Ensuring Credible Information: Content Moderation Techniques on Wikipedia
1. Patrollers

2. Sources
User Roles & Flags

- Unregistered or not logged in
- Registered new users
- Autoconfirmed users
- Administrators
- Bueraucrats
- Patrollers
- CheckUsers
- Autopatrolled
- Bots
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.
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)
Recent changes

This is a list of recent changes to Wikipedia.

Other review tools

Active filters
- Human (not bot) [X]  Page edits [X]  Page creations [X]  Logged actions [X]

_filter changes (use menu or search for filter name)_

_filters_

User Intent predictions  How do these work?
- Very likely good faith
  Highly accurate at finding almost all good-faith edits.
- May be bad faith
  Finds most bad-faith edits but with a lower accuracy.
- Likely bad faith
  With medium accuracy, finds an intermediate fraction of bad-faith edits.
- Very likely bad faith
  Very highly accurate at finding the most obvious bad-faith edits.

User registration and experience
- Unregistered
  Editors who aren't logged-in.
- Registered
  Logged-in editors.
- Newcomers
  Registered editors who have fewer than 10 edits or 4 days of activity.
- Learners
  Registered editors whose experience falls between "Newcomers" and "Experienced users."
- Experienced users
  Registered editors with more than 500 edits and 30 days of activity.

Contribution authorship
- Changes by you
  Your own contributions.
- Changes by others
  All changes except your own.

Automated contributions
- Bot
  Edits made by automated tools.
- Human (not bot)
  Edits made by human editors.

Significance
- Minor edits
  Edits the author labeled as minor.
- Non-minor edits
  Edits not labeled as minor.

Latest revisions
- Latest revision
  Only the most recent change to a page.
- Not the latest revision
  All changes that are not the "latest revision."

Type of change
- Page edits
  Edits to wiki content, discussions, category descriptions...
- Page creations
  Edits that make new pages.
- Category changes
  Records of pages being added or removed from categories.
- Wikidata edits
  Edits that originate on Wikidata.
- Logged actions
  Administrative actions, account creations, page deletions, uploads...

Advanced filters
ORES review tool

The ORES review tool is the key user-facing feature of the ORES extension, which provides objective revision evaluation services to automatically rate a revision's characteristics: likelihood it is vandalism, degree to which it might be damaging, likelihood of being good faith, likelihood it will be reverted, and overall quality. The review interface integrates the scores generated by the ORES service into MediaWiki's interface. ORES provides automated scoring of revisions in order to aid editors. For example, ORES can predict whether or not an edit is vandalism, as well as the overall quality level of an article. See ORES' documentation for more information about what types of scoring are available.

The default threshold is deliberately set low to capture almost all vandalism cases (so a number of false positives are also likely to occur). This is in contrast to anti-vandalism bots which set the threshold high to capture only the most obvious vandalism cases (and thus have few false positives). If you do not want to see the flag for most edits, you can simply change ORES sensitivity (see below).
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
Watchlist

- A **request for adminship** is open for discussion. [dismiss]
- Editors are invited to **provide feedback** on the candidates for the 2022 Arbitration Committee Electoral Commission until 23:59 October 15, 2022 (UTC). [dismiss]

You have 11 pages on your watchlist (excluding talk pages). Changes to pages since you last visited them are shown with solid markers.

Mark all changes as seen

Active filters

No active filters. All contributions are shown.

Filter changes (use menu or search for filter name)

Live updates

11 October 2022

- (diff | hist) . b Dharamshala; 23:35 (+191) . InternetArchiveBot (talk | contribs) (Rescuing 1 sources and tagging 0 as dead.) #IABot (v2.0.9.2) (Whoop whoop pull up - 10860)
11 октомври 2022

- (разл | ист) . дм Еманюел Макрон (03052772); 13:40 . Viskonsas (беседа | приноси) (Добавена междуязыковая препратка: bat-smg:Emanuilis Makrons)
- (разл | ист) . Уикипедия:Разговори; 05:30 . (-476) . 87.116.107.86 (беседа) →Актуализация (Етикети: Отменени, Редакция чрез мобилно устройство, Редакция чрез мобилно приложение)
- (разл | ист) . дм Брюксел (община) (Q239); 05:19 . Zubryckiy (беседа | приноси) (Променена междуязыковая препратка: от mdf:Брюксель на mdf:Брюссель)
- (разл | ист) . мб Уикипедия:Разговори; 01:39 . (-826) . Kerberizer (беседа | приноси) (Bot: Archiving 1 thread (older than 15 дни) to Уикипедия:Разговори/Архив/2022/септември)

10 октомври 2022

- (разл | ист) . Уикипедия:Разговори; 18:00 . (+458) . Rebelheartous (беседа | приноси) →Използване на шаблони "номинация" и "награден"
- (разл | ист) . Уикипедия:Разговори; 17:56 . (+448) . Rebelheartous (беседа | приноси) →Надкатегории
- (разл | ист) . дм Еманюел Макрон (03052772); 16:04 . Photoshoper97 (беседа | приноси) (Създадено изявление: Property:P8879: 3350923)
- (разл | ист) . дм Еманюел Макрон (03052772); 16:03 . Photoshoper97 (беседа | приноси) (Премахнато изявление: Property:P2604: 4108228)
- (разл | ист) . дм Уикипедия:Разговори; (Q16503); 15:19 . Yahva (беседа | приноси) (Добавена междуязыковая препратка: bnwikitext:उइकुड़ि:अलौचना)
- (разл | ист) . дм Чарлоте Ванкел (Q1780957); 14:31 . Lockal (беседа | приноси) (Променен обект: fixing adagp id)
- (разл | ист) . Уикипедия:Разговори; 14:23 . (+28) . Ted Masters (беседа | приноси) →Помощ за шаблон
- (разл | ист) . Уикипедия:Разговори; 14:20 . (+430) . Ted Masters (беседа | приноси)
- (разл | ист) . Уикипедия:Разговори; 14:11 . (+287) . Ted Masters (беседа | приноси) →Помощ за шаблон
- (разл | ист) . Уикипедия:Разговори; 14:05 . (+658) . Elizav22 (беседа | приноси) (Нова тема →Помощ за шаблон) (Етикети: Нова тема, Визуално)
- (разл | ист) . дм Джакомо Бала (Q368254); 11:44 . Lockal (беседа | приноси) (Променен обект: fixing adagp id)
- (разл | ист) . дм Карло Кара (Q168496); 11:25 . Lockal (беседа | приноси) (Променен обект: fixing adagp id)
- (разл | ист) . дм Фернан Леже (Q157183); 11:22 . Lockal (беседа | приноси) (Променен обект: fixing adagp id)
- (разл | ист) . дм Астгер Йорн (Q83578); 10:43 . Lockal (беседа | приноси) (Променен обект: fixing adagp id)
- (разл | ист) . дм Еrnst Лудвиг Кирхнер (Q229272); 09:32 . Bargioni (беседа | приноси) (Премахнато изявление: Property:P9251: 107820, batch #100664)
- (разл | ист) . дм Фернан Леже (Q157183); 09:32 . Bargioni (беседа | приноси) (Премахнато изявление: Property:P9251: 96453, batch #100664)
- (разл | ист) . дм Василий Кандински (Q61064); 09:30 . Bargioni (беседа | приноси) (Премахнато изявление: Property:P9251: 14845, batch #100664)
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
1. Patrollers

2. Sources
Jump to the list of frequently discussed sources.

The following presents a non-exhaustive list of sources whose reliability and use on Wikipedia are frequently discussed. This list summarizes prior consensus and consolidates links to the most in-depth and recent discussions from the reliable sources noticeboard and elsewhere on Wikipedia.

Context matters tremendously, and some sources may or may not be suitable for certain uses depending on the situation. When in doubt, defer to the linked discussions for more detailed information on a particular source and its use. Consensus can change, and if more recent discussions considering new evidence or arguments reach a different consensus, this list should be updated to reflect those changes.

Reliability is an inquiry that takes place pursuant to the verifiability policy and the reliable sources guideline. Note that verifiability is only one of Wikipedia's core content policies, which also include neutral point of view and no original research. These policies work together to determine whether information from reliable sources should be included or excluded.
Generally reliable in its areas of expertise: Editors show consensus that the source is reliable in most cases on subject matters in its areas of expertise. The source has a reputation for fact-checking, accuracy, and error-correction, often in the form of a strong editorial team. It will normally still be necessary to analyze how much weight to give the source and how to describe its statements. Arguments to exclude such a source entirely must be strong and convincing, e.g., the material is contradicted by more authoritative sources, it is outside the source's accepted areas of expertise (a well-established news organization is normally reliable for politics but not for philosophy), a specific subcategory of the source is less reliable (such as opinion pieces in a newspaper), the source is making an exceptional claim, or a different standard of sourcing is required (WP-MEDRS, WP-BLP) for the statement in question.

No consensus, unclear, or additional considerations apply: The source is marginally reliable (i.e. neither generally reliable nor generally unreliable), and may be usable depending on context. Editors may not have been able to agree on whether the source is appropriate, or may have agreed that it is only reliable in certain circumstances. It may be necessary to evaluate each use of the source on a case-by-case basis while accounting for specific factors unique to the source in question. Carefully review the Stability column of the table for details on the status of the source and the factors that should be considered.

Generally unreliable: Editors show consensus that the source is questionable in most cases. The source may lack an editorial team, have a poor reputation for fact-checking, fail to correct errors, be self-published, or present user-generated content. Outside exceptional circumstances, the source should normally not be used, and it should never be used for information about a living person. Even in cases where the source may be valid, it is usually better to find a more reliable source instead. If no such source exists, that may suggest that the information is inaccurate. The source may still be used for uncontroversial self-descriptions, and self-published or user-generated content authored by established subject-matter experts is also acceptable.

Deprecated: There is community consensus from a request for comment to deprecate the source. The source is considered generally unreliable, and use of the source is generally prohibited. Despite this, the source may be used for uncontroversial self-descriptions, although reliable secondary sources are still preferred. An edit filter, 869 (hist · log), may be in place to warn editors who attempt to cite the source as a reference in articles. The warning message can be dismissed. Edits that trigger the filter are tagged.

Blacklisted: Due to persistent abuse, usually in the form of external link spamming, the source is registered on the spam blacklist or the Wikimedia global spam blacklist. External links to this source are blocked, unless an exception is made for a specific link in the spam whitelist.

Request for comment: The linked discussion is an uninterrupted request for comment on the reliable sources noticeboard or another centralized venue suitable for determining the source's reliability. The closing statement of any RFC that is not clearly outdated should be considered authoritative and can only be overturned by a newer RFC.

Stale discussions: The source has not been discussed on the reliable sources noticeboard for four calendar years, and the consensus may have changed since the most recent discussion. However, sources that are considered generally unreliable for being self-published or presenting user-generated content are excluded. A change in consensus resulting from changes in the source itself does not apply to publications of the source from before the changes in question. Additionally, while it may be prudent to review these sources before using them, editors should generally assume that the source's previous status is still in effect if there is no reason to believe that the circumstances have changed.

Discussion in progress: The source is currently being discussed on the reliable sources noticeboard. Italic numbers represent active discussions (all discussions that are not closed or archived) on the reliable sources noticeboard. Letters represent discussions outside of the reliable sources noticeboard.

Shortcut: Abbreviated wikilink to the list entry for the source.
| Source                          | Status (legend) | List  | Last  | Discussions                                                                                                                                                                                                 |
|-------------------------------|-----------------|-------|-------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|
| **112 Ukraine**               | ✗               | 2019  | 2020  | 112 Ukraine was deprecated following a 2019 RFC, which showed overwhelming consensus for the deprecation of a slew of sources associated with Russian disinformation in Ukraine. It was pointed out later in a 2020 RFC that 112 Ukraine had not been explicitly discussed in that first discussion prior to its blacklisting request. Further discussion established a rough consensus that the source is generally unreliable, but did not form a consensus for deprecation or blacklisting. The prior blacklisting was reversed as of out of process. |
| ABC News                      | ✅              | 1     | 2     | There is consensus that ABC News, the news division of the American Broadcasting Company, is generally reliable. It is not to be confused with other publications of the same name.                                                                 |
| Ad Fontes Media               | ✅ 1            | 2 3 4 | 5     | There is consensus that Ad Fontes Media and their Media Bias Chart should not be used in article space in reference to sources' political leaning or reliability. Editors consider it a self-published source and have questioned its methodology. |
| Advameq (City-Data)           | ✗ 1            | 2019  | 2019  | Advameq operates content farms, including City-Data, that use scraped or improperly licensed content. These sites frequently republish content from Gale's encyclopedias; many editors can obtain access to Gale through The Wikipedia Library free of charge. Advameq's sites are on the Wikipedia spam blacklist, and links must be whitelisted before they can be used. WP:COPYLINES prohibits linking to copyright violations. |
| The Age                       | ✅              | 2021  | 2021  | The Age is a newspaper based in Melbourne, Australia. There is consensus that it is generally reliable.                                                                                                   |
| Agence France-Press (AFP)     | ✅              | 2020  | 2020  | Agence France-Press is a news agency. There is consensus that Agence France-Press is generally reliable. Syndicated reports from Agence France-Press that are published in other sources are also considered generally reliable. |
| Al Jazeera (Al Jazeera English, AlJazeera.com) | ✅ 1 | 2 3 4 | 5 6 7 8 9 | Al Jazeera is considered a generally reliable news organization. Editors perceive Al Jazeera English (and AlJazeera.com) to be more reliable than Al Jazeera's Arabic-language news reporting. Some editors say that Al Jazeera, particularly its Arabic-language media, is a partisan source with respect to the Arab-Israeli conflict. Al Jazeera's news blogs should be handled with the corresponding policy. |
| Alexa Internet                | ✅              | 2022  | 2022  | Alexa Internet was a web traffic analysis company owned by Amazon and discontinued as of May 2022. There is no consensus on the reliability of Alexa Internet. Rankings according to Alexa Internet, rankings of low-traffic websites are less reliable than rankings of high-traffic websites, and rankings of 100,000 and above are unreliable. A March 2022 RFC found no consensus on whether citations of Alexa Internet should be removed now that the service is defunct. Due to their instability, Alexa rankings should be excluded from infoboxes. |
| AllSides                      | ✗ 1            | 2022  | 2022  | In a 2022 RFC, editors found no consensus on the reliability of AllSides as a whole. A significant minority of users noted that AllSides has been referenced in reliable sources as an accurate source for media bias ratings, while another significant minority argued that its methodology, which is partly based on the opinions of users, makes it unsuitable for Wikipedia. There is general consensus that reliability varies among the website's articles and should be determined on a case-by-case basis; while the high-confidence ratings are generally reliable as they are reviewed carefully by experts, others depend on blind user surveys that some editors consider opinionated and less reliable. |
| AlterNet                      | ✗ 2            | 2019  | 2019  | There is consensus that AlterNet is generally unreliable. Editors consider AlterNet a partisan source, and its statements should be attributed. AlterNet's syndicated content should be evaluated by the reliability of its original publisher, and the citation should preferably point to the original publisher. |
| Amazon (WP:RSPAMAZON)         | ✗ 2            | 2021  | 2021  | User reviews on Amazon are anonymous, self-published, and unverifiable, and should not be used at all. Amazon is a reliable source for basic information about a work (such as release dates, ISBN, etc.), although it is unnecessary to cite Amazon when the work itself may serve as a source for that information (e.g., authors' names and ISBNs). Future release dates may be unreliable. |
Wikipedia: Nierzetelne źródła

Ta strona jest propozycją zalecenia edycyjnego. Może ona zostać w każdej chwili zmieniona przez każdego użytkownika na podstawie wniosków zawiartych w dyskusji, dopóki nie zostanie osiągnięty konsensus. Jakiekolwiek odniesienia do tej strony nie powinny opisywać jej jako zawierającej obowiązujące reguli lub zalecenia.

Nierzetelne źródła są źródłami wysoce wątpliwnymi, na które wikipedysty nie powinni powoływać się w artykułach, ponieważ w prawie wszystkich okolicznościach nie spełniają one wytycznych dotyczących wiarygodnych źródeł. Próby zapisania edycji zawierających powołanie się na te źródła mogą wywoływać filtr nadużyć.

Uznanie źródła za nierzetelne:
- wynika z zasad oceny źródeł w Wikipedii. Zasady te wymagają poszukiwania się rzetelnymi źródłami i określają, czym charakteryzują się rzetelne źródła,
- nie zmienia oceny źródła w świetle zasad oceny źródeł w Wikipedii, a jedynie dokumentuje tę ocenę. To znaczy: dane źródło może stać się rzetelne i wtedy należy usunąć je z tej listy,
- ma miejsce w celu unikania powtarzających się dyskusji oraz podniesienia świadomości wikipedystów na temat statusu danych źródeł.

Dyskusja o uznaniu źródła za nierzetelne

Można zaproponować uznanie źródła za nierzetelne. Źródło jest dopisywane do listy wtedy, gdy istnieje konsensus społeczności.

Dyskusja, która skutkuje uznaniem dla nierzetelne, może obejmować zmianę lub doprecyzowanie kwestii objętej konsensem (co skutkuje zmianą praktyki).

Porównaj istnieje wiele powodów, dla których źródło może być nierzetelne, konkretne powody odrzucenia różnią się w zależności od przypadku. Pierwszym źródłem formalnie uznanym za takie był ...

Wycofanie źródła różni się od jego zablokowania (czarnej listy), co zwykle ma na celu rozwiązanie problemów związanych ze spamem.

Lista nierzetelnych źródeł

<table>
<thead>
<tr>
<th>Źródło</th>
<th>Link do dyskusji</th>
<th>Uwagi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russia Today</td>
<td>rt.com</td>
<td>Czołowa tuba propagandowa Kremla, wielokrotnie krytykowana za rozpowszechnianie fałszywych informacji oraz propagowanie teorii spiskowych. 24 lutego 2022 roku Krajowa Rada Radiologicznej i Telewizji pozbawiła kanał licencji na nadawanie na terenie Polski.</td>
</tr>
<tr>
<td>Sputnik News</td>
<td>sputniknews.com</td>
<td>Konsensus społeczności Wikipedii wykazał, że strona jest pełna cechowo wprowadzanych w błąd, fałszywych informacji oraz służy oficjalnej rosyjskiej propagandzie. Autorzy poszczególnych wpisów z regulami nie utrzymują swojej tożsamości, co dodatkowo podważa wiarygodność serwisu.</td>
</tr>
<tr>
<td>City Population</td>
<td>citypopulation.de</td>
<td>Serwis zbiera różne dane populacyjne, w przypadku braku danych rzeczywistych przygotowuje dane przewidywane na bazie wcześniejszych spisów. Nie ma możliwości odróżnienia danych rzeczywistych od przewidywanych, co w przypadku państw, gdzie spisy są przygotowywane co kilka lat często mocno różnią się z prawdą.</td>
</tr>
<tr>
<td>Breitbart News</td>
<td>breitbart.com</td>
<td>Skrajnie prawicowy serwis informacyjny znany z wprowadzających w błąd artykułów i szerzenia teorii spiskowych.</td>
</tr>
<tr>
<td>Myśl Konserwatywna</td>
<td>myslkonserwatywna.pl</td>
<td>Skrajnie prawicowa propaganda, fake newsy, prokremiowski trolling. Rynsztokowy poziom wpisów (rozważania o konieczności „obałeniu nazistowskiej junty” celem przeprowadzenia przyszyłej „repolonizacji”). Ukraińscy czynni. 3. dnia rosyjskiej agresji na ten kraj</td>
</tr>
<tr>
<td>Polska</td>
<td>myslpolska.info</td>
<td>Skrajna postać polskojęzycznej propagandy rosyjskiej. Po zabójstwie Danii Duginej serwis przybrał czarno-białe barwy na znak żałoby po propagandystce, którą określił m.in. jako bohaterkę i człowieka wybitnego, pisząc o wojnie na Ukrainie, że zostanie wygrana (śl. przez Rosję). [3]</td>
</tr>
</tbody>
</table>
dimi@wikimedia.be

@dimi_z