Findings Report/**Incident Reporting**

Design Researcher // Daisy Chen
Project Requestors // Aishwarya Vardhana, Madalina Ana

Background

The Trust & Safety team is hoping to build an Incident Reporting system that can address the pain points of both the targets of harassment and responders on the wikis. Understanding the harassment ecosystem on the platform and current reporting workflows are crucial to identifying these pain points and needs. Conducting research with Admins on a couple pilot wikis will be an important first step, as well as allowing for an opportunity to socialize the concept of an incident reporting system within these communities.

Methodology

Recruitment during this time period has been challenging due to other prioritized community outreach efforts at the Foundation. The goal was to conduct as many interviews as possible with **Admins and Editors from Indonesian and Korean** wikis.

During these interviews, we asked participants to:

- narrate their experiences with responding to harassment on the wikis
- what steps they took (communication tools, technical tools, etc.) to do this work
- discuss, if they felt comfortable doing so, scenarios when they themselves were harassed on wiki and what transpired

Participant Profiles

Ultimately, we conducted:

- 3 interviews with Admins on Indonesian wiki
    - Id.wiki Admin, Bureaucrat, and member of Doxxing Task Force
    - Id.wiki Contributor, Patroller, and Admin
    - Id.wiki Admin
- 2 written response correspondences with Editors on Indonesian/Banjar wiki
    - Bjn.wiki Contributor
    - Id + bjn.wiki Contributor
- 1 interview with an Admin on Korean wiki
    - Kr.wiki Admin, Oversight committee, worked on Universal Code of Conduct

Findings

"contributing to Wikipedia means being prepared to be the target of harassment, insults, and so on"

Overall, the situation with harassment reporting is haphazard and unofficial.

Policies: unofficial; while there are adjacent policies like 'Resources for Editor privacy' or the Universal Code of Conduct, there are no official policies for harassment specifically, official ramifications, etc.

Reporting: there is no universal process and the existing processes are unclear, especially to newer users.

Status: the understanding around status of the reporting is also ad hoc.

Processing/review: Admins and other responders contend technically with tracking multiple channels of communication that could contain harassment reports, not to mention the various means through which harassment could occur.

Structurally, not all Admins and similar roles are comfortable dealing with harassment reports, and therefore lowers the number of individuals who can review and take appropriate actions. Smaller wikis face the additional hurdles of smaller #s overall of Admins/responders, and sometimes language barrier/translation issues.

Escalation: Responders do not always have clear paths for escalation in various cases. Within the wiki ecosystem, protective actions cannot always be taken in a timely manner. Outside the ecosystem, often it is unclear what can be done and when there are available options, protective actions need outside support (Foundation, affiliates, external enforcement).

Harassment

Frequency of Harassment

kr.wiki

Targeted harassment is not a rampant issue on Korean Wikipedia, but incidents are reportedly occurring at a rate of approximately 1-2x per week, mostly against Admins and popular Editors.

id/bjn.wiki

Targeted harassment is not a rampant issue on Indonesian Wikipedia; however, specific events and categories of content can spur a flurry of activity ranging from insults to physical stalking.

External force accelerating need: for id.wiki participants in particular, there is higher demand for this reporting system due to the upcoming 2024 elections.

Types/severity of Harassment

kr.wiki

1. Multiple spam accounts are created to send harassing messages to the target
2. Off-wiki wiki-related harassment is not a common occurrence
3. Participant's personal experience with harassment:
    a. Experienced retaliatory spam/threatening messages after blocking a user

id/bjn.wiki

- Personal insults/attacks on talk pages or village pump
- Doxxing (specifically mentioned x3)
- Physical abuse (stalking personal home, in-person harassment)
- Retaliatory harassment after Admins enforce rules/behaviors, and after Editors make contrary or controversial edits
- Off-wiki wiki-related harassment is comparatively more of a concern
- Participants' personal experience with harassment:
    - Almost all participants experienced retaliatory harassment
    - One participant mentioned feeling the need to preventatively avoid certain confrontations/topics and wishing there were a way to protect herself and others in a more formal and technical manner, especially in response to observing other users being doxxed and stalked

Other factors affecting type/severity

- Cultures and countries' norms/politics/laws on online interactions (harassment, doxxing, etc.) are varied
- Political concerns and 'temperature' differ (e.g. Editors who contribute to the page on the Communist part can be ID'd, doxxed, stalked, persecuted)
- What escalation options look like differ (reporting to an authority can be helpful or deleterious, depending on the situation)
- Related to official accountability; code of conduct and policies only apply to on-wiki spaces; off-wiki harassment on social media or in real life can be difficult to address

Outcomes of Harassment

- Though most participants didn't think that many people leave the platform due to harassment, it has still been noted anecdotally:
    - One moderator who hosted a wiki meeting who was attacked for mispronouncing a name was never active again
    - *"I didn't know the people who were being abused until I left the platform."*
- Harassment can cause a chilling effect; accountability is not always applied in a consistent fashion and within a reasonable amount of time. Additionally, escalations can be a free-for-all scenario especially without Foundation and/or outside institutional support.

Existing Workflows for Responders

kr.wiki

- **Initial report:** Difficult to catch all incident reports when they occur; emailed reports go to spam or are buried among other emails/notifications/spam messages
- **Tracking**: On-wiki recordkeeping and interaction/discussion of incidents exist
- **Review**: Debate for rights removal, (global) bans. Review of account logs to identify spam accounts
- **Escalation**: has required Foundation resources. Communication with staff can be difficult without a direct line to Trust & Safety and Legal teams in particular

id/bjn.wiki

1. **Initial report**: Contact can occur via many different channels, including but not limited to personal email/call, WMID contact email, Telegram/Whatsapp group chat or DM, Admin noticeboard.
2. **Tracking**: Unclear and unofficial.
3. **Review**: Typically initial warnings occur, and short blocks can be instituted after a clear infraction. After investigation and discussion, blocks and global bans can be enacted as a group decision; individual Admin actions are not encouraged. Responders can also notify of reports at the admin noticeboard, and wait for involvement/review/further action
4. **Escalation**: WMID has offered support in the past. More institutional power/support from the Foundation/affiliates/other resources can be helpful.
5. **Prevention/protection**: Admins have assisted targets in changing their usernames for safety purposes (especially when doxxing and stalking have occurred)

Concerns about existing processes

Shared

- Newer users and even possibly more experienced users don't know where to make the harassment complaint
- No official harassment policy
- Not enough external support for escalations
    a. *"Contacted Wikimedia Foundation's Trust & Safety team but they did not provide any specific information" - Id.wiki participant*

kr.wiki

- Often, collecting the information to conduct a thorough investigation can take many months to finalize and a much shorter timeline should be expected (~1 week)

id/bjn.wiki

- No specific places to discuss the report (although the Admin chat group feels supportive and is one shielded location)
- Not all admins are involved and responsive on the harassment reports issue
- Blocking is a public action; any user can see which Admin did the blocking (regardless of the consensus-based discussion that is involved in the review process)
    a. Retaliatory harassment is a main category of harassment on wikis; Admins feel the need to consider their own safety

Finally, a note that perceptions and impacts of harassment (frequency, severity, direct/indirect, etc.) all vary depending on the individual(s) involved. To further expound on this point, one participant provided many different anecdotes and salient reactions to harassment incidents (*observed, experienced, and anticipated*) in wiki spaces. Most participants acknowledged harassment as a very important issue, but even though they did not provide many personal opinions nor express as high a level of personal concern; it does indicate that the nuances of intersectionality are important ways to make our approach to creating this reporting system more safe, supportive and inclusive.

Workflows for Targets of Harassment

Shared

- No centralized, clear reporting mechanism for users
- The existing reporting methods are varied and scattered; contact is between individuals and/or a report of some kind is made to an email alias like emergencies@ for an unofficial review group of responders
- Most users would not know where to look/who to ask or trust to make a report

id/bjn.wiki

- Telegram, WhatsApp, and direct message/email/calls are the most common reporting channels
    - Reporting can also occur at the Village Pump
    - Most reports ask for mediation/resolution, not always for a direct user block

When responders are harassed, they:

- Don't respond

- Alert other Admins/Responders. Often, there is a warning, and a block if harassment continues. Stewards may be asked to issue a global block.
- Take preemptive defensive actions (most of which have direct and indirect deleterious effects), such as:
    - Request their uploads/contributions be deleted
    - Take extra precautions around edit/responder activities
    - Avoid participating in edit conflict/war
    - Change username *(this can cause issues for other responders who may need to see history for context)*

Recommendations

Essential elements of report submission/review

> *"make [the Incident Reporting system] simple. allow targets to quickly link a request to the harassment that occurred (many do not have the time to find the details to add to a report"*

- Username/individual doing the harassment
- Short description detailing reason for the report (a few examples could be presented via click-to-populate quick-select options, and one could also opt to add manually)
- Link(s) or photo(s) of the pages containing the harassment/threat
- Ability to select whether target would prefer to remain anonymous (i.e. Responder can/not get involved on-wiki with report-related communications, and therefore likely/avoid identify the target)
- Feature that accommodates users speaking different languages
- Connected communications channels to appropriate partners (Foundation Trust & Safety and Legal teams, Affiliate/Wikimedia chapter members, external organization contacts)

The first 3 bullet points could potentially be built-in/pre-populated via a highlight to report mechanism, similar to something like what the Android app team did with Share a Fact.

A report submission button would then generate an automated email from the system indicating receipt and link to check for status updates.

System-adjacent recommendations

- Centralized system to avoid Responders having multiple channels to monitor
- Chatbot to assist and direct users on-wiki
- Create an Admin noticeboard-style system where Responders can review pending requests
- More immediate (global) actions should be available to deploy, with at least temporary effect; for example, disallowing editing until the decision is made and it can be lifted or continued depending on the decision

Harassment-related recommendations

- Establish and disseminate official harassment policy/ies for wikis based on a document like the Universal Code of Conduct
    - When/whom to report to, appropriate consequences/actions following harassment reports, when to deescalate/involve Foundation/chapter/affiliate/legal representatives
    - *"official process may inherently reduce incidents" - id.wiki participant*
- Form an ArbCom of sorts on each wiki that is directly responsible for responding to investigating, and escalating harassment reports
    - Doxxing task force on id.wiki worked well; members delineated responsibilities of the work, and established relationships with external resources like SafeNet to establish better procedures around doxxing/stalking-type harassment

Theme: balance of safety / transparency and the potential creation of a 'privacy' layer

- Protect identity of targets when requested/needed so they have peace of mind
    - *"easily accessed by everyone [... but] report is not visible to everyone" - id.wiki participant*
    - *"be as transparent as possible while protecting anonymity" - kr.wiki participant*
- Allow for a location in the system where private communication (Responders with one another to discuss a case, Responder with target, etc.) can occur; alleviate more of the multi-channel status quo of harassment report-related communication
    - Id.wiki participants have considered the creation of a private wiki for this purpose (discussing critical issues, resolving reports). When the need arises, there can still be a record of communications, but it is not publicly available.
- Change policy on what individuals can see contributions and blocks
- Username changes and 'puppet' accounts: some participants have mentioned changing their username, assisting others with username changes, and/or the desire to create and use puppet accounts. However, especially with using puppet

accounts, they must be publicly announced/linked to the main accounts so in effect do not serve the ‘privacy’ use case.

- Participants have suggested a private layer wherein puppet accounts for certain users can be made with admin permissions, then approved and logged, but not publicly so. This can allow certain users (those with a history of being targeted, for example) to avoid being traced when editing on controversial topics
- Similarly, an ability to hide username change history when approved/appropriate