Reporting systems on English Wikipedia

A partial snapshot of current systems

Claudia Lo, Design Researcher, Anti-Harassment Tools Team

For the Community Health Initiative, November 2018

Contents

Summary

When we think of reporting systems, we tend to focus on the part that allows editors to make reports to some other party. While this is definitely a focus of any reporting system, it is not the only part. A reporting system is a loose assembly of related systems, coming together to convey information and facilitate action, with the primary purpose of allowing editors to report breaches of policy to a trusted party who can hopefully resolve the issue. Therefore, there are many areas for possible improvement outside of just reporting or flagging mechanisms.

Who is involved?

For reporting systems, we are concerned with three main groups of users (collectively, "involved users"):

- Moderators, who receive and handle reports.
- Reporters, who bring reports to moderators.
- Accused users, against whom reports are made.

None of these three are mutually exclusive, and we should expect some degree of overlap and fluidity between these categories.

There is a silent fourth group of observers: community members who may not engage in the reporting system, but whose feelings of safety, privacy, and belonging will be affected by the public-facing aspects of the entire reporting system.

Formal and informal systems

Current reporting systems on Wikimedia projects can be categorized as formal or informal. Formal systems are codified and supported by policy or code or both, designed to facilitate reporting. Examples of formal reporting systems include noticeboards or ArbCom.

Informal systems are various networks of communication, relationships, code and policy that can be repurposed for reporting purposes. Examples of informal systems include private off-wiki correspondence, or use of project-affiliated social spaces (e.g. IRC) to report misconduct.

Formal systems can be useful because they provide an obvious and structured way to bring a case to a moderator's attention. Ideally, formal reporting systems ease the burden of reporting cases while upholding important community values when dealing with reports, such as maintaining good faith for involved parties and committing to transparency in the way reports are handled and documented.

However, these systems can fail because of their rigidity and slower speed, making them unsuitable for emergencies, acute abuse, or particularly complex cases. The specific way in which formal reporting systems are set up for a given project may also be open for abuse or manipulation.

Informal systems are much faster and more flexible as they leverage existing relationships to reach moderators who could act upon an informal report. Ideally, informal systems can allow reporters to discreetly bring cases to moderator attention, especially for more complex cases that require speed and discretion. Because they are not bound by the structures of formal systems, they can accommodate more edge cases, involve different methods of mediation, and come to more complex resolutions.

However, informal systems are very opaque to those without lots of knowledge or deep involvement in the Wiki community. Their existence can also make both involved groups and observers uneasy, since they definitionally exist outside of the "official" system, and therefore are assumed to be less legitimate and not governed by the same community values.

On English Wikipedia, our current target of study, all easily-found reporting systems and most of the formal reporting system is public; the majority of public reporting spaces require the reporter to notify the accused user as a condition of use, which could potentially deter reporters from bringing new cases. Informal reporting systems are de facto private channels, yet their opacity means they are accessible only to experienced editors.

Formal and informal systems are not mutually exclusive, and ideally complement each others' strengths and weaknesses. The balance of formal-to-informal reporting system use will differ by project, based on their policy, available moderator labor, and project values. Because misconduct is a social issue, informal systems are unlikely to ever disappear. Formalizing previously informal networks can also be costly, causing these reporting paths to lose much of the flexibility and speed that makes them so advantageous in the first place.

In short, we must be careful and deliberate if we want to take an aspect currently handled primarily in an informal system (private reports) and turn it into a formal system. Any reporting system succeeds only with the trust and buy-in of everyone involved. An example of a system that has lost the trust of many participants is AN/I, on English Wikipedia.

Other considerations

In addition to these reporting systems, there are three major considerations at play: policy, values and labour.

Policies are the social and operation policies put into place in any given project, such as English Wikipedia's “standard” 31-hour block length used for a variety of infractions. These are the guidelines, best practices, procedures, and rules governing engagement with reporting systems, both formal and informal, on a project.

Values are the moral and social values prized by a given community, which shapes the design of reporting systems. Additionally, these social values provide a frame by which the community understands the authority and legitimacy of a reporting system in all aspects, from its perceived efficacy, to the validity of the outcomes it generates.

Lastly, labour concerns both the question of who performs the work necessary to keep these systems running, as well as the conditions under which that labour is performed. Major issues include providing proper support and training, dealing with volunteer retention, and managing burnout and other negative impacts of performing this dispute resolution work.

Takeaways

- The actual mechanism of bringing reports to the attention of moderators is only one of many areas for potential improvement.
- Any new system we design must be consistent with existing community values.
- Be mindful of the ambiguities around the term “harassment”, and difficulties in translating the term.
- Not all reports need to result in a block or other public sanctions. Not all reporters are the target of misconduct; reports can be a way to log bad behavior.

Who might use these systems?

For the purposes of reporting system work, we have identified three main categories of users who might engage with any potential reporting system.

Moderators: these are users who receive, deliberate and act upon reports. Moderators may or may not be administrators or other user groups, and this position is not always a named one. An administrator who receives and reads noticeboard reports, who concludes them by administering formal sanctions such as blocks, is a moderator. Equally, an influential editor with no user rights beyond that of any other editor may act as a moderator, receiving informal reports and acting upon them by engaging in conversation to reach a conclusion.

Reporters: these are users who bring reports to moderators. These users can be, but are not necessarily, the direct targets of misconduct or abuse; they could be bystanders, or be accused of misconduct themselves.

Accused users: these are users who have been named as engaging in misconduct or abusive behaviours.

None of these groups are mutually exclusive. A moderator might be a target of harassment, making them a reporter, yet be accused of misconduct and abuse of their power by others. Reporters bringing a case to a formal noticeboard may have their behavior scrutinized, in the process becoming accused of misconduct themselves. Part of the difficulty of designing a reporting system will be accounting for this overlap between involved users.

These groups do not exist in a one-to-one proportion. Any given report may be handled by any number of moderators. The number of moderators will be far smaller than that of accused users or reporters. One reporter may report multiple users, and an accused user may be reported by multiple people. As cases develop, other users may be pulled into the case to support, comment on, or criticize the other involved users, in which case we could call them “peripherally involved users”: not involved in bringing the case to the system, but nevertheless now engaged.

Formal reporting systems

Summary

A formal reporting system is a codified, documented structure or set of structures, accompanied by documentation (public or private) and supported with purpose-made products, pathways and spaces, whose overall purpose is to assist users who need to make and interact with reports. Formal systems do not require reporters to have existing familiarity with them, although familiarity will naturally help when navigating formal reporting systems.

The strengths of formal reporting systems include their purpose-designed nature, such that they ideally lower the barrier of participation for engaging in the reporting system, and ease the process of reporting generally. Their association with clear documentation both allows moderators to better keep track of and deal with reports, and provides a way to set precedents and leave traces in cases where a pattern of misconduct emerges. Formal reporting systems also lend reports legitimacy by codifying and claiming socially-approved values around reporting misconduct and abuse.

However, formal reporting systems necessarily have to set boundaries about what kind of behaviour qualifies as “report-worthy”, which means that they will not capture all dimensions of misconduct; the alternative is a system that defines misconduct so broadly as to lose all meaning. Formal reporting systems can also be slow as all reports ostensibly must or should go through the same process, which may or may not be suitable for a given case. Lastly, a formal reporting system must be trusted by all parties engaging in it in order to receive the legitimacy it needs to act: a system that is not trusted, whether it be by reporters, accused users, or moderators, is a system that will not function.

It is important to note that a visible and trusted formal reporting system will, by its existence, tell the community and any other onlookers whose safety and privacy is prioritized, and in what ways.

Existing examples

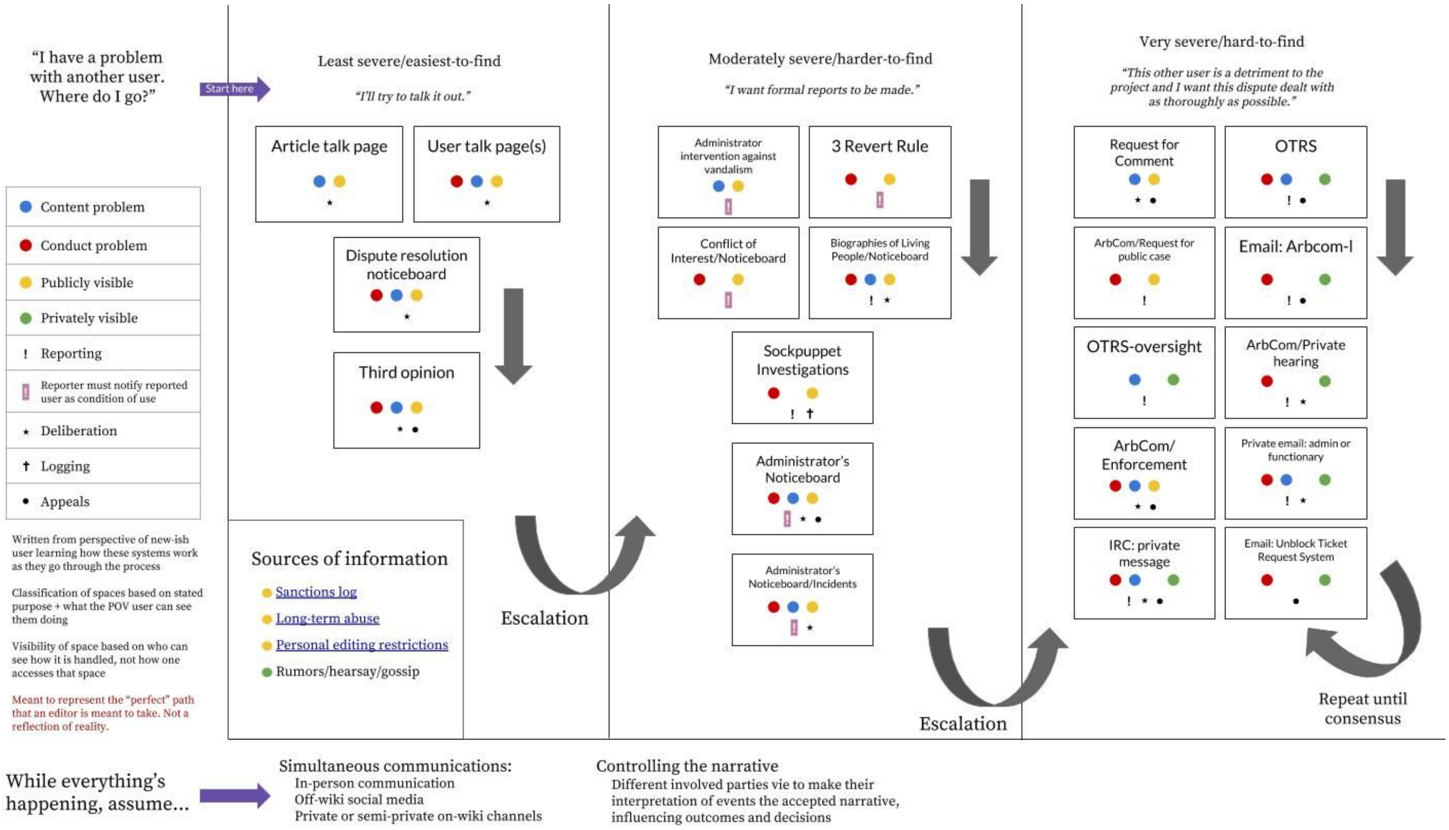

"I have a problem with another user. Where do I go?"
Start here
Content problem
Conduct problem
Publicly visible
Privately visible
! Reporting
! Reporter must notify reported user as condition of use
⋆ Deliberation
† Logging
● Appeals
Written from perspective of new-ish user learning how these systems work as they go through the process
Classification of spaces based on stated purpose + what the POV user can see them doing
Visibility of space based on who can see how it is handled, not how one accesses that space
Meant to represent the "perfect" path that an editor is meant to take. Not a reflection of reality.
Least severe/easiest-to-find
"I'll try to talk it out."
Article talk page
User talk page(s)
Dispute resolution noticeboard
Third opinion
Sources of information
Sanctions log
Long-term abuse
Personal editing restrictions
Rumors/hearsay/gossip
Escalation
Moderately severe/harder-to-find
"I want formal reports to be made."
Administrator intervention against vandalism
3 Revert Rule
Conflict of Interest/Noticeboard
Biographies of Living People/Noticeboard
Sockpuppet Investigations
Administrator's Noticeboard
Administrator's Noticeboard/Incidents
Escalation
Very severe/hard-to-find
"This other user is a detriment to the project and I want this dispute dealt with as thoroughly as possible."
Request for Comment
OTRS
ArbCom/Request for public case
Email: Arbcom-l
OTRS-oversight
ArbCom/Private hearing
ArbCom/ Enforcement
Private email: admin or functionary
IRC: private message
Email: Unblock Ticket Request System
Repeat until consensus
While everything's happening, assume...
Simultaneous communications:
In-person communication
Off-wiki social media
Private or semi-private on-wiki channels
Controlling the narrative
Different involved parties vie to make their interpretation of events the accepted narrative, influencing outcomes and decisions

Fig. 1. A non-exhaustive diagram representing the "ideal" path for resolving editor conflict on English Wikipedia. Accessible on Commons.

Figure 1 represents a non-exhaustive look at existing formal reporting systems and structures on English Wikipedia. From this diagram, we can see that many private systems double as spaces to manage the most serious cases, or otherwise expect some sort of public trail documenting previous attempts to compromise or discuss the issue at hand. However, this also means that for smaller, low-level disputes that nevertheless require privacy or sensitivity to deal with the report, the existing pathways are inadequate. They are also relatively unapproachable for newcomers. Secondly, many formal reporting spaces require the reporter to notify the accused user as a condition of use. While this helps ensure that all parties are informed of the report, putting this burden on the reporter can have a chilling effect, since reporting an incident already represents an escalation. By requiring the reporter to notify the accused user, retaliation and escalation is more likely to happen.

Administrator's Noticeboard/Incidents

For more detail, see the AN/I survey summary.

Commonly abbreviated to AN/I, this is a more general noticeboard nominally for "discussion of urgent incidents and chronic, intractable behavioral problems." Based on conversations with en-wiki administrators, as well as the results of our AN/I survey, this noticeboard is best suited to clear-cut policy violations. Unfortunately, many chronic and intractable behavioral problems have escalated to the point that they are no longer clear-cut. This contributes to the perception of AN/I as a "drama board", or a space where people are not interested in seeking a resolution or reaching a compromise so much as they are invested in fighting other users and trying to "win" the conversation[1]. Not only does this perception potentially discourage users from using AN/I[2], it also discourages administrators from becoming involved and trying to resolve open cases[3].

AN/I provides an interesting look at how some values of the en-wiki community, such as the desire for transparency and public participation, can clash severely with the needs of reporters, accused users and moderators. When faced with clear policy violation, it does allow a reporter to bring a quick, clear case to the attention of administrators. However, for complex interpersonal conflict, the public nature of the noticeboard, and their relatively unstructured nature, makes it extremely difficult to follow a case. This, in turn, makes it extremely difficult to resolve a case.

Specialized noticeboards

These include boards such as edit warring noticeboard, sockpuppet investigations, and other specific noticeboards. Each of these are venues where reporters can bring a claim against an accused user for breaking some specific aspect of policy or engaging in misconduct. Since

[1] To the point where one of the redirects to AN/I is WP:Dramaboard.
[2] See survey responses.
[3] From conversation at WikiConference NA 2018.

content disputes can escalate into conduct disputes, noticeboards that nominally deal in content disputes may also become de facto conduct reporting spaces.

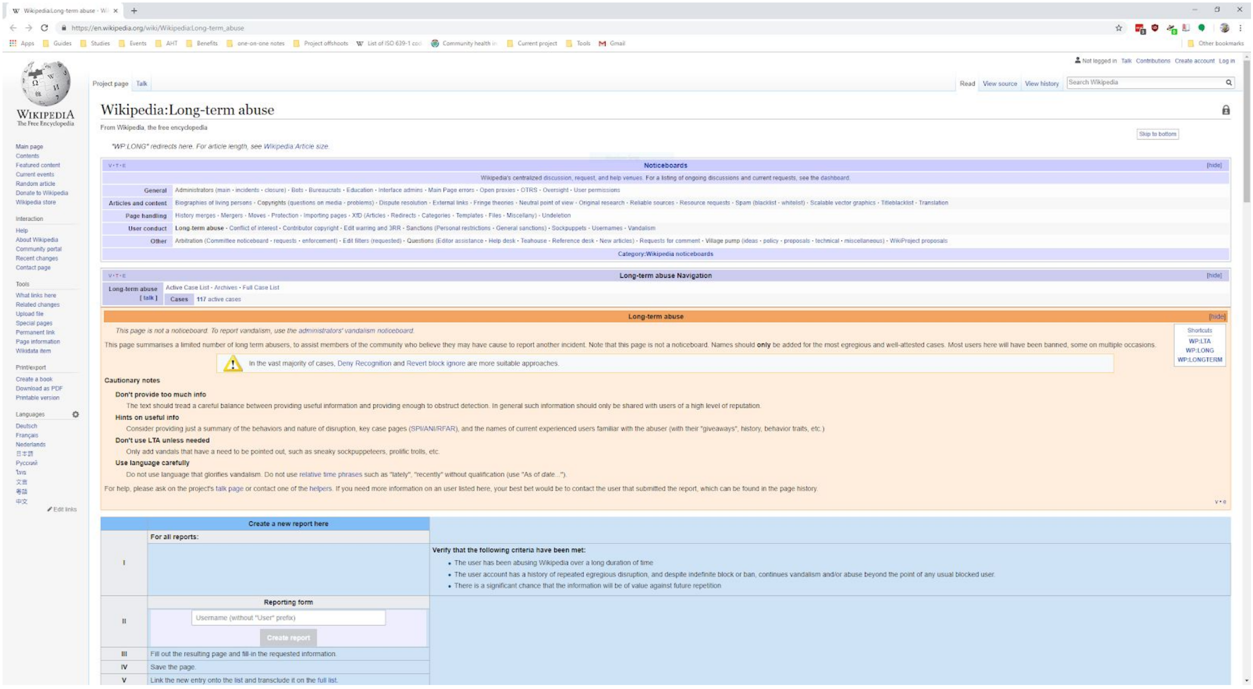
Not as prominent as AN/I, these specialized noticeboards are somewhat cushioned from the "drama board" reputation by their focused nature. However, that same focus plus the relatively open structure of reporting, and their underlying nature as essentially long talk pages, means that they still suffer from some of the same issues as AN/I. Long conversations are difficult to follow, archives are grouped only by date (making it potentially difficult to track long-term patterns of abuse) and they cannot deal well with cases involving breaches of multiple policies.

Long-term abuse, and other public logs

WP:Long-term abuse (or LTA) is an example of the documentation to which editors are accustomed. It is a truncated list of known long-term disruptive editors, many of whom are persistent vandals. Though one of the guidelines for use states that editors should not provide "too much info", to prevent vandals from using LTA as a way to learn bad behaviours, there is still a worry that it forms a "wall of fame" for long-term vandals.

This concern, expressed as WP:BEANS, formed the center of a Request for Comment in 2017, which proposed that the current LTA logs be moved to an off-wiki database accessible only to trusted users. They came to the conclusion that such a database would be useful, but first, they needed to determine how users would qualify as trusted enough to access it.

As it stands, the LTA page provides public information allowing editors to match patterns of abuse they see to known behaviours exhibited by long-term bad faith users. It is also an impromptu and non-indicative reporting space, thanks to the reporting template that can be seen in Figure 2.



Wikipedia:Long-term abuse

From Wikipedia, the free encyclopedia

"WP:LONG" redirects here. For article length, see Wikipedia:Article size.

Noticeboards	
	Wikipedia's centralized discussion, request, and help venues. For a listing of ongoing discussions and current requests, see the dashboard.
General	Administrators (main · incidents · closure) · Bots · Bureaucrats · Education · Interface admins · Main Page errors · Open proxies · OTRS · Oversight · User permissions
Articles and content	Biographies of living persons · Copyrights (questions on media · problems) · Dispute resolution · External links · Fringe theories · Neutral point of view · Original research · Reliable sources · Resource requests · Spam (blacklist · whitelist) · Scalable vector graphics · Titleblacklist · Translation
Page handling	History merges · Mergers · Moves · Protection · Importing pages · XfD (Articles · Redirects · Categories · Templates · Files · Miscellany) · Undeletion
User conduct	**Long-term abuse** · Conflict of interest · Contributor copyright · Edit warring and 3RR · Sanctions (Personal restrictions · General sanctions) · Sockpuppets · Usernames · Vandalism
Other	Arbitration (Committee noticeboard · requests · enforcement) · Edit filters (requested) · Questions (Editor assistance · Help desk · Teahouse · Reference desk · New articles) · Requests for comment · Village pump (ideas · policy · proposals · technical · miscellaneous) · WikiProject proposals
	Category:Wikipedia noticeboards

Long-term abuse Navigation		
Long-term abuse [talk]	Active Case List · Archives · Full Case List	
	Cases	117 active cases

Long-term abuse

This page is not a noticeboard. To report vandalism, use the administrators' vandalism noticeboard.

This page summarises a limited number of long term abusers, to assist members of the community who believe they may have cause to report another incident. Note that this page is not a noticeboard. Names should **only** be added for the most egregious and well-attested cases. Most users here will have been banned, some on multiple occasions.

⚠ In the vast majority of cases, Deny Recognition and Revert block ignore are more suitable approaches.

Cautionary notes

Don't provide too much info

The text should tread a careful balance between providing useful information and providing enough to obstruct detection. In general such information should only be shared with users of a high level of reputation.

Hints on useful info

Consider providing just a summary of the behaviors and nature of disruption, key case pages (SPI/ANI/RFAR), and the names of current experienced users familiar with the abuser (with their "giveaways", history, behavior traits, etc.)

Don't use LTA unless needed

Only add vandals that have a need to be pointed out, such as sneaky sockpuppeteers, prolific trolls, etc.

Use language carefully

Do not use language that glorifies vandalism. Do not use relative time phrases such as "lately", "recently" without qualification (use "As of *date*...").

For help, please ask on the project's talk page or contact one of the helpers. If you need more information on an user listed here, your best bet would be to contact the user that submitted the report, which can be found in the page history.

	Create a new report here	
I	For all reports:	Verify that the following criteria have been met: • The user has been abusing Wikipedia over a long duration of time • The user account has a history of repeated egregious disruption, and despite indefinite block or ban, continues vandalism and/or abuse beyond the point of any usual blocked user. • There is a significant chance that the information will be of value against future repetition
II	Reporting form Username (without "User" prefix) Create report	
III	Fill out the resulting page and fill-in the requested information.	
IV	Save the page.	
V	Link the new entry onto the list and transclude it on the full list.	

Fig. 2. A screenshot of WP:Long-term abuse. Note that there is a template form that serves as a form of reporting.

OTRS

While OTRS is not meant to be a reporting system, and is not advertised as such, it does occasionally handle reports of vandalism or issues with content. The Unblock Ticketing Request System, to be used if the regular method of using the `{{unblock}}` template on one's talk page is not feasible, is based on OTRS software and allows administrators to view unblock requests as tickets.

Pros and cons

Based on community consultation[4], some of the valued features of the existing formal system include[5]:

- Transparency in reporting systems, with public documentation for both public and private reports.
- Assumption of good faith for all involved users.

[4] See User reporting system § Things to keep.
[5] This list of "things to keep" relates to the current state of intertwined informal and formal reporting systems.

- The ability for reporters to influence where and who handles their reports, by picking particular reporting spaces.
- Accountability for moderators.[6]
- Retention of local project authority, allowing reports to be dealt with by local moderators in their native language.

Both systems rely on a combination of appeals to existing community values, policies, and precedent, as well as a track record of satisfactory outcomes, to earn that trust. For formal systems, this trust is earned through adherence to community values both nominally and in practice. For informal systems, the strength of the pre-existing relationship and status of the moderator or moderators involved is critical.

[6] Though this is not defined, we presume this is to do with the ability for the public to see and participate in most formal reports in the current system.

Informal reporting systems

Summary

An informal reporting system would be any *ad hoc* system(s) that allow reports to be made to authorities, whether or not these systems make use of tools that were meant to facilitate reporting. Informal systems strongly rely on the reporter's familiarity with community values and lines of communication.

The strength of informal systems is that, for knowledgeable reporters, it can be much faster, more flexible, and ultimately lead to desired outcomes more efficiently and at lower cost to the reporter. Because they are not constrained in the ways that formal reports are, in terms of how a report must be made or the parameters of said report, they allow a reporter to potentially have finer control over the reporting process. Informal systems can also paradoxically be easier to access for an editor that is well-connected but has had no prior motivation to seek out and learn to use formal systems, again because they leverage already-existing relationships.

However, because the existence of these systems are by nature opaque, they deeply rely on the reporter's social standing and can be extremely difficult for newcomers to these social circles to leverage. Additionally, their informal nature means that they are difficult to track and measure. The general lack of documentation associated with informal reports also means that an informal report may be treated as a discrete event; while this can be advantageous in the immediate term, it is detrimental when trying to record patterns of abusive behaviours in the longer term.

Existing examples

One common mode of informal reporting consists of contacting individual administrators. In our 2018 Wikimania roundtable on harassment, the three most commonly named channels for reporting were IRC, on-wiki talk pages, and email. None of these communication channels are designed for reporting misconduct, yet they are clearly being used as *de facto* reporting systems.

However, these informal systems are not just being used to report cases of harassment. One common repeated point was that these informal channels allowed them 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. Participants note that they use informal systems to communicate to other administrators and see if cases are already being handled elsewhere. The participants, who were highly knowledgeable editors, also used these informal systems 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.

Possible improvements

Some of the most immediate issues of these informal systems is their opacity. Without already knowing who to talk to, or existing familiarity with wiki-adjacent social spaces, it is difficult to access these informal systems. Secondly, it can be even harder to resolve a dispute in the process when that process is informal and, definitionally, somewhat shielded from outside eyes.

The flexible nature of informal systems can also be a flaw, in that it is difficult to ensure a consistent process and experience from one report to the next. It is also very difficult to ensure that the moderators handling informal reports are adequately supported; if a moderator acquires a reputation for being willing to handle complex harassment cases, they may find more cases redirected towards them. There is minimal support for moderators handling these complex tasks currently. Coupled with the unspoken-yet-common fact of abuse targeted at administrators, this can greatly accelerate burnout, paradoxically reducing a project's ability to handle harassment.

Other concerns

One of the social cornerstones of Wikipedia's reporting system are its policies. These govern the procedures for bringing reports, as well as dictate standards on how to conduct conflict resolution. Policy, much like the technical systems they complement, are reflections of a community's values and experience in handling this issue, in addition to their most immediate value as a guide on how to go about resolving disputes.

Values are the moral and social values prized by a given community, which shapes the design of reporting systems. Additionally, these social values provide a frame by which the community understands the authority and legitimacy of a reporting system in all aspects, from its perceived efficacy, to the validity of the outcomes it generates.

For example, the Wikimedia community highly prizes transparency. For reporting systems, this is interpreted as publicly-viewable processes, outcomes, and the identities of the involved users. Transparency in this case is not just a design consideration put into place to achieve a certain kind of efficiency or mode of operation, but a value to be strived for in the way the entire system operates.

Because the current reporting system aligns with a certain dominant interpretation of transparency, the system engenders a feeling of trust from its users. However, we know that the same commitment to transparency can be harmful and serves to chill the participation of other users who are not properly served by the system as it stands. Our current conundrum is the fact that, whatever changes we recommend, it must adhere to these values even as we change key features, otherwise it will not be trustworthy.

Lastly, labour concerns both the question of who performs the work necessary to keep these systems running, as well as the conditions under which that labour is performed. Administrator retention is a topic of enormous concern to administrators and functionaries, and burnout is a persistent spectre hanging over those who perform this mediating and moderating work. The volunteer nature of these roles means that there is no financial compensation, and little to no training or support. Depending on the size of the project, it can also be isolating work. And, because of the fluidity of categories and the value of objectivity across Wikimedia, a moderator who is themself a target of abuse may find it difficult to find help, especially if they are part of only a few admins, or are the only active administrator on their project.

Going forward

This quick overview of a single project shows that, for projects in which a reporting system makes sense, there is plenty of room for improvement along multiple lines. Chief among our concerns are the need to support existing workflows and workers, increasing system accessibility, investigating the factors affecting reporting system effectiveness, and questioning how to measure a reporting system's effectiveness.

Whatever products we can make, we must ensure that they are in keeping with community values, and are compatible with existing workflows so as to make it easier to adopt. On top of that, we want to make sure that we are attentive to labour issues, and consider how to better support and train moderators.

To make a more accessible reporting system, we want to make sure new and existing editors can both easily access public and private channels, aimed at dealing with cases of different severity. However, we want to make sure that, in making new formal private reporting systems, they are still as transparent as is practically feasible. We also want to think of ways to make existing reporting spaces easier to find and access, and consider how editors could find out about these spaces before a crisis forces them to both learn how to report, while dealing with conflict.

In order to do all of the above, we need to know more about what factors affect reporting system effectiveness. Could some design element of the existing system be impacting reporting rates, or closure rates? More fundamentally, we must carefully examine what kinds of metrics we look at to determine success. Reporting rate is not the same as report quality, or closure rates, or the time between first report to first action. The metrics we choose to present as a proxy for reporting system success could greatly impact the way we understand harassment, going forward.