must have to send them email. For example, autoconfirmation could be required to prevent the use of throwaway accounts. In case of autoconfirmed sockpuppets, a higher level, like "extended confirmed" on English Wikipedia, could be required.

**TBolliger_(WMF)**: We also want to build the ability to set which user groups can send emails. Currently the user preference for allowing direct emails is a tickbox — on or off. I think it might work best as a dropdown with a few options. Autoconfirmed, extendedconfirmed and admins, or only admins

**BethNaught**: One thing to keep in mind: when configuring the list of user groups, I think it ought to be hierarchical (i.e. every member of one group is a member of the previous). This is for usability, so the end user has a clear pattern of increasing protection strength.

**TheDJ**: Make it possible to block mail from entire usergroups.

**WereSpielChequers**: Probably the only access levels that this needs to be settable for are confirmed/autoconfirmed and Extended confirmed.

**Nyttend**: I prefer the idea of a dropdown, as I'm seeing it described here; the more options the better, as long as it doesn't take an inappropriate amount of developer time and doesn't cause problems when it's in use.

**Johnuniq**: Thanks, this is essential and overdue. Only a handful of cases are known to have received extreme harassment but a healthy community must take steps such as these to prevent easy abuse. The dropdown list is necessary because some people will be happy with autoconfirmed but others will need a far higher hurdle.

**TBolliger_(WMF)**: We are planning to build the ability to control direct emails from user groups very soon, likely next month [Sept 2017].

#### Some groups shouldn't mute

A few users pointed out that some groups should be universally reachable.

**Tokyogirl79**: There might be some pushback about the idea of admins blocking anyone but autoconfirmed accounts and higher from e-mailing them, but otherwise this is very good and I know that several users would take advantage of this.

**Gianfranco**: I'm a bit against this feature: even if admins are not required to enable the Special:Emailuser, there might be selected groups of users that we expect to leave an open door for legitimate emails from anyone in case of particular needs (i.e. checkusers).

**Tryptofish**: I worry that users might block emails that they really need to receive

## Users cannot easily find the mute function

The Anti-Harassment Tools team has expressed concern about this problem in the past but it does not appear to be top-of-mind for users in discussions about Wikimail or the general mute feature implementation. The current approach has been to unify multiple mute lists in the preferences with the help of a new page called Special:Mute. This addresses the information architecture of the website but not the e-mail touchpoint itself.

#### Problems with plaintext design

The mute function is difficult to find in e-mail because the Wikimail footer is cluttered. Wikimail messages should be delivered in HTML format like other notifications. Plaintext e-mails with links are difficult to visually scan because the URLs are unwieldy. In general, each link should be placed on a line of its own for readability in plaintext.

The current Wikimail footer needs to be radically simplified, especially as it relates to the mobile user experience. The language has grown more complex over time and doesn't match the documentation found in MediaWiki:Emailuserfooter.

The last paragraph should be a clear call to action: "Mute this user: username", rather than a confusing invitation to manage the other user's e-mail preferences. Information on privacy, security and abuse should be available on the mute page and from the initial Wikimail page. Legalese should be minimized. It doesn't appear that legal was involved in the original drafting process (on Phabricator) which simplifies the necessary editing.

General notification e-mails are already provided in HTML format but the footer could still be simplified. New designs should be tested, iterated and confirmed with analytics.

## Wikimail Proposals

Along with the mute feature proposed in 2016, the array of anti-harassment features proposed by the community over the past five years collectively reveal a defensive mindset. These improvements would help victims to evade Wikimail harassment but would do nothing to address the abusive behavior itself.

### Secondary e-mail address

The goal of this Wikimail proposal would be to avoid disclosing a secure e-mail address to harassers by assigning a second, less critical, address for use in messages.

This feature envisions a primary channel for official communications and a secondary channel for Wikimail with the expectation that the secondary e-mail address will eventually become compromised. If so, the user could simply change the Wikimail address while keeping their secure e-mail address for all other communications with Wikimedia.

**Dthomsen8**: An excellent idea. I can use a public ID for Wikipedia users, and a hidden ID for password problems and quiet communications with Administrators.

**TheDJ**: I'd also appreciate the option to use a different mail address for wikimail than the primary one coupled to my account.

**Vachovec1**: The underlying problem (exposing your e-mail address when answering to a wikimail) needs addressing. This seems like a decent solution.

**Raystorm**: The root problem is definitely an issue, would be good to fix it.

**Davey2010**: Support giving editors the option of adding a second email address but oppose making it a required thing.

**Addshore**: I would like to be able to attach multiple email addresses to my account.

**Mailer Diablo**: Should have been implemented long ago.

#### Robust account recovery

A secondary motivation for this feature is robust account recovery. Having two confirmed addresses would make it less likely for a user to be locked out of their account.

**Quiddity**: A user might lose access to one of their accounts for legitimate reasons: they stopped using that ISP, or attending that college, or working at that company.

**Gradzeichen**: With only one mail address, a user who loses this mail address, has no way to recover a password, as server admins have no way to identify the user. With two addresses, there is at least in principle a way to manually verify the user's identity.

**Platonides**: I see some benefit in storing multiple email addresses per account, mainly for the case when an email is no longer available.

#### Usability concerns

Many users sympathize with the core underlying problem regarding e-mail harassment and Wikimail but oppose this particular solution. The opposition stems from usability concerns about the implementation of preferences and the increase in complexity.

**TBolliger_(WMF)**: The proposed solution seems over-complicated, but the root problem of disclosing email addresses is definitely a problem worth looking into.

**Nemo_bis**: Just noting that there is no way we could afford a usability debacle such as this absurdly complicated interface.

**Platonides**: The proposal as stated is a preference nightmare.

**NickK**: Support that we should solve the problem of single email for everything (a more secure one is needed for password recovery than for answering spammers) but I oppose the proposed solution as overly complex to use and to manage.

**Murbaut**: This is good, but how if newbie can learn, it may be confused?

**Demian**: No confusing secondary email configuration. "Which one I use for what feature?" The name "Auxiliary" would sound alien for the everyday user.

**Dinoguy1000**: The level of configurability suggested seems like gross overengineering; the baseline should be that one email address serves as the recovery address, while the other serves to receive communications. Any level of configurability beyond that should only be undertaken with care, probably with a clear demonstration of need.

**TBolliger**: This probably shouldn't be present on account registration, and we'll need a way to keep the email preferences somewhat sane.

**Kakurady**: The UI looks unwieldy. Asking for two email addresses on registration (even though both are optional) is a cognitive burden for editors registering a new account. Perhaps call the auxiliary "account recovery email"

instead, and only gently prompt after a few days/edits.

**Anomie**: Oppose the overcomplicated solution described here. Neutral on the general idea of having separate recovery and EmailUser reply-to addresses.

**Poslovitch**: Support for the idea. Oppose due to its complexity to "manage" for a user who is not aware of all of this.

## E-mail aliases

Another proposed solution for the e-mail disclosure problem would be to eliminate personal e-mail addresses from Wikimail delivery altogether. An alternate e-mail alias would automatically be generated for each message.

#### One-way e-mail system

There are actually two different versions of this proposal. The original version described a delivery-only system similar to existing Wikimail but without disclosing the sender's address. Each e-mail would originate from a no-reply Wikimedia alias. The only way to respond would be to initiate a new message onsite using a link from the e-mail.

In this model, the e-mail would serve as a notification for a message but not as a two-way communication channel. This is essentially how the talk page notification model already works. GitHub and eBay use similar notifications for messaging.

**Nemo_bis**: So the real summary of this task is "send private messages to a user via a special page while only allowing replies from the same special page."

**Mattflaschen-WMF**: The simplest solution is to implement this task for everyone, and make it easy (e.g. a reply link) to reply by EmailUser.

**Bawolff**: If we do something like this, I think it would make sense to do foo@wikipedia.invalid so it's clearly a non-real email address.

#### Encouraging harassment

Many people recognized the potential for abuse in a system that anonymizes the sender of a Wikimail message. This would prevent recipients from blocking the address of known offenders and embolden the attackers.

**Aracali**: Wouldn't it work both ways and so enable anonymous harassing? Do not use email if you don't want your email address to be disclosed.

**NickK**: Sending an email and knowing that a person will not be able to answer it has a huge potential for bad-faith uses.

**Gianfranco**: I'm afraid this would encourage harassment. Currently you can sometimes find someone who insults you, threatens you or otherwise disturbs you, even if the interface tells them that their email address will be visible and their address has been confirmed. If you tell them they would be hidden, I'd bet they would feel more comfortable in acting badly.

**Hedwig in Washington**: I understand the idea but don't think enabling anon harassment is the answer. Pain in the ass, tho.

**Yann**: Might help harassment instead.

**Rschen7754**: Too easily abused.

#### Requires a blocking function

At the time of this proposal the mute feature for Wikimail did not yet exist and many users felt that such a mechanism would be essential. Mute functionality became available in 2017.

**KylieTastic**: This would just allow harassers to spam you emails that you could not easily block as it would all just be email from wikimedia. This could only work if we also had the "Allow users to restrict who can send them email" blocklist option as well.

**Gestrid**: I would support this if there is also a sure way on the Wikimedia side of things (as opposed to the email client side) to block certain users from sending you emails. As others have noted above, this could provide spammers with extra security as well, because their email addresses would be anonymous, too.

**Ryan Kaldari (WMF)**: Agree that this has the potential to embolden harassers and it should probably only be implemented if we also have some sort of blocklisting feature.

#### Breaks e-mail

Another argument against the one-way e-mail system is that it breaks the model of e-mail communication and prevents threading of messages.

**Nemo_bis**: This seems a very broken way to use email, but I understand that many people don't care about having tidy mailboxes where threading works and so on.

**Nemo_bis**: Why send an email if you don't want to see replies?

#### Forwarding alias system

The Anti-Harassment Tools team appears to have ignored or misinterpreted the original one-way e-mail proposal in favor of a more robust approach. This system would preserve the two-way nature of e-mail messaging by routing all replies through Wikimedia servers and generating an e-mail alias for both the sender and the recipient.

This is how projects like Craigslist maintain e-mail anonymity but such a system has never been formally proposed for Wikimedia.

**TBolliger_(WMF)**: If we decide to move forward with 2-way email relay we will probably begin researching other existing 2-way relay systems, such as Craigslist.

**Dolotta**: In my online auction days my e-mails with the other side of the transaction went to a randomly generated e-mail address connected with the person's account. Something on the order of random code@onlinecompany.com.

**Demian**: A common solution to this problem is to give a wiki email address to users (such as username@en.wikipedia.org), and forward/proxy emails to the private address. Might complicate configuring the server a bit.

**Mattflaschen-WMF**: It might be useful to allow direct replies while hiding the email addresses; that's doable, but a little more complicated.

**Jerodlycett**: This would simply open up people to being harassed anonymously. I could support it if they used a bounce email instead, something like mer-c@bounce.en.wikipedia.org which would bounce it to your actual email address.

**A Den Jentyl Ettien Avel Dysklyver**: The basic idea of having a "Dysklyver @ editor-en-wikipedia.org" email address to use instead of my normal email would be good.

**TheDJ**: Just give everyone their own temporary email alias every time a message is sent.

#### Bureaucratic / legal opposition

There may be policy concerns about the use of an e-mail relay system.

**TBolliger_(WMF)**: I've reached out to the WMF legal department to review this concept at a high level before proceeding any further [as of 2017].

**Gianfranco**: At a legal level, this would bring the Foundation to be responsible of this correspondence, and any time a user should need to act against harassers, or police legitimately requires quick help, WMF would

be bureaucratically called each time to reveal the hidden address, resulting in a useless heavy complication

**Tiggerjay**: Using an anon service would open up WMF/ORTIS to increased workloads by becoming the proxy for abuse and complaints. Currently abuse can be deferred off to the email provider, but once WMF gets in the middle of it, we now have to play middle man.

**Gradzeichen**: Wikipedia is supposed to be an international project. By the legislation of one or more states in the world forwarding mail may be considered offering a mail service, which by some legislation requires the mail provider to verify the mail user by the user providing a telephone number or a surface mail address.

**Platonides**: The problem with a forwarder is that when people start eg. sending emails with viruses attached (or simply spam) to @private.wikipedia.org addresses, they will go out from WMF IPs, which are then categorised as a source of virus/spam.

#### Implementation questions

The primary advocate for a competing proposal (secondary e-mail) raised several questions about the implementation of this relay system:

- ➔ Will you be able to answer an email only once, twice, or send an unlimited number of answers?
- ➔ Will you only be able to answer from the address MediaWiki sent the message to, from a number of known addresses or from any address?
- ➔ How long will you be able to send the answer(s)? an hour, a day, a week, a month, a year, ten years?
- ➔ If one party gets blocked in a single wiki project (i.e. for email spamming) can you still answer with this system? Can the blocked party contest the block by mail?
- ➔ The system can only be used while the wiki servers are running and online and not if the WMF servers are (temporarily) blocked themselves in a region of the world.

### Wiki messaging system

Several users mentioned their preference for an on-wiki messaging system as an alternative solution to the Wikimail harassment problem. Like the e-mail relay concept, this has never been formalized as a community proposal.

#### Structured discussions

Flow, also known as structured discussions, was a discussion and collaboration system for

Wikimedia projects. The system featured forum-style group discussion tools which were ultimately deemed unsuitable as a replacement for talk pages but might be a better fit for Wikimail communication.

**Max Semenik**: We could use Structured Discussions for private messaging. You would get a nice interface to follow threads, built-in customized messaging (including an option to not receive emails) and most of the code is already here.

**Boing! said Zebedee**: A private messaging system that does not use email addresses would help greatly with the problem.

**Izno**: That proposal seems like a similar interface to what reddit currently does with private messages, which isn't crazy to me.

**Jeblad**: I would propose that all interactions with other users goes on a separate thread, where some (all) interactions are private and anonymous by default. When a user writes a private message only a transcript is sent to the recipient, and both must agree on letting the thread be non-anonymous or non-private. Yes this can be implemented as part of the Flow-system.

**enL3X1**: PMing on the IRC doesn't fulfill the needs.

#### Monitoring of abusive content

An on-wiki messaging system would allow administrators to more easily document abusive activity reported by users or automatically flagged by bots.

**Izno**: Implementing a wikitext (or Flow) based private message system seems like a good idea and could still be viewable by administrators.

#### Flagging / moderation

Wikimail harassment is not easy to flag or report. Even muting a user requires a number of steps beyond the effort required on platforms like Twitter.

**John Broughton**: Is there some way to report the email as harassment? If not, there should be. Wouldn't it be a good idea to add at the end of every email? ["If you want to report this email as inappropriate, forward it to bademail@wikimedia.org, with a brief explanation."]

Wikimedia researcher Claudia Lo analyzed reporting systems on English Wikipedia for the Community Health Initiative in a 2018 report. She also wrote a competitive analysis of peer-dependent reporting systems such as Reddit and Facebook Groups the following

year. Both papers should inform future anti-harassment efforts for Wikimail.

#### Limits of automatic filters

Algorithmic approaches to harassment detection have improved over the years, no doubt aided by the arms race in spam filter development. This type of detection is a common subject in academic research but still faces important limitations.

In their 2016 project Understanding Personal Attacks on Wikipedia, authors Wulczyn and Thain developed a tool called Wikidetox to detect personal attacks on Wikipedia. They built a classifier that analyzes talk page comments and outputs the probability that a comment contains a personal attack.

Researchers have shown that automatic detection efforts are easily outmaneuvered by subtly modifying an otherwise highly toxic phrase in such a way that an automated system will assign it a significantly lower toxicity score.

Another researcher noted that too many spelling errors rendered the sentiment features of their detection model ineffective.

## Identifying harassment

The Wiki messaging system concept outlined in the previous section would open the door to an array of proactive measures against Wikimail harassment. Even in the absence of such a complex solution it might be helpful to brainstorm other proactive measures to identify and deal with harassment as a counterpoint to the status quo.

#### Policy limitations

One of the most striking aspects of Wikimail abuse is the relative lack of detection systems. There appears to be no way to identify what types of messages are passing through the system. This makes it impossible to filter out harassment, extortion and other abusive content or even to understand the scope of the problem.

But this is a policy choice more than a technical limitation. The current policy states:

Emails sent using Wikimedia's mail facility are private. Their contents cannot be read by administrators or anyone else, even to check for appropriateness, so only limited help is possible if the feature is being abused.

Any message originating from Wikimedia servers could theoretically be accessible, if not by humans (due to limitations of scale) then at least by the same types of algorithms that identify spam. Ebay uses this type of system to detect violations of their terms of service. The decision not to intervene is a policy choice

which avoids some of the bureaucratic entanglements identified earlier in this report.

This is exacerbated by the policy decision prohibiting victims of harassment from disclosing the contents of the harassing messages they receive:

*Do not post the email on-wiki without permission. You should not post the email itself without permission (although you can describe briefly in summary what it contains or shows). This is partly due to copyright concerns, given that Wikipedia pages can be re-used by anyone.*

Addressing Wikimail harassment will require a combination of both policy innovation and technical innovation. At a minimum there should be a way to document this type of abuse without running afoul of copyright(!) concerns.

#### Active vs passive solutions

Filing complaints publicly has its own set of downsides. Public reports of harassment can be viewed as aggressive in some cultures. As mentioned earlier, this has been a longstanding issue for French Wikipedia.

**Nattes**: Without documentation it is difficult to address problems. And we should also be able to notify the relevant communities regularly with updates on the types of behaviors that create problems. Any attempt to do this locally on French wiki has been deleted. Public reporting is not possible because it creates backlash and is seen as abusive itself.

Muting a user is a private alternative to publicly reporting abuse or attempting to have the user blocked or banned. It is essentially a passive response. Even the error message shown to muted users is less aggressive than on other social media platforms. It obscures the fact that the user has even been muted.

Reporting and flagging are baseline requirements for most modern platforms where users interact but the reporting process on Wikipedia is not straightforward for harassment complaints. Each wiki has its own system for dealing with this issue. Claudia Lo's 2018 analysis on reporting provides an overview of these formal and informal channels.

The ongoing Reporting System project is beyond the scope of this analysis but flagging could provide a middle ground between active and passive responses to harassment.

#### Metadata analysis

The Anti-Harassment Tools team published a wiki page on the User Mute feature that offered an interesting hypothesis:

*If making the mute features more visible increased the frequency of the mute lists being used, that would be a good indicator that there is more on-wiki harassment than we thought.*

They published initial results in 2019 but a targeted analysis of mute preferences and Wikimail usage might also help to inform anti-harassment detection, especially when combined with the User Interaction Timeline project.

- ➔ Number of users who mute a particular user
- ➔ Timestamp clusters of muting
- ➔ Mute events coinciding with other abuse
- ➔ Flagging as a precursor to muting
- ➔ Ratio of messages between any two users

This potentially intersects with the AHT team's work on Wikihounding from 2018. They theorized that harassment could be analyzed by time, frequency and location before looking at context. This is relevant to Wikimail harassment because of the policy restrictions that prohibit any analysis of message content.

**Gestrid**: When an email is sent using Special:EmailUser, the fact that an email was sent (nothing else) is recorded in a log somewhere. We could somehow use that log to monitor emails incoming to one of these anonymous email addresses.

The Wikihounding report also mentioned a 30 page document (with notes) on canonical cases assembled for the Support and Safety Team.

#### Acknowledging harassment

In their 2017 paper Classification and Its Consequences for Online Harassment, researchers from the University of Michigan School of Information examined policies related to harassment and abuse. They shared three insights relevant to Wikimail harassment:

- ➔ For victims, labeling experiences as "online harassment" provides powerful validation of their experiences.
- ➔ For bystanders, labeling abusive behaviors enables bystanders to grasp the scope of this problem.
- ➔ For online spaces, visibly labeling harassment as unacceptable is critical for surfacing norms and expectations around appropriate user behavior.

Like many of the studies referenced in this report, the 2017 analysis was somewhat limited by its focus on an American audience. Other cultures may prioritize competing norms which influence their approach to Wikimail harassment. This type of abuse may not be

equally distributed across wikis or even among editors on any particular wiki.

The Wikimedia community reacts to online harassment as if it were an unavoidable force of nature. Beyond the technical solutions mentioned in this report, future policies should be geared toward acknowledging and validating instances of harassment and fostering norms that discourage abuse.

The goal should be an affirmative end to Wikimail harassment.

## Next steps

Future editor and administrator surveys should include a section focused on Wikimail harassment, e-mail mute and other proposed solutions.

It's also important to understand the extent of the problem across individual language communities, starting with analytics and moving to surveys and semi-structured interviews with administrators and editors.