Selected Papers of #AoIR2024:
The 25th Annual Conference of the
Association of Internet Researchers
Sheffield, UK / 30 Oct - 2 Nov 2024

AUTOMODERATOR AS AN EXAMPLE OF COMMUNITY-DRIVEN DEVELOPMENT

Claudia Lo
Wikimedia Foundation

Sam Walton
Wikimedia Foundation

Introduction

Rushes to adopt the latest technologies to the field of community moderation are deeply lopsided (Franco et. al., 2023; Weng et. al. 2023). The closed-door nature of product development at the majority of tech companies means that the logic underlying the creation of new features is generally opaque. Many alternative approaches to design have been articulated that take a more collaborative approach, including design justice (Costanza-Chock, 2020), participatory design (Kensing and Blomberg, 1998; Callon, 2004), and social design (DiSalvo et al, 2011).

We present the development and deployment of the Wikimedia Foundation's Automoderator[1], an automated anti-vandalism tool, as an example of a product developed through community-driven design principles. Product development for the Automoderator project included a participatory process, where the product team integrated volunteer feedback and direction on a continuous basis, rather than writing software in isolation and deploying rapidly. While it is helpful to analyze and critique the design approach laid out here, we should recognize that these practices do not explicitly follow any of the collaborative frameworks above, and also do not necessarily represent the practices of every Wikimedia Foundation product team.

Content Moderation on Wikipedia

On average Wikipedia receives 4 edits every second across more than 300 languages (Wikistats, n.d.). While the Wikipedia community and Wikimedia Foundation has invested in automating the process of patrolling recent edits to Wikipedia articles, the

[1] https://www.mediawiki.org/wiki/Moderator_Tools/Automoderator

Suggested Citation (APA): Lo, C and Walton, S. (2024, October). *Automoderator as an Example of Community-Driven Development*. Paper (or panel) presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

task remains time consuming. There is a constant stream of changes from contributors ranging from small single-word edits, to entire new articles. While some edits may require careful human review, some low quality edits are obvious vandalism, and can be reverted immediately. Such edits may not individually take up much volunteer time, but in aggregate editors can spend a lot of time dealing with routine edit reverts.

Since the early days of Wikipedia in 2001, volunteers have created automated tools to more easily identify and automate parts of this workload. In specific, volunteer-created anti-vandalism bots were a major inspiration for the Automoderator project. Such bots are important to these communities in their efforts to combat vandalism on their projects (Geiger and Halfaker, 2013). However, they are only available on a very small minority of Wikipedias, and each requires a technically proficient volunteer to create, configure, and maintain it. This fractured ecosystem puts strains on individual volunteers, who are duplicating each other's work, and become single points of failure for systems which the wider community considers vital.

Beyond demonstrating a desire for automated anti-vandalism in general, these community-created bots gave clear directions for Automoderator's development based on their successes and failures. These existing tools steered Automoderator's development towards the values of accuracy, transparency, and human control over all its actions.

Developing and Deploying Automoderator

Before any code was written for the Automoderator software, the Moderator Tools product team engaged with a wide range of volunteer contributors to understand what features it should have, how trustworthy it needed to be, which language communities might be interested in trialing it, and what configuration options it ought to have. The team specifically contacted the authors of older anti-vandalism bots, to learn about the solutions they had built, and how Automoderator might be able to build on their learnings.

At the same time, the goals and key metrics for the project were shared publicly and comments and feedback were invited[2] and integrated from a wide range of volunteers. A public page was created as soon as the project was first scoped[3], which was kept up to date with progress updates and opportunities for feedback throughout.

During its development, the Automoderator team developed a testing tool in the form of a spreadsheet, allowing any user from any community to compare how their personal judgements would compare against Automoderator's judgements. This allowed community members to assess the tool's efficacy before full deployment. It also allowed them to see how Automoderator might impact their wikis, without needing to wait for the fully finished tool. These tests also allowed the development team to gauge whether volunteers considered the prototype to be reliable enough for deployment.

[2] Visible on the public discussion page https://www.mediawiki.org/wiki/Talk:Moderator_Tools/Automoderator

[3] https://www.mediawiki.org/w/index.php?title=Moderator_Tools/Automoderator&oldid=5950762

The team also integrated functionality from the Community Configuration extension[4], an extension that allows Wikimedia communities to customize MediaWiki features for their specific needs, into Automoderator. This addition enables Wikipedia administrators to control Automoderator's behavior directly (Walton, 2023). It also provides administrators with immediate control over whether Automoderator is running at all. All of Automoderator's user-facing text can be customized per-wiki, and editors have access to a monitoring dashboard[5] to review statistics about its behavior, and its estimated false positive rate (Velaga, 2024).

Each Wikipedia language edition must initiate the deployment of Automoderator by making a request to the product team. By giving the option to opt-in to use the product on their wikis, Automoderator's product team gave control over the timing of its deployment to the communities. At the time of writing, Automoderator is operational on five Wikipedia projects. It uses a language-agnostic ML model[6] that predicts the probability of an edit being reverted (Trokhymovych et. al, 2023). A further round of user testing is ongoing with a more highly performant multilingual ML model (Walton, 2024).

Discussion

The above approach to product development hews closer to community-collaborative design approaches, but there are some trade-offs to this product development approach.

Time. The most obvious trade-off is time. The need to continually collect feedback, and optionally translate both ways, creates additional work and dependencies. Moreover, this level of continuous feedback requires committing to maintaining long-term relationships with chosen communities. For the Automoderator project, facilitating multilingual collaborative development depended upon:

- access to automated translations for quick replies,
- access to multilingual support for design research, with professional translators and interpreters,
- engaging with community members fluent in English or hiring community members to help facilitate discussions in their communities,
- constant monitoring of the central product page, as well as discussions happening in other community fora in their own languages.

Deep knowledge of Wikipedia patrolling operations. The highly technical nature of Automoderator, as well as its use-case, meant that the product team needed extensive familiarity with the context of moderation on Wikipedia. This added an additional layer of complexity to the development process as well as the continual feedback cycle.

Accessibility. The population of users who do respond to requests for feedback are not guaranteed to be representative, either of the end-users of the product or of their

[4] https://www.mediawiki.org/wiki/Community_Configuration

[5] https://superset.wmcloud.org/superset/dashboard/unified-automoderator-activity-dashboard/

[6] https://meta.wikimedia.org/wiki/Machine_learning_models/Production/Language-agnostic_revert_risk

community at large. The creation of the testing spreadsheet was a conscious effort to broaden the pool of respondents, but to an extent, the population of community members who have the inclination, time and interest to respond is self-selecting.

Equity. While the model presented in this paper has significant advantages in providing opportunities to users for engaging in the product development process, it does not entirely address the challenges of equitably participating in such processes. For example, we expect that users with technical and economic advantages will be able to more actively engage in such a product development process.

We hope that by sharing more about a contemporary alternative system for product development, we demonstrate that it is possible to conduct product development in a more collaborative manner within the technology world.

References

Callon, M. (2004, March 1). *The role of hybrid communities and socio-technical arrangements in the participatory design*. https://www.semanticscholar.org/paper/The-role-of-hybrid-communities-and-socio-technical-Callon/1b69de82480500cdc37dcb3134452e3706679046

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

DiSalvo, C., Lodato, T., Fries, L., Schechter, B., & Barnwell, T. (2011). The collective articulation of issues as design practice. *CoDesign*, *7*(3–4), 185–197. https://doi.org/10.1080/15710882.2011.630475

Franco, M., Gaggi, O., & Palazzi, C. E. (2023). Analyzing the Use of Large Language Models for Content Moderation with ChatGPT Examples. *3rd International Workshop on Open Challenges in Online Social Networks*, 1–8. https://doi.org/10.1145/3599696.3612895

Kensing, F., & Blomberg, J. (1998). Participatory Design: Issues and Concerns. *Computer Supported Cooperative Work (CSCW)*, *7*(3–4), 167–185. https://doi.org/10.1023/A:1008689307411

Moderator Tools/Automoderator. (n.d.). MediaWiki. Retrieved November 22, 2024, from https://www.mediawiki.org/wiki/Moderator_Tools/Automoderator

Moderator Tools/Automoderator/Multilingual testing. (n.d.). MediaWiki. Retrieved November 22, 2024, from https://www.mediawiki.org/wiki/Moderator_Tools/Automoderator/Multilingual_testing

Utilisateur:Salebot. (2023). In *Wikipedia*. https://fr.wikipedia.org/w/index.php?title=Utilisateur:Salebot&oldid=208729762

Velaga, K. (2024, July 8). *⚓ T369488 Develop a unified Automoderator Activity Dashboard (v1)*. https://phabricator.wikimedia.org/T369488

Walton, S. (2023, October 20). *⚓ T349374 How will communities configure Automoderator?* Wikimedia Phabricator. https://phabricator.wikimedia.org/T349374

Walton, S. (2024, May 22). *⚓ T365581 Use multilingual revert risk model in Automoderator on supported wikis*. Wikimedia Phabricator. https://phabricator.wikimedia.org/T365581

Weng, L., Goel, V., & Vallone, A. (2023, August 15). *Using GPT-4 for content moderation | OpenAI* [Blog]. OpenAI Blog. https://openai.com/index/using-gpt-4-for-content-moderation/

Wikistats. (n.d.). *Wikimedia Statistics—All wikis - Edits*. Retrieved November 26, 2024, from https://stats.wikimedia.org/#/all-projects/contributing/edits/normal|bar|2022-10-01~2024-12-01|~total|monthly