# Wikimedia Foundation comments on UNESCO Guidelines for regulating digital platforms 2.0[1]

08 March 2023

## *Section One: General comments on the overall draft 2.0 Guidelines*

We are pleased that UNESCO has addressed many of the comments that the Wikimedia Foundation and various other organizations submitted about draft 1.0. The improvements we see in draft 2.0 reflect this commitment to a multistakeholder approach. We appreciate that many recommendations to draft 2.0 align with our values, mission, policies, systems, processes, and public policy positions.[2] In particular, we applaud: the inclusion of the recognition of the role of all relevant stakeholders in maintaining an enabling environment for freedom of expression and the right to information; the emphasis placed on the State's duty to protect freedom of expression and refrain from imposing disproportionate measures; and, the improvement of the regulatory system recommendations, in which we see reinforced obligations of transparency and accountability, among others.

Despite the positive changes, the guidelines still remain focused on the model of large social media platforms, and do not consider potential impacts to decentralized community-led governance and content moderation models such as that of Wikipedia and the other Wikimedia projects. In this model, the information is added, organized, and edited by a decentralized community of volunteers who engage in open debate to reach consensus around content decisions and policies, without interference from the Foundation. This model is empowered by existing protections from liability for user-generated content, and can be threatened by restrictive regulations that require platforms to make more top-down decisions. The Wikimedia volunteer community experiences the consequences of technology regulation everyday, facing both unintended consequences of misguided policies alongside harms caused by laws that are inconsistent with human rights standards.

Another general concern we have is that the guidelines impose overly restrictive and/or burdensome compliance requirements that could be difficult for small or nonprofit platforms.

[1] https://unesdoc.unesco.org/ark:/48223/pf0000384031.locale=en

[2] The Wikimedia Foundation is the nonprofit organization that hosts Wikipedia and other free knowledge projects. The vision of our free knowledge movement is a world in which every single human being can freely share in the sum of all knowledge. To this end, we support a community of more than 300,000 volunteers around the world, who contribute to the Wikimedia projects by adding, editing, and verifying content across 300 languages, all for free and without ads.

Some recommendations would be particularly onerous, especially if each government or jurisdiction sets divergent standards for platform operators such as the Foundation.

With these concerns in mind, the Foundation has worked to comment on draft 2.0 of the guidelines in a manner that reflects and highlights both the Wikimedia model and the experience of our global communities and volunteers. We trust that this stage of the process will correct the weaknesses that we continue to underline in the guidelines, and that the final result will translate into a public policy orientation that supports a public interest, community-led internet, and also reinforces a broader digital ecosystem that enables the enjoyment of human rights.

## *Section Two: General comments on the drafting process*

We thank UNESCO for launching a consultation process for draft 2.0 and subsequent drafts. This process could be more effective if UNESCO publishes a clearer work plan, in which it makes public the consultation dates to follow as well as the different forms of participation. In addition, future consultations would benefit from adopting a multilingual and multigeographical approach.

## *Section Three: The objective of the Guidelines*

### Comments on paragraph 10(a)

> *10.a The scope of these Guidelines includes digital platforms that allow users to disseminate content to the wider public, including social media networks, messaging apps, search engines, app stores, and content-sharing platforms. Bodies in the regulatory system should define which digital platform services are in scope, and also identify the platforms by their size, reach, and the services they provide, as well as features such as whether they are for-profit or non-profit, and if they are centrally managed or if they are federated or distributed platforms.*

Regarding the scope of the recommendations, it still seems that the approach remains relatively broad. A blanket approach to regulation will not be practical given the diversity of platform models and communities that exist online today. The guidelines should, therefore, avoid overgeneralization by ensuring that any norm does not unduly restrict freedom of information, including access to information and the ability to produce and share information.

Consequently, the guidelines should caution that when regulatory bodies define digital platforms considering the different features mentioned in the paragraph, it will be crucial to contemplate whether different standards or caveats for platform type should be included in enforcing the guidelines.

## *Section Four: The Regulatory System*

Comments on paragraph 46a

> *46. To fulfill the goal of regulation, the regulatory system should have the following powers:*
>
> *a. Establish Standardized reporting mechanisms and formats. Ideally,reports should be made annually in a machine-readable format.*

We are concerned about how an excessive application of these powers may disrupt the editing processes in the Wikimedia projects. The content on the projects is written, curated, and moderated by a community of volunteers who also collectively decides on the content policies they will be enforcing. This community governance model has grown more complex and effective over time, enabled by liability protections for platforms that have allowed the Foundation to support these efforts without having to assert influence over actual content. Applied too broadly, the regulatory requirements imagined in the guidelines threaten to disrupt this carefully developed system.

In particular, the guidelines should caution against any mandates committing platforms to specific moderation practices, and should encourage remedies for violations that do not involve removal of liability protections.

Comments on paragraph 46b

> *46. To fulfill the goal of regulation, the regulatory system should have the following powers:*
>
> *b. Commission off-cycle reports if there are exigent emergencies, such as a sudden information crisis (such as that brought about by the COVID-19 pandemic) or a specific event which creates vulnerabilities (for example,elections or protests).*

Unreasonable application of off-cycle reporting may have the inadvertent effect of burdening smaller, not-for-profit platforms with excessive demands. The justification for off-cycle reporting is envisioned quite broadly at the moment, including examples that can both last for an indefinite amount of time—such as the COVID-19 pandemic—and that would occur with substantial frequency for a global platform—such as elections and protests. While it is important to research the impact of such events on platforms, it is also important to consider that producing even cyclical reports may already be burdensome for platforms that are operating at a lower cost or with fewer personnel than the largest social media websites. In this case, the addition of off-cycle reports based on broadly defined “emergency” situations could prove entirely overwhelming and, in some cases, could be used in bad faith to suppress certain types of content.

If the recommendation remains, we suggest that it include a much more narrow definition of emergency, and potentially limit the platforms from which such reports can be demanded.

Comments on paragraph 46e

> *46. To fulfill the goal of regulation, the regulatory system should have the following powers:*
>
> *e. Establish a complaints process that offers users redress should a platform not deal with their complaint fairly, based on the needs of the public they serve, the enforcement powers they have in law, their resources,and their local legal context.*

We wonder what it means when “a platform does not handle a user complaint fairly.” We worry that the vagueness of this phrase could lead to excessive and inappropriate intervention of regulatory bodies. The Foundation empowers the Wikimedia projects’ volunteer community to develop and implement their own conflict resolution policies and processes, which are subject to transparent and participatory debate, and have been shown to be largely effective.

In the interest of protecting this model, we recommend defining more clearly under what circumstances a complaint process could be triggered: for example, once a user has exhausted the platforms’ internal mechanisms for resolving complaints, and the policies for resolving disputes have been properly applied.

## *Section Five: Responsibility of digital platforms*

### *Principle 1. Platforms respect human rights in content moderation, and curation*

### *Content moderation and curation policies and practices*

Comments on paragraph 54

> *54. Content moderation and curation structures and processes should be applied consistently and fairly across all regions and languages.*

It is worth noting that Wikimedia projects are not organized by markets and national jurisdictions, but by language communities. One of the great things about Wikipedia and the other projects is that different language communities can set their own rules to address issues that may not even be relevant in other communities and/or groups of people speaking the same language. Indeed, we have observed that this community-led content moderation model is more effective and resilient against harmful information. However, while Wikipedia and the other projects have clear policy guidelines supporting consistency, this model does not allow for perfect consistency or fairness in moderation, among other things, because what is considered fair in one region may not be considered fair in another.

Instead, we recommended that the paragraph focus on guiding moderation processes to effectively apply the policies and processes adopted by the various platforms, rather than require consistency and fairness, as this seems to be a challenging standard to measure and achieve.

### *Human content moderation*

Comments on paragraph 60

> 60. *Human content moderators should be adequately trained, sufficiently staffed, fluent in the language concerned, vetted,and psychologically supported. Platforms should further put in place well-funded and -staffed support programmes for content moderators to minimize harm caused to them through their reoccurring exposure to violent or disturbing content while at work.Where possible and when it would not negatively impact human rights or undermine adherence to international norms for freedom of expression, human moderation of content should take place in the country or region where it is published to ensure close awareness of local or national events and contexts, as well as fluency in the language concerned.*

We commend the recommendation's aims to improve human moderators' well-being and work conditions. However, we are concerned that its wording is not sufficiently nuanced to distinguish between volunteer moderators—i.e., individuals who contribute and improve information online on their own time and/or motivated by their own initiative—and paid moderators. Failure to recognize the diversity of the digital ecosystem when designing regulatory recommendations, as noted above, can be disruptive to community-led content moderation models such as that of Wikipedia and the other Wikimedia projects.

Furthermore, asking that content moderation be done in the national jurisdiction or regions where it is published is a risky request. The recommendation, as drafted, presents significant security problems for human moderators. For example, volunteer editors have occasionally become victims of legal prosecution or political reprisal for edits made on Wikipedia that were considered inconvenient by national authorities.

Therefore, we believe UNESCO must go back to the drawing board and evaluate the implications of this recommendation considering the diversity of existing content moderation models and the risks of the localization requirement.

### *Data access for research purposes*

Comments on paragraph 72

> 72. *Digital platforms should provide access to non-personal data and anonymised data for vetted researchers that is necessary for them to undertake research on content to understand the impact of digital platforms. This data should be made available through automated means,such as application programming interfaces (APIs),or other open and accessible technical solutions allowing the analysis of said data.*

We are concerned that this paragraph uses the concept of anonymization. Studies have shown that successfully anonymizing data is impossible for any complex dataset. The most advanced technological efforts have yet to ensure complete data deidentification, which can almost certainly be re-identified by cross-referencing with other readily available datasets. A more useful concept is pseudonymization.

Accordingly, considering the technical limitations upon complete anonymization, we recommend that the requirement for anonymization be replaced by pseudonymization and aggregated or de-identified data. In addition, aggregated and de-identified data are accepted practices to remove identifying factors from the data used for further research and studies.

## *Principle 3. Platforms empower users*

### *Media and information literacy*

Comments on paragraph 77

> 77. *When reporting to the regulatory system, platforms should demonstrate their overall strategy related to media and information literacy and the actions they have taken to advance on it.There should be a specific focus inside the digital platform on how to improve the digital literacy of its users, with thought given to this in all product development teams. The digital platform should consider how any product or service impacts user behavior beyond the aim of user acquisition or engagement.*

It is misguided for the guidelines to recommend that platforms demonstrate their general strategy on media and information literacy to regulatory systems. Indeed, platforms should play an essential role in this regard, but it is not their duty to carry out literacy activities. Instead, the recommendation should invite platforms to support and collaborate with the media and information literacy strategies and actions of governments and other stakeholders such as civil society.

Even so, we would like to emphasize that both the Foundation and Wikimedia volunteer communities develop and support various media and information literacy initiatives so as to build the critical skills necessary to access, understand, create and/or participate in content using digital media.

## *Principle 5. Platforms conduct human rights due diligence*

### *Human rights safeguards and risk assessment*

Comments on paragraph 92

> 92. *Digital platforms should be able to demonstrate to the regulatory system the system or process they have established to ensure user safety while also respecting freedom of expression, access to information,and other human rights.*

The implementation of these guidelines should not result in governments imposing immense regulatory burdens on digital platforms, which would likely vary from one national jurisdiction to another. The time and resources needed to comply with such disparate legal requirements would harm the ability of community-led, nonprofit platforms to serve those communities, since they would create a regulatory environment in which it is legally and financially unfeasible to operate. Instead, the guidelines should encourage platforms to adopt human rights due diligence practices tailored to the unique context of their model and encourage transparency into these practices so that they are visible to the general public.

Comments on paragraph 94

> 94. *Apart from periodic assessments, risk assessments should also be undertaken:*
>     a. *Prior to any significant design changes, major decisions,changes in operations, or new activity or relationships;*
>     b. *To protect the exercise of speech by minority users and for the protection of journalists and human rights defenders;*
>     c. *To help protect the integrity of electoral processes;*
>     d. *In response to emergencies, crises,or conflict or significant change in the operating environment.*

We find this recommendation to be overly prescriptive. As currently drafted, particularly concerning subparagraph c, this recommendation could force platforms to conduct risk assessments for every election, no matter the level (national, sub-national, municipal, etcetera), in every jurisdiction where they operate, which would be resource-intensive and financially burdensome for community-led platforms.

Instead, this section should encourage platforms to consider these issues in a less prescriptive manner, which would allow community-led, nonprofit platforms to identify and mitigate such human rights risks in ways consistent and feasible within their own context. We recommend modifying the text as follows:

> 94. Apart from periodic assessments, **due diligence** should take into account [...]

Comments on paragraph 96

> *96. Platforms can create spaces to listen, engage,and involve victims, their representatives, and users from minorities to identify and counter illegal content and content that risks significant harm to democracy and the enjoyment of human rights, to identify opportunities and systemic risks in order to then promote solutions and improve their policies. Consideration should be given to the creation of specific products that enable all relevant groups to actively participate in the strengthening of counter-narratives against hate speech.*

Creating new, legally-mandated products are burdensome and resource-intensive for community-led, nonprofit platforms that instead develop products based on community needs and priorities. Instead, we propose to modify the second sentence of this paragraph as follows:

> 96. [...] Consideration should be given to the creation of specific **spaces or opportunities** that enable all relevant groups to actively participate [...]

### *Specific measures to fight gendered disinformation and online gender-based violence*

Comments on paragraph 98a

> *98. To fight gendered disinformation and online gender-based violence,digital platforms should:*
>
> *a. Conduct annual human rights and gender impact assessments,including algorithmic approaches to gender-specific risk assessment, with a view to identify the systemic risks to women and girls and to adjust regulations and practices to mitigate such risks more effectively.*

Annual “human rights and gender impact assessments” may not be feasible due to capacity and resources constraints, especially for community-led nonprofit platforms with limited financial resources. Instead, impacts on women—as well as transgender, nonbinary, and gender-fluid people (a fault that we note in the guidelines is that it does not acknowledge the existence of these people, who are all disproportionately impacted by implicit and explicit sexual and/or gender discrimination, especially online)—should be analyzed and included in broader periodic human rights risk assessments. These impacts should also be analyzed in ongoing human rights due diligence efforts that platforms carry out in a more sustained cadence throughout their operations to understand the potential human rights impacts of new products, tools, business arrangements, and others as such questions arise.

### *Specific measures for the integrity of elections*

Comments on paragraphs 99-101

> *99. While electoral bodies and administrators need to ensure that the integrity of the electoral process is not affected or undermined by disinformation and other harmful practices, digital platforms should have a specific risk assessment process for any election event. Such risk assessments should also consider the users,the level of influence that advertising messages may have on them,and the potential harm that may come out of such messages if used against specific groups, such as minorities or other vulnerable groups.*
>
> *100. Within the assessment, digital platforms should review whether political advertising products, policies, or practices arbitrarily limit access to information for citizens,voters, or the media,or the ability of candidates or parties to deliver their messages.*
>
> *101. Digital platforms should also engage with the election's administrator/regulator (and relevant civil society groups), if one exists, prior to and during an election to establish a means of communication if concerns are raised by the administrator or by users/voters.Engagement with other relevant independent regulators maybe necessary according to the particular circumstances of each jurisdiction.*

As we explained before, this recommendation could require any platform to perform a risk assessment for each election at every level (national, sub-national, municipal, etc.) in every jurisdiction in which they operate. This would pose a considerable financial challenge, to the point of being unfeasible, for platforms hosted by nonprofit organizations like the Foundation. A better way to approach this issue is for the guidelines to invite platforms to understand the risks around elections by establishing processes or mechanisms in which they listen to and engage with civil society organizations and other relevant stakeholders in these jurisdictions.