

Patrolling on Wikipedia

Jonathan T. Morgan, Wikimedia Research



WIKIMEDIA
FOUNDATION

https://meta.wikimedia.org/wiki/Research:Patrolling_on_Wikipedia/Report

tl;dr

I want to understand how WMF can help editors address current and future challenges to **content integrity**.

I read a bunch of papers about Wikipedia vandalism.

I interviewed editors who do anti-vandalism patrolling work and/or build tools that other editors use to detect, report, and address vandalism.

Project overview

Goals

- Understand how patrolling tools and workflows *differ across small and large Wikipedias*
- Understand differences between *fast vs slow, single-wiki vs cross-wiki* workflows
- Identify *limitations and gaps* in current infrastructure that create vulnerabilities

Scope

- Workflows of editors who patrol on Wikipedia(s)
- Tools these editors use in the course of their work
- Current study excludes workflows and tools that are specific to Commons and WikiData patrolling

Methods

- **Review of research literature** on patrolling and vandalism
- **Interviews with 4 Wikipedia editors** (including tool developers/maintainers, local and cross-wiki admins)


Why research patrolling?

- **Editors patrol recent pages and edits** to ensure that Wikimedia projects maintains high quality as new content comes in.
- **Patrolling is Wikipedia's first line of defense** against disinformation, copyright infringement, libel and slander, threats, and other forms of vandalism.
- **Patrolling is supported by tools:** special userrights, Mediawiki software features, bots, gadgets, noticeboards, dashboards, and more.
- **Patrolling tools and activities vary** from project to project.

Terminology

I will use terms borrowed from the domain of cybersecurity like *threat model*, *attack vector*, and *structural vulnerability*.

I will also use the term ‘patrolling’ and ‘anti-vandalism’ somewhat interchangeably to describe the activity of reviewing (usually) recent content contributions for quality assurance purposes.



The screenshot shows the Wikipedia:Patrolling page. At the top, there is a navigation bar with the user name 'Jmorgan (WMF)', a notification bell, and a list of links: 'Talk', 'Sandbox', 'Preferences', 'Beta', 'Watchlist', 'Contributions', 'Log out', and the time '19:30:14'. Below this is a secondary navigation bar with 'Project page', 'Talk', 'Read', 'Edit source', 'View history', 'More', and 'TW'. A search box on the right contains the text 'Search Wikipedia'. The main heading is 'Wikipedia:Patrolling', followed by the subtitle 'From Wikipedia, the free encyclopedia'. A section titled 'You may look for:' contains a bulleted list of links: 'Wikipedia:New pages patrol', 'Wikipedia:New pages patrol/patrolled pages', 'Wikipedia:Recent changes patrol', 'Wikipedia:Vandalism', 'Wikipedia:Counter-Vandalism Unit', and 'Wikipedia:Cleaning up vandalism/Tools'. At the bottom, a box labeled 'Categories:' contains the link 'Wikipedia patrols'.

Patrolling tools

Default toolset (all wikis have these)

- Special: pages
- Elevated user rights
- Diff, history, and discussion pages
- Standard MW extensions

Extended toolset (differ across wikis)

- Bots
- Gadgets, userscripts and custom extensions
- Assisted editing programs
- On-wiki reports and triage boards
- On-wiki noticeboards
- External comms channels
- Web applications

Patrolling Fast and Slow



Fast patrolling

Features

- Instinctive, heuristic-based decision-making
- Usually an individual activity
- Performed by dedicated patrollers
- Well-defined workflows

Purpose

- Review of most/all new changes to the wiki
- Remove obvious vandalism quickly
- Stop attacks in real time

Key tools

- Special:Recent changes
- Abuse filters
- Patroller user right
- Assisted editing programs
- Anti-vandal bots
- Real-time recent changes (RTRC)

Slow patrolling

Features

- Deliberative, context-sensitive decision-making
- Individual or collaborative activity
- Performed by a wide variety of editors
- Complex or ill-defined workflows

Purpose

- Fill in gaps in fast patrolling
- Review recent(ish) *or* historical edits that are related to content I'm personally invested in
- Assess time-consuming judgement calls
- Investigate suspicious patterns of behavior

Key tools

- Watchlists
- Related changes
- Editor/edit/page histories and logs
- Checkuser user right
- Noticeboards, IRC channels, mailing lists
- Triage dashboards and worklists

Threat model

Fast patrolling

- **Patroller userright is too easy to obtain.** Vandals sneak in and start patrolling each others edits to avoid scrutiny
- **Patroller userright is too hard to obtain.** Not enough trusted editors engage in fast patrolling and vandalism slips through

Slow patrolling

- **Serendipitous and ad hoc.** Depends on active, trusted editors watching the right pages and following up on suspicious edits

Overview

Important factors

- Type of vandalism
- Project size by articles, edits, pageviews
- # active registered editors
- # editors with elevated permissions
- Availability of specialized patrolling tools

Key tools

- Bots (e.g. ClueBot_NG)
- On-wiki reports, worklists, and noticeboards (e.g. AN/I)
- Gadgets, userscripts, specialized extensions (e.g. Twinkle)
- Assisted editing programs (e.g. Vandalfighter)
- External comms channels (e.g. IRC, listserves)
- Web apps (e.g. CopyPatrol)

Overview

Important factors

- Type of vandalism
- Project size by articles, edits, pageviews
- **# active registered editors**
- **# editors with elevated permissions**
- **Local availability of specialized patrolling tools**

Key tools

- Bots (e.g. ClueBot_NG)
- On-wiki reports, worklists, and noticeboards (e.g. AN/I)
- Gadgets, userscripts, specialized extensions (e.g. Twinkle)
- Assisted editing programs (e.g. Vandalfighter)
- External comms channels (e.g. IRC, listserves)
- Web apps (e.g. CopyPatrol)

Threat model

Large Wikipedias

- Sockpuppets and IP-hopping
- Sleeper accounts
- Account hacking and zombie accounts
- Meat puppets and tag-teams
- Brigading

Threat model

Large Wikipedias

- Sockpuppets and IP-hopping
- Sleeper accounts
- Account hacking and zombie accounts
- Meat puppets and tag-teams
- Brigading

Small Wikipedias

All of the large Wikipedia threats, plus...

- Lack of access to the best available tools
- Fewer local tool-builders (and maintainers)
- Fewer local editors with elevated userrights and subject matter expertise
- Fewer editors available to counter high-volume attacks in real time
- Greater risk of being hijacked by insiders
- Fewer abusefilters
- Target of opportunity for vandals whose efforts have been stymied on large Wikipedias
- Lack of local reporting and remediation forums

Overview

Important factors

- Wikimedia content is highly integrated across projects through articles and search
- Some content is surfaced/transcluded across projects (e.g. images, WikiData values)
- Microcontribution workflows allow cross-project editing

Default tools

- Global IP block logs
- CentralAuth log
- Global Userrights

Community-created tools

- IRC bots
- Private IRC, mailing lists, and wikis
- Global support request noticeboards
- Global user contributions tools
- Global spam blacklist

Threat model

Wikipedias

- Cross-wiki 'related changes' are invisible by default
- Most blocks (accounts, IPs, and IP ranges) and bans (accounts) are local
- Global noticeboards have limited i18n
- Global reporting workflows are high-touch and time-consuming for all parties involved

Threat model

Wikipedias

- Cross-wiki ‘related changes’ are invisible by default
- Most blocks (accounts, IPs, and IP ranges) and bans (accounts) are local
- Global noticeboards have limited i18n
- Global reporting workflows are high-touch and time-consuming for all parties involved

WikiData and Commons

All of the Wikipedia threat, plus...

- Massive-scale transclusion of local content into Wikipedias creates unpredictable attack vectors
- Direct edits to local content may be immediately visible on Wikipedias, with no local edit trace
- Recent microcontribution workflows has dramatically increased edit volume by unaffiliated editors

Recommendations



Technological interventions

- Cross-wiki watchlists & related changes
- Central incident databases for vandalism
- Social media inbound traffic reports

Further research

- **Sockpuppet detection:** *(ongoing)*
- **Zombie accounts:** what kind of edits do resurrected accounts perform?
- **Patrolling of microcontributions:** how and to what extent are Commons and WikiData patrolling these contributions?
- **Cross-wiki vandalism:** when and how frequently do vandals edit multiple wikis in the same edit session?
- **Coordinated disinformation case studies:** can we collect rich descriptions of previous disinformation campaigns?
- **Social media traffic vs. vandalism:** can we model the relationship between traffic spikes and suspicious edit patterns?
- **Reference changes:** can we identify source additions, removals, or replacements associated with vandalism?
- **Unreliable sources:** can we identify known (or probable) disinformation websites?