WIKIMEDIA
FOUNDATION

May 24, 2023

Re: Arcom queries regarding "hateful content" moderation

Dear Arcom,

We are pleased to offer our responses to your recent queries, regarding moderation of hateful content. Thank you again for the extended deadline.

In this document, your questions (translated to English) are presented in bold, followed by our observations.

REPORTS

1. **"Have you encountered any challenges including online hate motives in the reporting mechanism available to users of your service? Thank you for specifying."**

Wikimedia Foundation, Inc. (hereafter "WMF") response: We have not encountered such challenges - our standard mechanism is flexible enough.

Challenges, if any, are more likely to arise during the subsequent analysis of the content that is reported - rather than in adapting the reporting mechanism. As ARCOM will appreciate, it is challenging to determine, with precision:

- the nature of the content (and intent of the uploader(s), and those subsequently interacting with it);
- how it is qualified, whether under applicable laws, or applicable rules of the website; and
- what a proportionate, fundamental rights -respecting response should be.

Of course, there is nothing new about that challenge.

2. **"Among the reports you received concerning the French version of your service in 2022 and which related to a reason corresponding to the definition of hateful content within the meaning of article 6-4 of law n ° 2004-575 of June 21, 2004 for confidence in the digital economy, how many of them came from:**
    1) **Platform users**
    2) **Trusted flaggers you work with if applicable**
    3) **Public authorities (administrative or judicial)"**

WMF response: We do not recall any such content (i.e., "in-scope content") being reported to us during 2022, in relation to any French-language version of the platforms we host.

As Arcom likely appreciates, hate speech would generally not be tolerated (i.e., would be deleted/sanctioned) by users, directly and autonomously, and this could explain the low reporting rate (in effect, there would be nothing to "escalate" to the Wikimedia Foundation). Also (and while this is also a broad generalization, to which exceptions can no doubt be found), our projects do not tend to welcome the same type of persons that might use traditional social media as a podium for hate speech.[1]

Note that it is not entirely clear to us exactly what content corresponds to the definition of hateful content within the meaning of article 6-4 of law n ° 2004-575 of June 21, 2004 for confidence in the digital economy. Firstly: that defined term ("contenu(s) haineux") does not seem to be used in the cited legal provision. Secondly: the legislator appears to have been *extremely* inclusive, and indirect, when including content into the scope of that provision. When we have attempted to follow the many references in that provision, through to their underlying statutory provisions, then it appeared to us that although most of them unquestionably concern hate speech/"hateful" content, this is less clear for some of them, such as bestiality - which turn more on immorality/obscenity than "hate."

We have not found a comprehensive definition on Arcom's website, either.

Accordingly, while our answers above (and those elsewhere in this document) are given in good faith, we still think it is appropriate to offer a general caveat regarding our underlying uncertainty.

EXERCISE OF MODERATION

3. **"How many moderation actions relating to content corresponding to the definition of hateful content within the meaning of the aforementioned article 6-4 did you carry out in 2022, broken down by type[2] ?"**

WMF response: by conscious design and as a result of the way the Wikimedia projects are run, it is rare for us (the Wikimedia Foundation, as platform operator) to take moderation actions. We call these "Office Actions".

We do not recall taking any Office Actions in 2022 in relation to in-scope content.

4. **"How many actions did you take in 2022 against users abusing reporting tools, broken down by type?"**

WMF response: We do not recall taking action against users abusing reporting tools, in relation to in-scope content, in 2022.

Of course, we added the senders of obvious commercial spam (i.e., advertisements) and malicious emails (e.g., phishing) to email blocklists. These normally do not come from users, and are quite different from content reports.

[1] See, e.g., https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Neutralit%C3%A9_de_point_de_vue

[2] "In particular (non-exhaustive list): warning, withdrawal of content, access restriction measures (by age, by geographical area), measure to limit the visibility of content or an account, demonetization measure, suspension or deletion of an account."

5. **"What is the proportion of the actions referred to in questions 3 on the one hand, and 4 on the other hand, taken after human decision (i.e. actions which do not result solely from an automatic process)?"**

WMF response: Not applicable (see answers to Q3 and Q4).

Note that Office Actions are never automatically taken, and spammers are also added to blocklists manually.

6. **"Following the actions referred to in questions 3 on the one hand, and 4 on the other hand, what is the rate of internal appeals from users accessing your services from France, and what were the results of these (percentage upholding original decision, percentage overturning original decision)?"**

WMF response: Not applicable (see answers to Q3 and Q4).

7. **"What are the number, location and working language(s) of the people assigned to the processing of reports and appeals from users of the French version of your online hate service (provide information valid for the month of December 2022)?"**

WMF response: A geographically distributed team of around 45-50 persons was available to process reports of this nature. However, for many, this is often not their core/routine activity - they would be involved only as substitute/supplemental resources if the "front line" needed the extra support. We expect that most matters of this nature are handled by 2-6 people (approximately), but discussion with the wider team is common.[3]

Our team is inherently multilingual, with staff able to cover the *extremely* wide range of languages in which our projects are available. They do this due to native fluency, schooling, machine translation, and/or with the help of other persons within the Wikimedia Foundation who can speak the requisite language.

Team members are not specifically assigned to specific language versions of our projects - this would be inefficient, especially since a report could (for example) highlight content that is found both on the French and English versions of Wikipedia. We instead have a central intake for reports, and then triage them to teammates according to their availability and suitability.

Note that the Wikimedia *community* operates a volunteer-run, self-selecting community helpdesk, the Volunteer Response Team (VRT), which can (for example) be contacted via info-fr@wikimedia.org. This is an effective, self-governing, and highly multilingual system

[3] As Arcom will observe, current conditions - e.g. the DSA's fundamental preservation of a "notice and takedown" (i.e. reactive) model, rather than veering into proactive moderation - allow us to closely examine and debate the merits of takedown requests, and the best way to deal with them. Wherever possible, this includes subsidiarity to the wider community, or at least direct engagement with it. This is critical to fundamental rights protection, and to the continuing success of the Wikimedia projects (which attract and retain committed volunteers because they are *empowered*, rather than having policies, duties and decisions routinely imposed on them). This may seem artisanal, but its effectiveness is proven beyond doubt. Proactive moderation obligations - in Europe or elsewhere - threaten this, and would compromise community autonomy.

that users, trusted flaggers, etc., can use to ask questions, report problems, etc., that are then dealt with by fellow users of the websites. For reasons of privacy, resource efficiency and respect for those users' autonomy, the Wikimedia Foundation does not actively monitor VRT activity (in much the same way that, we imagine, Facebook does not closely supervise or keep statistics on messages received, and actions taken by civilian "admins" of individual Facebook Groups). We do not have data on (for example) VRT members' languages or location.

COOPERATION WITH PUBLIC AUTHORITIES

8. **"In 2022, how many suspicions of offense were the subject of a transmission from your part to the competent public authorities, in particular to the public prosecutor, and for what reasons (give, if applicable, the only main reason)"**

WMF response: It is unclear to us whether the question is asking about in-scope content, or more broadly; and, whether it is asking about reports made to French authorities, or more broadly.

9. **"What procedures and human and technological resources have you implemented to respond to administrative or judicial authorities as soon as possible?"**

WMF response: Our legal@wikimedia.org inbox is monitored by a geographically-distributed team, and is the preferred destination for reports; however, it may be in practice even quicker to contact one or more other users of the website, either by discussing the issue with them directly (e.g., here), or emailing VRT (e.g., info-fr@wikimedia.org).

If there is an immediate threat to life, our emergency@wikimedia.org address can be used, including by relevant law enforcement agencies.
Terrorist and Violent Extremist Content (TVEC) can also, in urgent cases, be notified as per the instructions here.

PREPARING FOR THE IMPLEMENTATION OF THE RSN/DSA

10. **"What actions has your organization taken to prepare the service for the application of the European regulation on digital services (RSN/DSA)?"**

WMF response: For an overview, please see https://diff.wikimedia.org/2023/05/04/wikipedia-is-now-a-very-large-online-platform-vlop-under-new-european-union-rules-heres-what-that-means-for-wikimedians-and-readers/

We expect that other Digital Services Coordinators would have a similar interest to Arcom, and we are considering ways in which we can efficiently provide transparency here. Hypothetically, if we organized a seminar/"open day," later this year, to talk about the DSA and its impact for us, and how we are preparing - would Arcom be keen to attend?

11. **"Have you faced any difficulties or issues (of all kinds: understanding, method, calculation of indicators, modification of the product, etc.) in the implementation of the aforementioned article 6-4 which would be likely to new encountered in the context of that of the RSN, or which would have allowed you to anticipate the latter? Which ones?"**

WMF response: While we do anticipate some DSA implementation challenges, these are not specifically informed by our narrow experience with art. 6-4.

12. **"Do you work with third parties based in France that you recognize as trusted online hate flaggers? If yes, which ones ? If so, why did you choose to collaborate with them?"**

WMF response: No.

Note, also, that our entire "movement" ethos - indeed, the strength of the systems that ensure the success of these projects - is based around individual citizens being empowered to create, curate, and govern platforms.

While a Wikimedia Foundation partnership with trusted "hate speech" flaggers is not impossible, it would still be a considerable challenge. One reason is that it could be seen, by some, as contrary to that ethos, if the Wikimedia Foundation specifically seeks out and allies with local/national special interest groups in order to "Office Action" in-scope content on our projects. Doing so bypasses/undermines community autonomy, and would also precipitate legitimate questions: Why one "trusted flagger" agency, but not another? Why from France, but not from every other country in the world? Etc.

Flaggers (of any type) are of course welcome to report content to us (and we are aware of our DSA obligations in this regard), but it seems less likely that - unless our own communities encourage us - we would actively seek to partner with any specific "flaggers."

Some other forms of "collaboration" could perhaps be envisaged: e.g., providing flaggers (of any kind) with educational materials explaining how they can efficiently raise concerns (for instance, explaining the difference between on-wiki reporting, emailing VRT, and emailing the Wikimedia Foundation).

This is not a concrete and definitive view on this issue. We will ultimately be guided by the Wikimedia community's own (evolving) wishes and suggestions, and our own assessment of systemic risks and mitigations (amongst other DSA obligations).

* * *