PUBLIC COPY

Note: Budgets and staffing may change over the course the development of this project.

Anti-Harassment Tools For Wikimedia Projects

Summary

- **What we are building**:
  A more advanced system to reduce harassing behavior on Wikipedia and block harassers from the site. We will build new tools that more quickly identify potentially harassing behavior, and help volunteer wiki administrators evaluate harassment reports and make good decisions. Paired with existing tools that we're improving and redesigning, this new system will streamline the way we combat "trolling," "doxxing," and other menacing conduct on Wikipedia.
- **Who will benefit**:
  By providing the volunteer community and Wikimedia Foundation staff with robust tools for detecting, reporting and evaluating harassment, we will enable them to patrol Wikimedia communities more effectively and be able to identify and block harassing users. This in turn will create stronger, more effective, inclusive and diverse communities, benefitting current and future editors alike.
- **How long it will take:**
  28 months for research, community input, design and implementation of a suite of tools for detecting harassment, improving the reporting system, evaluating reported cases, and blocking harassers.
- **How much it will cost:**
  Staff and infrastructure costs related to specific blocking and anti-harassment tool implementation over three years would be approximately $2.1 million.
- **Resources required to make it happen:**
  5 staff members: 4 on the Community Tech product team and 1 from the Community Engagement team.
- **Outcomes at the end of Year 3:**

- Automatic detection systems based on text analysis and models of social interaction, which flag and surface harassment problems to Wikipedia administrators -- the volunteer editors (also called “wiki administrators”) charged with maintaining behavior norms.
- An improved reporting system, which encourages targets of harassment to reach out for assistance.
- A suite of tools that helps Wikipedia administrators to investigate and manage harassment reports.
- A sophisticated blocking and security system that will make it more difficult for users who are blocked from Wikipedia to come back by masquerading under other names and IP addresses.

Introduction

Harassing behavior on Wikipedia undermines the very foundation of the site. "Trolling," "doxxing," and other menacing behaviors are serious burdens to Wikipedia's contributors, impeding their ability do the writing and editing that makes Wikipedia so comprehensive and useful. We are building an overarching system to reduce harassing behavior and block harassers from Wikipedia. Central to this system: Tools that more quickly identify potentially harassing behavior -- that analyze editors' online comments to other editors; that recognize the context of these interactions; and that alert potential harassment cases to Wikimedia staff and to volunteer administrators. Paired with existing editing tools that we’re improving and redesigning, this new system will streamline the way we combat harassment on Wikipedia. It will make it easier and more secure for editors to report harassment. It will make it easier for administrators to review harassing behavior. It will prevent blocked users from coming back under an alias (“sock-puppetry”) and greatly reduce other banned and inappropriate uses of the projects. Collectively, they are a new start in our effort to contain behaviors that were never welcome in the first place.

Harassment on Wikipedia

This project addresses the major forms of harassment reported on the Wikimedia Foundation's 2015 Harassment Survey, which covers a wide range: content vandalism, stalking, name-calling, trolling, doxxing, discrimination -- anything that targets individuals for unfair and harmful attention.

Harassment takes different forms on Wikipedia than it does on other major websites. For the most part, Wikipedia users aren't writing about their personal lives and identities. There are no personalized avatars, and users are discouraged from posting pictures of themselves, or writing in detail about their own lives.

On Wikipedia, harassment usually begins as a content dispute between editors. The trigger could be an article about a current event involving a controversial issue, like gun violence, discrimination laws, or a terrorist attack. A contributor might be criticized for pushing a particular agenda or bias, and that argument can turn personal. The attacker can pick up on a target's personal attributes -- their gender, race, religion, sexual orientation, political affiliation -- based on something that they've shared, or just an assumption based on the user's edit history. Harassment based on identity is often mixed up with criticism based on point of view or editing competence, which makes it difficult to detect and resolve appropriately.

Commonly, the harassment takes the form of written comments, posted on the article's discussion page, on site-wide forum pages, on user profile pages, or even in the short "edit summaries" that editors use to explain why they made a particular change to a page.

The forms of harassment most commonly reported on the 2015 survey were: content vandalism (26% of responses), trolling/flaming (24%), name-calling (17%), discrimination (14.5%), and stalking (13%).

Content vandalism and stalking are similar -- another user "following you around" the site by looking at your contributions, and editing, reverting or commenting on your work to an

extent that makes you feel targeted -- although any single example may seem reasonable to an observer. For editors building a reliable and accurate encyclopedia, resolving disputes through discussion is a critical part of the process -- but a harasser who intentionally uses those systems to target an individual editor can discourage them from participating at all.

That version of harassment is harder to spot, and according to the 2015 survey, it may account for about a third of the reported harassment. Detecting this behavior will require looking at context and patterns of interactions, rather than analyzing individual edits.

The Personal Cost of Harassment

What does harassment actually look like? It can be a comment like this one, made to a female editor on a Wikipedia talk page: "The easiest way to avoid being called a c*** is not to act like one." Harassment also looks like this comment: "All q****s will be shot! You f***ing fa****, I hope you burn in hizzell!" But, as noted earlier, harassment takes on even more menacing forms, like stalking, where the harasser follows the contributor to every forum on Wikipedia and leaves upsetting comments; like "doxxing," where harassers publish private information about contributors; and like death threats.

None of these harassing actions are allowed on Wikipedia, and the Wikimedia Foundation has mechanisms in place that protect editors and ban abusive actions, but harassment happens anyway -- just as it happens across the wider Internet. Thirty-eight percent of respondents in our 2015 Harassment Survey said they had personally experienced harassment. Twenty-seven percent of the poll's female respondents say the harassment they faced lasted more than a year. The experience has a stultifying effect on contributors. Fifty-four percent decrease their participation on the project where they experienced the harassment. Eleven percent completely withdraw from editing or writing articles, leaving behind all they worked for. Other editors do their best to confront the harassment,

exemplified by Emily Temple-Wood, a first-year medical student at Midwestern University who has been editing Wikipedia since she was 12.

Temple-Wood has originated almost 400 articles, and improved hundreds of others, and in 2012 she founded WikiProject Women Scientists -- a collective of contributors who write and maintain biographies on women scientists. Temple-Wood has regularly received salacious emails, many of them sexual, and earlier this year, she took a new tack: For every harassing email she got, Temple-Wood would start a new biography on a women scientist, or would coordinate with another editor from WikiProject Women Scientists to write the biography.

“I was just so frustrated,” she says. “I was like, ‘I need to do something productive with this rage rather than sitting around and being angry – that doesn’t solve anything.’ ”

We agree. We’re strengthening tools that will help our Support and Safety team – the team that ensures our community of editors has a safe and supportive environment to work in – better solve the harassment problem on Wikipedia. We’re building tools that will give Wikipedia editors more immediate power to stop harassment where it starts. We’re developing new tools that will more immediately detect harassing comments, and more immediately lead to editor and staff interventions that end the harassment. We’re conducting new research that will unearth even more ways to confront evasive harassers, which the lead analyst in our anti-harassment project will study upon his or her hiring in March 2017. We plan to scale all our tools over the next three years, and make these tools a new layer of a more comprehensive system to tamp down harassment on Wikipedia. We take these actions with the encouragement of the Wikimedia Foundation's Board of Trustees, which issued a statement of support in November 2016 for a "proactive" approach to deal with harassment on Wikipedia. We won’t eliminate harassment 100 percent. No system can do that. But we can pare it down – way, way down – with a more streamlined approach. That’s what we’re doing now.

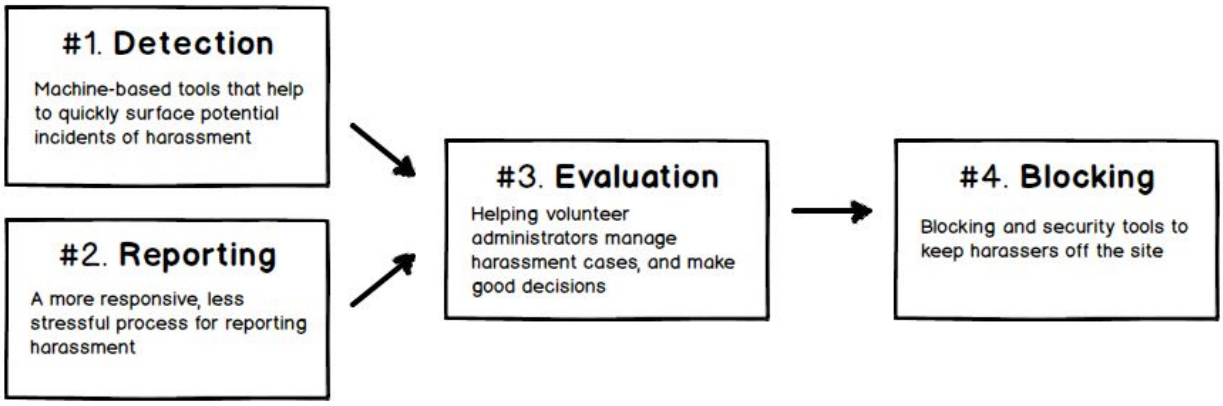
Community input

Community input will be essential to the success of this program. **This will occupy a significant percentage of time and work, and our plans will change in response to community feedback and ideas.** We have an incredibly passionate, intelligent and opinionated community, and working with them will make the project better and more successful. Sometimes, it will also make the project slower and more unpredictable.
The community work will include:

- Socializing our goals, making sure that interested people understand and agree with the premise of our work.
- Generating and refining ideas with community stakeholders. Our active contributors are detail-oriented, and articulate about the way that they work and the tools that they need.
- Conversations about freedom of expression vs. political correctness. It's very important that this project is seen as addressing the kinds of abuse that everyone agrees about (obvious sockpuppet vandalism, death threats) the kinds of abuse that people will differ over (gender, culture, etc.). The project will not succeed if it's seen as only a "social justice" power play.
- Dealing with proverbial foxes. Opening this process up to the community means that we're inviting both foxes and chickens to talk about security features for the henhouse. This dynamic is unavoidable, and creates patches of uncertainty in the schedule.

Focus Areas

This diagram illustrates the four interrelated focus areas in this project: Detection, Reporting, Evaluation, and Blocking, and how they interact. The team will be working on independent projects in all four of these areas.



#1. **Detection**: Automatically detecting harassing behavior, and surfacing the cases for review by volunteer administrators and Foundation staff. This area will include both analysis of written content and modeling social interactions between users.

#2. **Reporting**: A more accessible and less stressful system that helps targets of harassment reach out for assistance. Under the current system, users report suspected harassers on a public "incidents noticeboard", where anyone can comment on the reported harassment before the case is even investigated, which often creates a chaotic and contentious atmosphere around the harassment cases.

#3. **Evaluation**: Helping wiki administrators and staff to evaluate, manage and make good decisions about the cases raised in #1 and #2. Currently, investigating a report of long-term harassment is hindered by the limited features of on-wiki tools, which are great at tracking an individual's contributions, but don't help administrators understand the interactions between different users.

#4. **Blocking & security**: Tools to help wiki administrators and staff keep harassers off the site. One of the MediaWiki tools available to our volunteer administrators is the "block" – the technical ability to stop a problematic user from editing our sites. Our blocking framework is based solely on IP addresses – a system developed in the early 2000s that technically proficient users can easily circumvent. We will update these blocking tools to reflect the current state of the art technology.

These four focus areas interact with each other, but we can work on them independently and make iterative improvements. For example, we can begin work immediately on improving some blocking and security tools (#4), based on blocking the harassers that we currently know. As we build tools and processes to better identify harassers (#1 and 2), that feeds into further iterative work on evaluation (#3) and blocking (#4).

These tools will require some investment in the Support and Safety team for use in anti-harassment, as well as training of volunteer administrators and other community members.

This work falls within the remit of Community Tech, the Wikimedia Foundation product team focused on meeting the needs of active Wikimedia contributors for improved, expert-focused curation and moderation tools. The team has recently done some work on AbuseFilter and improving the existing blocking system. Community Tech's Product Manager, Danny Horn, and Engineering manager, Ryan Kaldari, each have more than ten years of wiki community/product development experience. The team also includes Leon Ziemba, a Wikipedia CheckUser with substantial blocking/anti-vandal experience. Department Head Toby Negrin held technical and product management roles in anti-abuse at Yahoo, and this experience complements Ryan, Danny, and Leon's wiki work.

Schedule of Work

The plans below refer to the four focus areas – Detection, Reporting, Evaluation, and Blocking. There will likely be ongoing work in each area during any given time period, in one stage or another.

The projects that we work on will be informed by a deep analysis of the problems and investigations of promising solutions, with a range of stakeholders.

Please note that these plans will change through the lifetime of the project, based on community input. (See "Community input" section above.) We will, of course, communicate any substantive changes to the plan.

1st year (4 months): March to June 2017, FY 2016-2017

Community input:
During the first four months of the grant period, the product manager and community advocate will be working closely with the volunteer community, to gain acceptance and generate enthusiasm for the overall project goals. This will include:

- Initial presentation on project goals, tying work on anti-harassment tools to fighting other disruptive behavior, including content vandalism and conflict of interest editing.
- Organize and run 5 live-stream online workshops with important stakeholders, to learn more about current needs. Three sessions will be tailored for wiki administrators who handle disruptive behavior. Two sessions will be specifically for users who have experienced or witnessed long-term harassment on Wikimedia projects.
- Organize and run an in-person workshop at the Wikimedia Hackathon in late May, to test approaches and generate ideas.
- Hold a project-wide online community consultation in June to build agreement and support for the project in the community as a whole.

- The community consultation concludes with a report delivered in July, synthesizing learnings and laying out a plan for year two.

The software development work that we do during this period will consist of improvements to existing tools that already have widespread community support. We will also begin research and analysis on some of the larger issues.

Detection:

- Rewrite AbuseFilter extension, which automatically flags or blocks edits according to a pre-set schema, to make it possible to write new plug-in modules.
- Improve ProcseeBot, a volunteer-written tool that searches internal and external black and whitelists for networks and open proxies, to make the tool stable and maintainable.

Evaluation:

- Begin research to define user interactions that are likely to involve content vandalism and stalking.

Blocking:

- Research connection information and metadata such as IP address, proxy, location, network, etc., that we could use to improve the blocking system, for use in year 2.

2nd year (12 months): July 2017 to June 2018, FY 2017-2018

Community input:

- Build on 1st year's community work to build support for making changes to the reporting system. There is general agreement in the community that the current system is chaotic, stressful and prone to reaching the wrong outcomes, but building consensus for changing the system is a big project. This will involve the same kind of steps that are outlined above – presenting initial ideas, conducting online workshops and feedback sessions, conducting in-person workshops at 2 large Wikimedia

conferences, and on-wiki community consultation. The goal for this year is to have a finished prototype for the new reporting system, which can be put into place in year 3.

- In year 2, we will also work with the community on creating a new user group for volunteer administrators who want to work specifically on harassment cases. The community advocate will work with the Support and Safety team to provide training and support for these volunteers.

Detection:

- Design and build a more user-friendly interface for AbuseFilter, to expand the pool of volunteers who can keep the filters up-to-date.
- Build AbuseFilter modules for violent threats and disclosures of personal information.
- Improve existing anti-spoof tools, which detect submissions that use numbers and symbols to work around content bans, to reduce mapping limitations and build a more robust spoof library. Build AbuseFilter module using new anti-spoof tools.
- Prototype using social interaction modeling to surface potential harassment incidents to wiki administrators.

Evaluation:

- Design and build Interaction history tool, which will allow wiki administrators to understand the interaction between two users over time, and make informed decisions in harassment cases.
- Create a private system for wiki administrators to collect information on users' history with harassment and abuse cases, including user restrictions and arbitration decisions. (In the current system, administrators need to search in many different places to find this historical data.)

Blocking:

- Design and build Per-page blocking tool, which will help wiki administrators to redirect users who are being disruptive without completely blocking them from

contributing to the project; this will make wiki admins more comfortable with taking decisive action in the early stages of a problem.
- Make global CheckUser tools work across projects, improving tools that match usernames with IP addresses and user agents so that they can check contributions on all Wikimedia projects in one query.
- Use analysis of connection information and metadata to build more robust sockpuppet blocking tools.

3rd year (12 months): July 2018 to June 2019, FY 2018-2019
It is likely that we will further refine our third-year plan in year 2 of the project, after the Senior Analyst has assessed the progress and community input that emerges when the project is undertaken. As with any complicated project, some parts of the program will go faster than anticipated and some slower, and year 3 will probably include some measurable percent of time working on pieces that require extra attention.

Some of the work in year 3 will involve internationalizing tools that were built in year 2, making them available for all project languages.

Detection:
- Build tools that use the social interaction modeling to surface content vandalism and stalking problems to wiki administrators and staff, based on previous research and prototypes.

Reporting:
- Implement the new harassment reporting system, as defined in year 2. Monitor use of the new system, and make improvements.

Evaluation:
- Design and build a dashboard system for wiki administrators to help them manage current investigations and disciplinary actions.

- Build cross-wiki tools that allow wiki administrators to manage harassment cases across wiki projects and languages.

Blocking:

- Identify new sources of data and signals to incorporate in mainline blocking tools.

Resources Needed

5 new, full-time staff:

A **senior analyst** -- an experienced harassment/abuse fighter familiar with current internet architecture and attacker techniques, to profile and understand WP specific attacker techniques

A product manager to understand existing and desired anti-harassment workflows that can be addressed via the data/understandings generated.

A back-end engineer and **a front-end engineer** to build new tools and augment existing tools, with the direction of the analyst and PM.

A community advocate to work with community members on implementing the new suite of anti-harassment tools and measures:

- Facilitating regular online community consultations about tools, policies, values and the project as a whole
- Coordinating with product manager on incorporating community feedback in project plans
- Attending several wiki conferences per year to present on plans, and facilitate in-person discussions
- Leading trainings for wiki administrators on updated tools and practices

- Organizing a new user group, volunteer administrators who specialize in evaluating harassment cases
- Documenting existing practices

Total Cost: $2.1 million