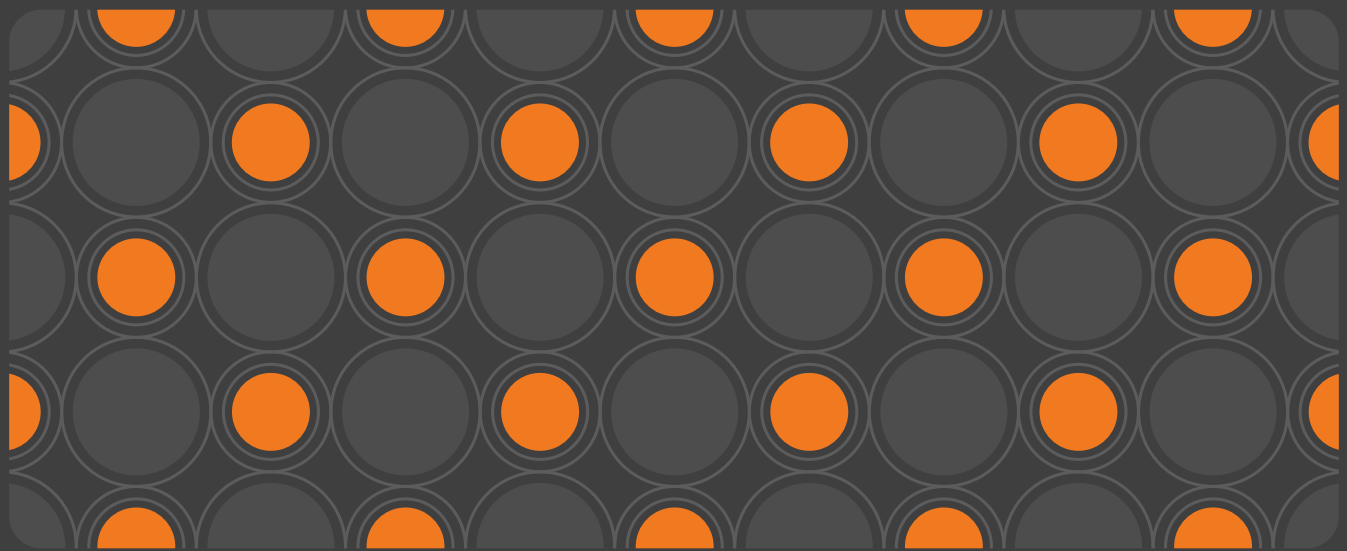


2020
JULY



Assessing the Human Rights Impacts of Wikimedia Free Knowledge Projects

ARTICLE ONE



Table of contents

Foreword from the Wikimedia Foundation	3
I. Executive Summary	7
II. Introduction	13
III. Scope & Methodology of the Assessment	16
IV. Salient Human Rights Risks	19
1. Harmful Content	21
A. Attacks on Individuals Profiled	22
B. Misrepresentation of Historical Facts	23
C. Project Capture	24
D. Dangerous Content	24
2. Harassment	27
A. Harassment within the Volunteer Community	28
B. Harassment of Foundation Staff	31
3. Government Surveillance & Censorship	34
A. Online Surveillance of Wikimedia Volunteers and Readers	35
B. Requests for User Data	35
C. Government Censorship	36
4. Risks to Child Rights	38
A. Privacy and Reputation Risks	39
B. Exposure to Harmful Content	39
C. Child Sexual Exploitation Material	39
D. Harmful Contact	40
5. Limitations on Knowledge Equity	42
A. Gender Equity	43
B. Racial & Ethnic Diversity	44
C. Accessibility	44
D. Knowledge Equity of the Global South	45
V. Conclusion	47
Appendix	50



Foreword from the Wikimedia Foundation

July 2022

The Wikimedia Foundation is the global nonprofit that makes knowledge free and accessible to everyone around the world by hosting and supporting volunteer-run projects. This includes Wikipedia, which currently offers over 55 million articles across 300 languages, all for free and without commercial advertisements. A worldwide community of volunteers contributes, edits, and moderates content across Wikimedia projects based on a robust set of standards and norms volunteers have created and regularly enforce.

Wikipedia and Wikimedia projects occupy a unique space in today's internet ecosystem: our projects leverage a decentralized, volunteer-governed model to create reliable and neutral knowledge for the public. Together, volunteers who edit the content on Wikimedia projects prioritize accuracy and verifiability over the virality of content. This has enabled Wikipedia and Wikimedia projects to become widely-trusted sources of information to people around the world.

Wikimedia Foundation and Human Rights

The Foundation believes knowledge is a human right. Wikimedia projects provide channels and platforms through which everyone, everywhere, has the right to share and access knowledge freely. Free knowledge, along with the fundamental right to freedom of expression, empowers people to exercise many other rights enshrined in the Universal Declaration of Human Rights, including the rights to education, artistic expression, economic advancement, and political participation.

As host of these projects, the Foundation is therefore committed to respect the human rights of all those who seek, receive, and impart knowledge on Wikimedia projects.

About this Human Rights Impact Assessment Report

This Human Rights Impact Assessment (HRIA) report reflects the Foundation's focus on protecting and advancing the human rights of those who use and contribute to Wikimedia projects. It was carried out in 2020 by Article One, a specialized strategy and management consultancy with expertise in human rights, responsible innovation, and sustainability. The purpose of the assessment was to better understand whether and how Wikimedia projects, platforms, and activities might cause inadvertent human rights harms to Wikimedia volunteers, Foundation employees, readers, and others affected directly or indirectly by free knowledge projects. Only by identifying and understanding how possible human rights harms occur can the Foundation work to mitigate and prevent them in the future.

The assessment was completed by [Article One](#) and submitted to the Foundation in July 2020. Unfortunately, due to capacity constraints and disruptions caused by the COVID-19 pandemic, the publication of this report has been significantly delayed. It was also important to take the time—working with colleagues across the Foundation—to ensure that the public version of this report would not expose volunteers, Foundation employees, or any other people who interact with Wikimedia projects to additional harm. Furthermore, the Foundation has made investments in its ability to respond meaningfully to the recommendations outlined in the report (more details are outlined on the following page).

The Foundation and Article One partnered to carry out a comprehensive review of the HRIA report in order to identify any content that could either endanger individuals or empower malicious actors to misuse Wikimedia projects. When possible, efforts were made to revise or generalize such content. Some content was removed completely when the risks outweighed the benefits of publishing it. This HRIA report is, therefore, a redacted version of the original.

Wikimedia Foundation Investments in Human Rights Work

Since Article One submitted the report to the Foundation in mid-2020, the Foundation has invested in its capacity to respond to recommendations in the report and, where possible, move forward on recommendations that were already aligned with organizational priorities and other project roadmaps.

Key steps include:

Developing a Universal Code of Conduct: In 2020, the Foundation began the process of codeveloping with volunteer communities a [Universal Code of Conduct](#) for Wikimedia platforms. This was a recommendation of the Wikimedia Movement Strategy conversations between Foundation staff and volunteers, which coalesced in May of that same year. It also aligns with a recommendation from this report. This policy outlines basic standards for acceptable behaviors on Wikimedia projects, without tolerance for harassing behaviors. Initial research and consultations with Wikimedia communities occurred between June and December 2020. The Universal Code of Conduct was approved by the Board of Trustees in February 2021. Enforcement guidelines for this policy are currently under development.

Recruiting Human Rights Expertise: The Foundation hired a Human Rights Lead in January 2021 to build a team and program dedicated to upholding and defending the safe contribution of the movement's volunteers. Additionally, the Foundation hired a Vice President for Global Advocacy, with deep expertise in policy, human rights, and digital authoritarianism, in October 2021 to lead the organization's efforts to promote policies that advance an online ecosystem that upholds human rights. The Foundation also created the position of Senior Human Rights Advocacy Manager in early 2022 to help coordinate the tactical implementation of commitments made in the Foundation's human rights policy, in addition to managing continued work toward addressing issues raised by the initial HRIA report. These investments in staffing have better equipped the Foundation to be more forward-looking when responding to, respecting, and advancing human rights.

Establishing Human Rights Leadership: The Foundation established in May 2021 a Human Rights Steering Committee, composed of senior-level leaders from across the Foundation, to create an integrated, organization-wide implementation of the Foundation's Human Rights Policy and due diligence practices

Strengthening Human Rights Resources for Volunteers: Starting in 2021 and deepening into 2022, newly hired Foundation staff with human rights expertise have:

- ▶ Worked to establish better channels and mechanisms for volunteers and affiliates to report human rights concerns to Foundation staff who are able to take action and respond to threats;
- ▶ Trained at-risk volunteers on digital security skills and best practices, with one-on-one consultations and the development of a multilingual toolkit, so that these volunteers can better protect their privacy and safety online;
- ▶ Advanced work with the [Voices Under Threat](#) program, which aims to support volunteers contributing to our projects in challenging or high-risk regions and predates this HRIA report. In the past two years, we have developed more multilingual resources for volunteers and developed regular office hours dedicated to digital security for communities;
- ▶ Partnered with global human rights organizations to build out local, regional, and international capacity to support at-risk volunteers and communities; and,
- ▶ Developed the Foundation's first Crisis Response Protocol to provide organization-wide support to threatened volunteers and to coordinate efforts across departments.

Approving a Human Rights Policy: The Foundation developed a [Human Rights Policy](#), which was approved by the Board of Trustees in December 2021. This policy commits the Foundation to four key activities necessary for addressing and mitigating human rights risks, including:

- ▶ Conducting ongoing human rights due diligence;
- ▶ Tracking and publicly reporting on our efforts to meet our human rights commitment;
- ▶ Working with partners, the private sector, and governments to advance and uphold respect for human rights; and,
- ▶ Providing access to effective remedies when human rights harms have occurred.

Continuing Due Diligence: Following the submission of the HRIA report, the Foundation commissioned two additional human rights impact assessments, including a child rights impact assessment in April 2022, and a product-level human rights impact assessment in May 2022. The Foundation plans to publish these reports as well.

Engaging with Wikimedia Communities: The Foundation launched a [series of public and private dialogues](#) in May 2022 across various communities of volunteers and Foundation staff in order to better understand the needs of community members facing human rights challenges as well as how the implementation of the Human Rights Policy can support those needs. Dialogue with volunteers remains ongoing.

Mitigating the Impacts of Disinformation: Given the broad potential impact disinformation can have on freedom of expression throughout Wikimedia projects, the Foundation has undertaken a series of interventions to address and mitigate this risk, including:

- ▶ **Cultivating Internal Expertise:** The organization has cultivated a team of dedicated disinformation specialists who support communities in researching, identifying, and addressing disinformation on Wikimedia projects. An Anti-Disinformation Strategy Lead was also hired in April 2022 to coordinate internal and external efforts to counter disinformation and propose effective policies on these issues.
- ▶ **Establishing Event-Specific Monitoring:** To address risk around significant political events that attract disinformation, the Foundation has also dedicated additional resources to monitoring for such content as needed. For example, the Foundation established a cross-functional [Disinformation Task Force](#) to support volunteers in evaluating and responding to any disinformation attempts during the November 2020 US presidential election. This effort also extends to the commissioning of reports to better understand the impacts of disinformation campaigns, including [one organized on Croatian Wikipedia](#), which was published in June 2021.
- ▶ **Developing an Institutional Strategy:** Informed by these experiences, the Foundation is working to develop an institutional strategy to address disinformation on Wikimedia projects, and has strengthened its advocacy efforts to promote public policies addressing disinformation online.
- ▶ **Expanding Dedicated Research and Tool Building:** The Foundation's Research and Product teams are engaged in expanding dedicated research on disinformation, also in cooperation with volunteer communities, in order to better understand the issue on the platform and provide helpful tools for content moderation.

These key steps have laid the groundwork for the Foundation to scale up its work on addressing and mitigating human rights risks in the coming years, and reinforced human rights values in the Foundation's DNA.

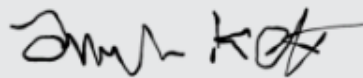
It is important to acknowledge, however, that much remains to be done to address the risks identified in this assessment and to live up to the commitments outlined in our Human Rights Policy. The Wikimedia Foundation must focus on establishing, implementing, and disclosing our human rights policies and related impact assessments. To meet the commitments of our new Human Rights Policy, the Foundation is committed to carrying out human rights due diligence on an ongoing basis and being transparent about the findings of those processes.

Looking Forward

We hope that our inaugural HRIA report will support all stakeholders in the Wikimedia movement to better understand the human rights risks and threats that we jointly face, and thereby inform the work required to address those risks. In the coming months and years, the Foundation will continue to work with volunteers to unpack the findings of this HRIA report and to determine how we can best move forward together to advance human rights for Wikimedia projects.

We further hope that this HRIA report will be instructive to other non-profit organizations that operate online platforms for a public interest purpose. Regardless of business model, online platforms that support human rights standards and values should protect and respect the human rights of their contributors, readers, audiences, and other communities whose lives are affected by their operation. This responsibility necessarily includes a focus on due diligence and accountability to affected communities.

Finally, we also hope that the publication of this report, combined with the steps we have taken in the two years since the HRIA report was completed, will be seen as evidence of our sincere commitment to genuine and honest—even if sometimes difficult—dialogue about how Wikimedia’s free knowledge projects can truly fulfill the movement’s vision of a world in which everyone, everywhere, can share knowledge freely.



Amanda Keton
General Counsel
Wikimedia Foundation
July 12, 2022

*Assessing the Human Rights Impacts
of Wikimedia Free Knowledge Projects*



Executive Summary



I. Executive Summary

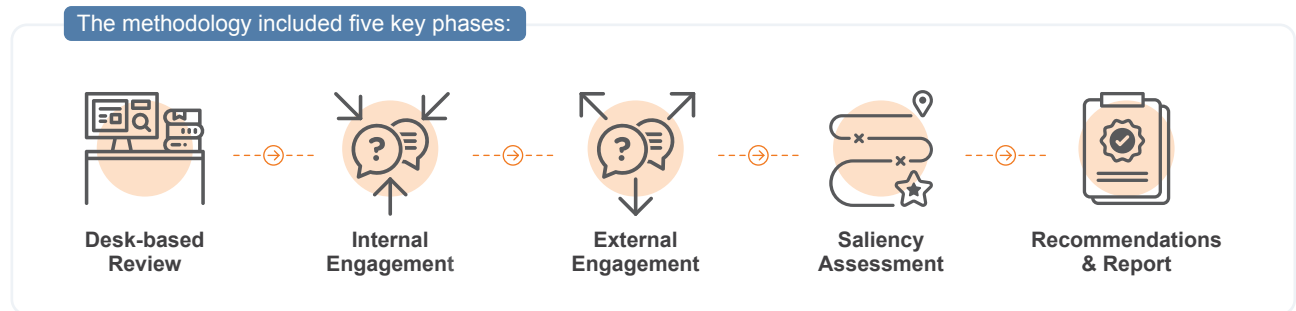
The Wikimedia Foundation (the Foundation) hosts free knowledge projects and protects *“the values and policies that allow free knowledge to thrive.”*¹ This vision advances the fundamental right to access and impart information globally. To ensure that Wikimedia projects continue to advance respect for human rights, the Foundation has commissioned this human rights impact assessment (HRIA) with the goal to:

- 1 Surface salient human rights risks across its free knowledge projects;
- 2 Mitigate actual and potential risks related to its projects around the world, including avoiding harm to rightsholders²;
- 3 Support the Foundation in becoming a member of the Global Network Initiative (GNI); and
- 4 Better align with stakeholder expectations regarding ongoing human rights due diligence.

International human rights standards provide a powerful framework for the Foundation to understand and address the risks associated with its free knowledge projects. Human rights standards, including the International Bill of Human Rights³ and the UN Guiding Principles on Business and Human Rights (UNGPs)⁴, offer a global, broadly accepted set of values to inform the Foundation’s approach to surfacing and mitigating human rights risks.⁵ Indeed, human rights are inherent, inalienable, interdependent, and indivisible: they cannot be granted or taken away, the enjoyment of one right affects the enjoyment of others and as such they must all be respected.

▣ Scope & Methodology of the Assessment

The HRIA follows guidance from the UNGPs and the GNI Principles⁶ and is based on Article One’s award-winning methodology for and experience in conducting human rights impact assessments. While the assessment focused on risks associated with Wikimedia free knowledge projects, we have outlined where risks had a direct impact on Foundation staff.



Article One conducted a desk review of public and private information on the Foundation and Wikimedia projects, including news reports, research assessments and confidential information shared with Article One under a non-disclosure agreement (NDA). We supplemented the desk review with interviews of 17 Foundation staff and six leading external experts.⁷ The findings were then validated with six volunteers from across the globe. Human rights risks were mapped to the Universal Declaration of Human Rights (UDHR) and the Convention on the Rights of the Child (CRC). In addition, Article One assessed Wikimedia’s responsibility for risks that this assessment has identified, including whether the Foundation may have caused, contributed to, or may be directly linked to a harm.

¹ <https://wikimediafoundation.org/about/>

² Rightsholders include anyone potentially impacted by a product or service, including project volunteers, readers and others impacted by each free knowledge project

³ <https://www.ohchr.org/en/what-are-human-rights/international-bill-human-rights>

⁴ https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

⁵ See Appendix I for an overview of the UNGPs

⁶ <https://globalnetworkinitiative.org/gni-principles/>

⁷ This report includes quotes from engaged stakeholders. While Article One made every effort to quote directly, the quotations were edited, at times, for ease of understanding. To promote transparency during the interviews, Article One committed to non-attribution of quotes.

It is important to note that the public version of this report is a joint effort between Article One and the Wikimedia Foundation, based on a full HRIA independently conducted by Article One and submitted to the Foundation in July 2020. As with all impact assessments it remains a snapshot in time, highlighting human rights risks and corresponding management practices from 2020. It does not include actions the Foundation has taken since the assessment was submitted or additional risks that may have materialized in the last two years. Article One and the Wikimedia Foundation jointly edited this public version of the report to protect the safety and security of Foundation staff and the larger volunteer community.

▣ Salient Human Rights Risks

The HRIA found that Wikimedia’s free knowledge projects help advance the realization of multiple human rights, including the right to free expression and to impart and access information. At the same time, the Foundation faces five categories of salient human rights risks related to its free knowledge projects.

These salient human rights risks included:



1 Harmful Content

Harmful content can impact a range of rights including security of person (UDHR 3), the right to access information (UDHR 19), to take part in government (UDHR 21), to be free from unlawful attacks on one’s honor and reputation (ICCPR 17), and to truth (Resolution 2005/66).

The assessment found four types of harmful content that impact human rights:

- A. **Attacks on individuals profiled**, including the vandalism of biographies of living persons, doxing, and spreading hate speech;
- B. **Misrepresentation of historical facts**, including hosting conspiracy theories, white washing, unreliable sourcing, sock puppeting and meat puppeting⁸;
- C. **Project capture**, involving the potential spread of government sponsored and/or nationalist-leaning disinformation across the free knowledge projects; and
- D. **Dangerous content**, including content that can contribute to self-harm or harm to others.

These risks are especially salient on Wikipedia given its efforts to provide encyclopedic knowledge and the degree to which the project is used globally.

⁸ We recognize that sock puppeting and meat puppeting represent tactics to bypass community consensus processes, rather than inherent forms of misrepresentation.

2 Harassment

Harassment can take many forms, including gendered attacks on volunteers who identify publicly as female, transgender, or non-binary; doxing of personal information as well as threats of violence. At its most fundamental level, online harassment can impact on the right to be treated with dignity, but it can also impact on the right to: non-discrimination (UDHR 2), right to security of persons (UDHR 3), privacy (UDHR 12), expression (UDHR 19), assembly (UDHR 20), participation in cultural life (UDHR 27), and to be free from unlawful attacks on one's honor and reputation (ICCPR 17).

The assessment found two primary types of harassment that impact human rights:

- A. **Harassment within the volunteer community**, predominantly of minority voices on the knowledge projects, including abusive speech, doxing, defamation, and blackmailing; and
- B. **Harassment of Foundation staff by volunteers**, both on and offline.

3 Government surveillance and censorship

Human rights are being challenged around the world, especially in relation to free expression, freedom of the press, internet blackouts, internet content controls, and crackdowns on human rights defenders. For the Foundation, these infringements may impact on the rights to: security of persons (UDHR 3), to be free from torture (UDHR 5), privacy (UDHR 12), expression (UDHR 19), assembly (UDHR 20), and to take part in government (UDHR 21).

The assessment found three primary risks related to government surveillance and censorship:

- A. **Online surveillance of Wikimedia volunteers and readers**, especially in countries with restricted internet freedoms or authoritarian governments and on topics considered taboo in those countries;
- B. **Requests for user data**, including formal and informal government requests to the Foundation. There is an increasing risk for community members who handle non-public data to receive requests directly from government officials; and
- C. **Government censorship**, ranging from blocking certain sections of articles to intermittently blocking access to Wikipedia as a whole.

4 Risks to child rights

Editors of all ages are welcome to contribute to Wikimedia projects. Some volunteers who have served in important roles in the Wikimedia communities of editors later disclosed that they had been minors at the time. However, risks to children remain, and may impact on the right to: dignity (UDHR 1), privacy (UDHR 12), free expression (UDHR 19), education (UDHR 26), protection from harmful content (CRC 17), protection from sexual exploitation and abuse (CRC 34), and to be free from unlawful attacks on one's honor and reputation (ICCPR 17).

The assessment found four primary risks related to child rights:

- A. **Privacy and reputation risks**, given that children are at times the focus of content on knowledge projects. This includes child activists and celebrities, who may be subject to smear campaigns;
- B. **Exposure to harmful content**, including content that “promotes substance abuse, racial hatred, risk-taking behavior or suicide, anorexia or violence;”⁹
- C. **Child sexual exploitation material** that has been found on Wikimedia platforms;¹⁰ and
- D. **Harmful contact**, such as grooming children to perform sexual acts on and offline and to purchase illegal products, which may occur on Wikimedia talk pages.

⁹ UNICEF: “Children’s Rights and the Internet From Guidelines to Practice” (2016)

¹⁰ Article One interview with Wikimedia staff member in June 2020; note that action is in progress, including an automated analysis of Wikimedia content and AFAIK

5 Limitations on knowledge equity

The Foundation is committed to making information more accessible and bringing forward knowledge left out by systems of privilege and power.¹¹ In pursuit of these ambitions, the Foundation has faced a series of challenges that can impact on the rights to: be free from discrimination (UDHR 2), expression and information (UDHR 19), and cultural participation (UDHR 27).

The assessment found four primary risks related to limitations on knowledge equity:

- A. **Gender Equity**, including having a disproportionately small number of women contributing to projects; a harassing environment for many female, LGBTQ+, and non-binary volunteers; and underrepresentation of women, LGBTQ+, and non-binary individuals in content;
- B. **Racial and Ethnic Diversity**, including the underrepresentation of historically underrepresented racial and ethnic groups as subjects of articles and contributors,¹² poor retention of minority editors, and harassment of ethnic minorities in volunteer communities;
- C. **Knowledge Equity of the Global South**, due to the limited accessibility of the platform in the global south and underrepresentation of the world's languages; and
- D. **Accessibility of knowledge projects for those with visual, audial or other disabilities and impairments, and communicating knowledge in ways beyond writing.**

▣ Recommendations

Article One developed a suite of recommendations to address each category of salient risks. **We recognize the need to engage and secure input from Wikimedia's vast volunteer base and as such recommend that the Foundation consult with volunteers and other experts to determine the best path forward.** Priority recommendations include:

Strategies for the Foundation

1. Develop a standalone Human Rights Policy that commits to respecting all internationally recognized human rights by referencing the International Bill of Human Rights.
2. Conduct ongoing human rights due diligence to continually assess risks to rightsholders. A Foundation-level HRIA should be conducted every three years or whenever significant changes could have an effect on human rights.
3. Develop rights-compatible channels to address human rights concerns, including private channels, and ensure alignment with the UNGPs' effectiveness criteria.

Harmful Content

1. Develop an audit protocol to assess projects that are at high risk of capture or government-sponsored disinformation.
2. Develop a Content Oversight Committee (COC) to review content with a focus on bias and have the ability to make binding editorial decisions in line with ICCPR 19.
3. Continue efforts outlined in the Knowledge Integrity¹³ white paper to develop: a) a machine-readable representation of knowledge that exists within Wikimedia projects along with its provenance; b) models to assess the quality of information provenance; and c) models to assess content neutrality and bias. Ensure that all AI/ML tools are designed to detect content and action that would be considered illegal under international human rights law, and that the response aligns with the three-part ICCPR test requiring that any restriction on the right to free expression be legal, proportional, and necessary.
4. Provide access to a geotargeted suicide prevention hotline at the top of the articles on Suicide Methods.

¹¹ Berkman Klein Center for Internet & Society. "Will Wikimedia Exist in 20 Years?" (2017)

¹² Fast Company. "Black History Matters, So Why is Wikipedia Missing So Much of It?" (2015)

¹³ Leila Zia, Isaac Johnson, Bahodir Mansurov, Jonathan Morgan, Miriam Redi, Diego Saez-Trumper, and Dario Taraborelli. 2019. Knowledge Integrity - Wikimedia Research 2030. doi.org/10.6084/m9.figshare.7704626 [CC BY 4.0]

Harassment

1. Develop and deploy training programs for admins and volunteers with advanced rights on detecting and responding to harassment claims.
2. Commission a “social norms marketing” research project to assess what type of messaging is likely to reduce and prevent harassing comments and actions.
3. Explore opportunities to rate the toxicity of users, helping to identify repeat offenders and patterns of harassment. Consider awards for projects with the lowest toxicity levels.
4. Consider developing admin metrics focused on enforcing civility and applying the forthcoming Universal Code of Conduct (UCoC).
5. Ensure that the (UCoC) and its accompanying governance mechanism is reviewed by human rights experts, including experts on free expression and incitement to violence.

Government surveillance and censorship

1. Continue efforts underway as part of the IP-masking project to further protect users from public identification.
2. Develop awareness-raising tools and programs for all volunteers to understand and mitigate risks of engagement. Tools should be made publicly available and should be translated into languages spoken by volunteers in higher risk regions.¹⁴

Risks to child rights

1. Conduct a child rights impact assessment of Wikimedia projects, including conducting interviews and focus groups with child contributors across the globe.
2. Create child safeguarding tools, including child-friendly guidance on privacy settings, data collection, reporting of grooming attempts, the forthcoming UCoC as well a “Child’s Guide to Editing Wikimedia Project” to help advance the right of children to be civically engaged.

Limitations on knowledge equity

1. Support retention by developing peer support and mentoring for under-represented contributors.
2. Engage stakeholders on how the “notability” requirement may be shifted to be more inclusive of oral histories, and to identify what definitions resonate with under-represented communities.
3. Adapt Wikimedia projects to be more accessible via mobile phones.

¹⁴ Higher risk regions can be determined based on historical knowledge from the Foundation combined with country rankings on human rights and internet freedom, including, for example, Freedom House’s [Freedom on the Net](#) report.

*Assessing the Human Rights Impacts
of Wikimedia Free Knowledge Projects*



Introduction





II. Introduction

The Wikimedia Foundation (the Foundation) hosts free knowledge projects and protects “the values and policies that allow free knowledge to thrive.” This vision advances the fundamental right to access and impart information globally.

To ensure that Wikimedia projects continue to advance respect for human rights, the Foundation has commissioned this human rights impact assessment (HRIA) with the goal to:

- 1 Surface salient human rights risks across its free knowledge projects;
- 2 Mitigate actual and potential risks related to its projects around the world, including avoiding harm to rightsholders¹⁵;
- 3 Support the Foundation in becoming a member of the Global Network Initiative (GNI); and
- 4 Better align with stakeholder expectations regarding ongoing human rights due diligence.

International human rights standards provide a powerful framework for the Foundation to understand and address the risks associated with its free knowledge projects. Human rights standards, including the International Bill of Human Rights¹⁶ and the UN Guiding Principles on Business and Human Rights (UNGPs)¹⁷, offer a global, broadly accepted set of values to inform the Foundation’s approach to surfacing and mitigating human rights risks.¹⁸ Indeed, human rights are inherent, inalienable, interdependent, and indivisible: they cannot be granted or taken away, the enjoyment of one right affects the enjoyment of others and as such they must all be respected.

Importantly, many of the key concerns raised by the Foundation, the community of editors, and other stakeholders over recent years — including harassment on Wikimedia platforms, concerns regarding potential disinformation campaigns, surveillance of the Wikimedia community by state actors and so on — touch on fundamental rights protected under the international human rights framework.

The public version of this report is a joint effort between Article One and the Wikimedia Foundation based on a full HRIA independently conducted by Article One and submitted to the Foundation in July 2020. As with all impact assessments, it remains a snapshot in time, highlighting human rights risks and corresponding management practices from 2020. It does not include actions the Foundation has taken since the assessment was submitted or additional risks that may have materialized in the last two years. Article One and the Wikimedia Foundation jointly edited this public version of the report to protect the safety and security of Foundation staff and the larger volunteer community.

¹⁵ Rightsholders include anyone potentially impacted by a product or service, including project volunteers, readers and others impacted by each free knowledge project

¹⁶ <https://www.ohchr.org/en/what-are-human-rights/international-bill-human-rights>

¹⁷ https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

¹⁸ See Appendix I for an overview of the UNGPs

▣ UN Guiding Principles on Business & Human Rights

In 2011, the UN Human Rights Council unanimously endorsed the UNGPs. The UNGPs recognize the state's ultimate duty to protect, and business' responsibility to respect human rights. These principles include guidance for both states and companies related to three core pillars:



Pillar 1, the State Duty to Protect, recognizes the State's duty to protect its citizens against corporate human rights abuses. Protection is best accomplished through robust laws that align with international human rights standards and a strong rule of law that ensures their enforcement.

Pillar 2 calls on companies and other organizations operating in similar capacities to publish a policy commitment in support of human rights and to "know and show" their respect for human rights by acting with due diligence. This includes:

1. Assessing actual and potential impacts, including through human rights impact assessments;
2. Integrating the findings of the assessment across the entire business and taking appropriate action to address adverse impacts; and
3. Tracking and communicating performance.

As part of the due diligence expectation, the UNGPs recognize that companies may need to prioritize which actual and potential impacts to address. However, these impacts should not be prioritized based on the company's relationship to an impact, but rather on its saliency, specifically on the degree of risk to rightsholders. Indeed, a key differentiator of the UNGPs is the focus on risks to rightsholders, rather than on risks to the business or organization.

Pillar 3 outlines the obligations of both states and companies to provide access to effective remedies in cases of human rights infringements. If the company is found to have caused or contributed to an impact, the company may be obligated to provide or facilitate access to a remedy. If the company is directly linked to an impact through a business relationship, there is no obligation to provide or facilitate access though the company may use its leverage to help ensure a remedy is provided.

*Assessing the Human Rights Impacts
of Wikimedia Free Knowledge Projects*



Scope & Methodology of the Assessment

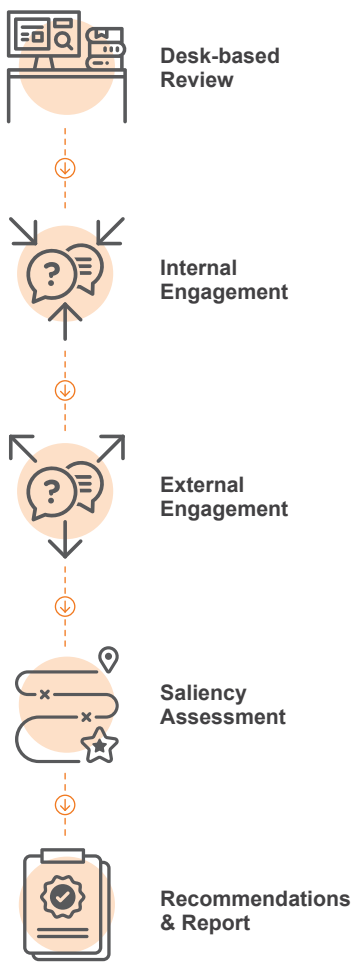


III. Scope & Methodology of the Assessment

This Human Rights Impact Assessment (HRIA) follows guidance from the UNGPs and the Global Network Initiative (GNI) Principles and is based on Article One’s award-winning methodology for and experience in conducting human rights impact assessments.

The assessment, conducted in 2020, focused on risks associated with Wikimedia free knowledge projects. Where risks had a direct impact on Wikimedia Foundation staff, we have outlined them. However, the assessment does not focus on potential impacts related to other areas of the Foundation, including human resources management and supply chain related risks.

The methodology included five key phases:



Article One conducted a desk review of public and private information on Wikimedia projects, including news reports, research assessment and confidential information shared with Article One under an NDA. We supplemented the desk review with interviews of 17 Foundation staff and six leading external experts.¹⁹ The findings were then validated with six volunteers from across the globe. The assessment was completed in July 2020. Additional risks and mitigations that occurred after that date are not included in this report.

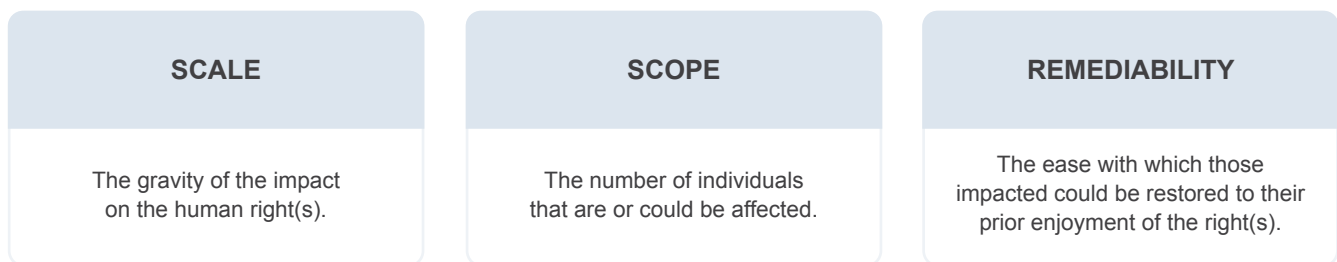
Human rights risks were mapped to the Universal Declaration of Human Rights (UDHR) and the Convention on the Rights of the Child (CRC). These rights include the following, as outlined in the UDHR:

- | | |
|---|---|
| Article 1 Right to Equality | Article 16 Right to Marriage and Family |
| Article 2 Freedom from Discrimination | Article 17 Right to Own Property |
| Article 3 Right to Life, Liberty, Personal Security | Article 18 Freedom of Belief and Religion |
| Article 4 Freedom from Slavery | Article 19 Freedom of Opinion and Information |
| Article 5 Freedom from Torture and Degrading Treatment | Article 20 Right of Peaceful Assembly and Association |
| Article 6 Right to Recognition as a Person before the Law | Article 21 Right to Participate in Government and in Free Elections |
| Article 7 Right to Equality before the Law | Article 22 Right to Social Security |
| Article 8 Right to Remedy by Competent Tribunal | Article 23 Right to Desirable Work and to Join Trade Unions |
| Article 9 Freedom from Arbitrary Arrest and Exile | Article 24 Right to Rest and Leisure |
| Article 10 Right to Fair Public Hearing | Article 25 Right to Adequate Living Standard |
| Article 11 Right to be Considered Innocent until Proven Guilty | Article 26 Right to Education |
| Article 12 Right to Privacy, Family, Home and Correspondence | Article 27 Right to Participate in the Cultural Life of Community |
| Article 13 Right to Free Movement in and out of the Country | Article 28 Right to a Social Order that Articulates this Document |
| Article 14 Right to Asylum in other Countries from Persecution | Article 29 Community Duties Essential to Free and Full Development |
| Article 15 Right to a Nationality and the Freedom to Change It | Article 30 Freedom from State or Personal Interference in these Rights |

¹⁹ This report includes quotes from engaged stakeholders. While Article One made every effort to quote directly, the quotations were edited, at times, for ease of understanding. To promote transparency during the interviews, Article One committed to non-attribution of quotes.



To evaluate the relative priority of salient risks, Article One considered the likelihood and the severity of each risk based on:



In addition, we applied guidance from the UN Office of the High Commissioner for Human Rights (OHCHR) to determine Wikimedia’s responsibility for surfaced risks, including whether the Foundation may have caused or contributed to, or may be directly linked²⁰ to a harm. An organization may **contribute** to an adverse impact if it:²¹

- **Incentivized harm**, including whether the organization’s actions or omissions (failure to act) make it more likely that someone else will cause the harm.
- **Facilitated the harm**, for example where the organization adds to conditions that make it possible for someone else to cause harm.
- **Failed to adequately conduct human rights due diligence** in line with the UNGPs.
- **Knew or should have known** about the adverse impact.

Article One then developed a series of recommendations to support the Foundation in maximizing its positive human rights impacts and mitigating adverse impacts related to its free knowledge projects.

This public version of the report was jointly developed by the Wikimedia Foundation and Article One. It has been edited to protect the safety and security of Foundation staff and the larger volunteer community. The edits have included removal of:

- * Specific examples where the human rights of individuals involved may be a risk if the examples were made public;
- * Information that could empower malicious actors;
- * The causal relationship to harm given that additional mitigations have occurred since the report was submitted in July 2020; and
- * Private, confidential research conducted by the Foundation.

Despite these edits, we believe this report provides a holistic overview of the relevant human rights risks related to the Foundation and Wikimedia free knowledge projects. We hope this report contributes to broader awareness of human rights and management of risks at the Foundation.

²⁰ An organization may be directly linked to a human rights impact that is caused by an entity with which it has a business relationship through its own operations, products or services.

²¹ OHCHR: [“OHCHR response to request from BankTrack”](#) and B Tech [“Taking Action to Address Human Rights Risks Related to End-Use”](#)

*Assessing the Human Rights Impacts
of Wikimedia Free Knowledge Projects*



Salient Human Rights Risks



IV. Salient Human Rights Risks

The HRIA found that Wikimedia’s free knowledge projects help advance the realization of multiple human rights, including the right to free expression and to impart and access information. At the same time, the Foundation faces salient human rights risks related to free knowledge projects.

Based on the findings of the assessment, Article One has developed five key categories of risk:

-  **Harmful content**
-  **Harassment**
-  **Government surveillance and censorship**
-  **Risks to child rights**
-  **Limitations on knowledge equity**

For each category of risk, this HRIA report includes:

- An overview of the risk, outlining how it intersects with the protection of human rights
- Analysis of how the risk has manifested on Wikimedia’s free knowledge projects
- Existing mitigation measures
- Recommendations to mitigate actual and potential risk

As it relates to our recommendations, Article One recognizes the need to engage and secure input from Wikimedia’s vast volunteer base. We recognize the many benefits of a grassroots approach to governance while outlining the instances where there is need for greater support and oversight from the Foundation. Where we have highlighted these needs, we recommended that the Foundation consult with volunteers and other experts to determine the best path forward.



Risk
Harmful Content

- Overview
- Analysis
- Mitigation Measures
- Recommendations

Online platforms, including social media platforms and Wikimedia’s free knowledge projects, have been used by bad actors to disseminate harmful content. This phenomenon includes attacks on individual people such as political leaders,²² misrepresentation of historical facts,²³ and the dissemination of disinformation. Importantly, Wikimedia’s nonprofit model does not incentivize “viral” content and prioritizes accurate, unbiased information. As such, Wikipedia and Wikimedia projects have established volunteer-led policies and systems to guard for neutrality and reliability of content on the site. These systems largely work to prevent and address bias, misinformation, and disinformation. While in most cases edits are identified and rectified quickly by volunteers, the increasingly sophisticated tools available to bad actors warrants constant vigilance.²⁴

For harmful content to be considered a potential infringement on human rights, it must limit the ability for people to access information, to be free from defamation, or to ensure their security. Satire for example, would not be considered harmful content while disinformation including all “forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit”²⁶ would. Importantly, we recognize and acknowledge the guidance from

OHCHR in its Resolution 2005/66 that the “right to the truth may be characterized differently in some legal systems as the right to know or the right to be informed or freedom of information.” While traditionally, the right to truth has been associated with the need to end impunity when it comes to gross human rights violations, the language of Resolution 2005/66 allows for an understanding of the right to truth to be inclusive of the right to be informed, a right that some types of harmful content such as disinformation may infringe upon.

The right to truth and information is especially salient for Wikipedia given the platform’s position as an online encyclopedia – placing a clear expectation that users will be accessing reliable and truthful information.²⁷ Indeed, Wikipedia has become essential to the work of journalists, students and inquisitive minds, helping to form our understanding of the world.²⁸ According to the MIT Technology Review, Google ratings of Wikipedia verify its dominance as an online encyclopedia which in turn “means that the content of [Wikipedia] articles really matters.”²⁹ This is reinforced by the increasing use of digital assistants, such as Amazon’s Alexa and Google’s Google Home, which rely on Wikipedia to answer user queries.³⁰

Human Rights	Justification for Inclusion
Dignity (UDHR 1)	Risk that dangerous content (e.g., suicide methods) contributes to self-harm or harm to others.
To security of person (UDHR 3)	Harmful content may contribute to offline harm, such as physical attacks on individuals or groups cited in Wikipedia pages, for example.
To access information (UDHR 19)	According to OHCHR, disinformation is “often designed and implemented so as to mislead a population, as well as to interfere with the public’s right to know and the right of individuals to seek and receive, as well as to impart information and ideas of all kinds.” ³¹ This is an especially salient issue for Wikimedia given its focus on knowledge.

²² Wikipedia: “Wikipedia is Not a Reliable Source” (2020)

²³ Haaretz, “The Fake Nazi Death Camp: Wikipedia’s Longest Hoax Exposed” (2019)

²⁴ Wikimedia Foundation: “The nationalist takeover of Croatian Wikipedia”

²⁵ European Commission: “A multi-dimensional approach to disinformation” (2018). It is important to note that disinformation does not include misleading advertising, reporting errors, satire and parody, or clearly identified partisan news and commentary.

²⁶ Human Rights Resolution 2005/66: “Right to the truth”

²⁷ This expectation exists despite statements by Wikipedia that “it is not a reliable source.

Wikipedia can be edited by anyone at any time. This means that any information it contains at any particular time could be vandalism, a work in progress, or just plain wrong.” Wikipedia: “Wikipedia it is not a reliable source”

²⁸ MIT Technology Review: “Wikipedia and the Meaning of Truth” (2008).

²⁹ MIT Technology Review: “Wikipedia and the Meaning of Truth” (2008).

³⁰ Voicebot ai: “Voice Assistants Alexa, Bixby, Google Assistant and Siri Rely on Wikipedia and Yelp to Answer Many Common Questions about Brands”

³¹ OHCHR: Joint Declaration On Freedom Of Expression And “Fake News”, Disinformation And Propaganda

Human Rights	Justification for Inclusion
<i>To take part in government (UDHR 21)</i>	For an election to be free and fair, voters need to have accurate information about the parties, candidates, and issues when they vote. According to OHCHR, “without a well-informed electorate, it is impossible to guarantee that elections genuinely reflect the will of the people.” ³² If Wikimedia projects include false and misleading information about candidates and parties, it may influence a voter’s opinion and in turn vote.
<i>Freedom from unlawful attacks on one’s honor and reputation (ICCPR 17)</i>	In so much as harmful content relates to a specific individual – for example a public figure – it is designed to harm that person’s reputation, often in pursuit of other goals.
<i>The right to truth (Resolution 2005/66)</i>	The right to truth may be understood to be inclusive of the right to be informed, a right that some forms of harmful content, including disinformation, may infringe upon.

Overview

Wikimedia’s free knowledge projects allow anyone sharing its vision to participate in its efforts to share in the sum of all knowledge.³³ The open contribution model has powerful benefits, but also opens up the risk of “*special interests to introduce bias and misinformation.*”³⁴ Article One’s assessment found four forms of harmful content or actions that could be found to infringe on human rights. These forms go from the micro – attacks on single places – to the macro – capture of full projects.

- A. Attacks on individuals profiled
- B. Misrepresentation of historical facts
- C. Project capture
- D. Dangerous content

These risks are especially salient on Wikipedia given its efforts to provide encyclopedic knowledge and the degree to which the project is used globally.

Analysis

A. Attacks on Individuals Profiled

Wikipedia can be edited by anyone at any time, thus its content is forever a work in progress. Given this, some edits and new content may not meet the robust community standards that volunteers have created and regularly enforce around neutral point of view, reliability of sources, and verifiability on the site. In some instances, edits are intentionally malicious and do not meet these standards.³⁵ On Wikipedia, vandalism is the act of editing a project in a malicious manner that is intentionally disruptive. Vandalism includes the addition, removal, or modification of the text or other material that is either humorous,

nonsensical, a hoax, or that is of an offensive, humiliating, or otherwise degrading nature.³⁶

Biographies of living persons, subjects that happen to be in the news, and politically or culturally contentious topics are especially vulnerable to these issues.³⁷ For example, in October 2016, both Hillary Rodham Clinton and Bill Clinton’s Wikipedia pages were vandalized, and pornographic images were added to their articles by an internet trolling group.³⁸ Other examples include anonymous editors removing or hiding the page names of individuals on the “List of Transgender People” who are cited as Catholic and adding homosexual references and

³² OHCHR: “Human Rights and Elections” (1994); In the [Joint Declaration On Freedom Of Expression And “Fake News”, Disinformation And Propaganda](#) OHCHR also outline that “the human right to impart information and ideas is not limited to ‘correct’ statements, that the right also protects information and ideas that may shock, offend and disturb, and that prohibitions on disinformation may violate international human rights standards, while, at the same time, this does not justify the dissemination of knowingly or recklessly false statements by official or State actors.”

³³ Wikimedia Foundation: “[Knowledge Integrity](#)” (2019)

³⁴ Wikimedia Foundation: “[Knowledge Integrity](#)” (2019)

³⁵ Wikipedia: “[Wikipedia is Not a Reliable Source](#)” (2020)

³⁶ Wikipedia: “[Vandalism on Wikipedia](#)” (2020)

³⁷ Wikipedia: “[Wikipedia is Not a Reliable Source](#)” (2020)

³⁸ Slate: “[Hillary Clinton Wikipedia Page Vandalized With Pornographic Images, Pro-Trump Message](#)” (2016)

personal attacks to Philippine media personality Mo Twister's biography.³⁹

Another category of attacks includes doxing, where personal information of living people is released without their consent. In September 2018, for example, the personal information of three prominent United States senators was added to their respective Wikipedia articles during the hearing of Supreme Court Nominee Judge Brett Kavanaugh.⁴⁰ The information included their home addresses and phone numbers. While the edits were removed from Wikipedia by the community shortly afterwards, they were screenshotted and disseminated via Twitter. This example highlights the potential for Wikipedia to be weaponized through cross-platform distribution.

Given that acts of vandalism span from light trolling to harmful disinformation, the level of impacts on human rights can vary. In its most extreme forms, including doxing and spreading hate speech, these acts can impact on the right to privacy (UDHR 12) and security of persons (UDHR 3), including the right to freedom from harassment and discrimination (UDHR 2).

B. Misrepresentation of Historical Facts

Given Wikimedia's grassroots governance approach, much of the content on each of the projects is up for debate and discussion. While the projects include guidance to editors on the type of content that is appropriate, including the need for it to be robustly sourced, there are ongoing concerns that some editors have been successful in pushing false narratives, especially against minority and marginalized groups. This has been an increasing risk in the last four years, driven in part by a rising nationalist sentiment globally and the increased use of social media platforms to "coordinate actions in a distributed fashion across multiple platforms."⁴¹

For example, Haaretz reported on the "Warsaw Concentration Camp" Wikipedia article which was first published in 2004, calling it "Wikipedia's longest hoax" as the article contained misinformation and conspiracy theories about the camp until 2019.⁴² The article formerly suggested the "estimates of the camp's victims are well above 212,000, mainly Poles and several thousands of non-Polish," exaggerating the total death toll, and underrepresenting the portion of Jewish prisoners relative to non-Jewish Poles.⁴³ According to Haaretz, the narrative is harmful because it minimizes "Polish cooperation and collaboration with the Nazis in the persecution of Jews."⁴⁴

Haaretz reported, while the theory "has failed to make headway in academia or the world media, on Wikipedia it has thrived."⁴⁵

Another type of misrepresentation on the platform is whitewashing, in which an article is written or edited to "to gloss over or cover up vices, crimes or scandals or to exonerate by means of a perfunctory investigation or through biased presentation of data."⁴⁶ One example comes from Wikipedia's article on Afrikaner Weerstandsbeweging (AWB), a South African Neo-Nazi group. According to research from Southern Poverty Law Center (SPLC), the edit history of the article reveals a long pattern of edit-warring on efforts to characterize the group as "neo-Nazi white supremacists" and the inclusion of accurate descriptions of the group's use of Nazi imagery. As of 2018, the page's editors had effectively eliminated any reference to the group's violent past.⁴⁷

SPLC's research found several weaknesses and tactics used by volunteers to push historical narratives that counter factual evidence.⁴⁸ These include:

1. **Sources:** The "Reliable sources/Perennial sources" list,⁴⁹ while an important contribution, leaves room for debate and the potential for ideologues to rely on inappropriate sources to support their edits.⁵⁰ This allows experienced contributors to manipulate their references to support biased views.⁵¹
2. **Sock Puppeting:** Sock puppeting is the abuse of multiple accounts to skirt bans⁵² and other administrative actions and promote the idea that a "view has wider support" than it does. According to SPLC, a white nationalist who co-founded Rightpedia, a far-right free encyclopedia, created more than 140 Wikipedia accounts in the past 10 years to push his point of view on Wikipedia.
3. **Canvassing/Meat Puppeting:** This includes recruiting other like-minded individuals to edit content. The SPLC highlights that the proliferation of far-right online spaces, such as white nationalist forums and alt-right boards, has "created a readymade pool of users that can be recruited to edit on Wikipedia en masse."

While each of these actions in a specific instance may not be considered a violation of human rights, the cumulative impact of the types of harmful content described above may result in the spread of uninformed narratives that impact on a reader's view of the world around them, thereby impacting on the right to information (UDHR 19) and potentially other rights including the right to participate in government (UDHR 21).

³⁹ Wikipedia, "Wikipedia: Most Vandalized pages" (2020)

⁴⁰ Washington Post: "Fight over Kavanaugh nomination finds its oddest front yet: Wikipedia pages" (2018)

⁴¹ Wikimedia Foundation: "Knowledge Integrity" (2019)

⁴²⁻⁴⁵ Haaretz, "The Fake Nazi Death Camp: Wikipedia's Longest Hoax, Exposed" (2019)

⁴⁶ Wikipedia, "Whitewashing (censorship)" (2020)

^{47, 48} The Southern Poverty Law Center: "Wikipedia wars: inside the fight against far-right editors, vandals and sock puppets" (2018)

⁴⁹ Wikipedia: "Reliable sources/Perennial sources"

⁵⁰ Wikipedia does maintain a noticeboard for fringe sources and theories.

⁵¹ Harvard Business Review: "How Wikipedia Keeps Political Discourse from Turning Ugly" (2016)

⁵² It is important to note that using multiple accounts to skirt bans is against Wikipedia's rules.

C. Project Capture

Staff from across the Foundation raised serious concerns regarding the potential spread of government sponsored and/or nationalist-leaning disinformation across the free knowledge projects. Staff largely recognized these risks are more likely in countries with limitations on free expression and on projects with a limited pool of volunteers and a limited number of native language speakers.

The most extensive case of project capture to date occurred on Croatian Wikipedia which, according to research supported by the Foundation, was “*dominated by ideologically driven users*” who “*have held de-facto control over the project for more than a decade.*”⁵³ According to the research these users have “*intentionally distorted the content presented in articles, abused power, and systematically obstructed otherwise accepted global Wikipedia community practices.*”⁵⁴

The Foundation’s research on Croatian Wikipedia found that politically motivated bias has been inserted into the project in three key ways:⁵⁵

- ◆ **Selection bias:** selectively including and excluding content regardless of its notability or topical relevance;
- ◆ **Framing bias:** contextualizing factual claims in articles in non-neutral ways to mislead readers; and
- ◆ **Source bias:** supporting factual claims with unreliable sources to promote a specific agenda.

The ability of the broader Croatian Wikimedia community to place checks on nationalist editors was limited. According to the research, ideologically driven users had been using “*on-wiki positions of power to attract like-minded contributors,*

silence and ban dissenters, manipulate community elections and subvert Wikimedia’s and the broader movement’s native conflict solutions mechanisms.”⁵⁶ Importantly, while the capture of Croatian Wikipedia is the most well-known, the author of the report warns that “*there could be similar attempts of project capture in other languages.*”⁵⁷

Organized and systematic disinformation campaigns such as what has taken place on Croatian Wikipedia are the most likely to result in significant adverse human rights impacts. These include not only impacts on the right to information and free expression (UDHR 19), but also may contribute to a cumulative impact on the right to participate in government (UDHR 21) by reducing the ability for the population to be adequately informed and the right to truth (Resolution 2005/66), especially if it erases prior human rights violations against specific groups or communities.

D. Dangerous Content

Content Warning: The following paragraph contains discussion of suicide, which some readers may find distressing.

A final area of harmful concern relates to content that could be used to harm oneself or others. One clear example is Wikipedia’s “Suicide Methods” page which provides guidance on a variety of approaches to suicide – from slitting one’s wrists to suffocation.⁵⁸ Suicide prevention advocates in the UK raised concerns that this content is not paired with appropriate resources to connect readers with mental health professionals, including for example national suicide hotlines.⁵⁹

*have developed over the years robust socio-technical strategies for defining, identifying, and addressing threats to the integrity (e.g., neutrality, verifiability, overall ‘quality’) of knowledge they create and curate. These communities have policies, processes, as well as tools for dealing with issues of vandalism, non-neutral language, conflicts of interest, promotional editing, sockpuppets, and spam.*⁶¹

■ Risk Mitigation Measures

Wikimedia communities are empowered to self-govern, reducing both the need and the ability for the Foundation to step in. Indeed, Wikimedia communities often push back against what some perceive as Foundation overreach.⁶⁰ According to Foundation research, communities:

⁵³⁻⁵⁷ The Case of Croatian Wikipedia: Encyclopaedia of Knowledge or Encyclopaedia for the Nation? Available at: https://upload.wikimedia.org/wikipedia/commons/1/14/Croatian_WP_Disinformation_Assessment_-_Final_Report_EN.pdf

⁵⁸ Wikipedia: “Suicide methods”

⁵⁹ Interview with Foundation staff in July 2020

⁶⁰ One email communication with Wikimedia Foundation staff

⁶¹ Wikimedia Foundation: “Knowledge Integrity” (2019)

However, as the Foundation’s own research points out, the governance approaches have largely been developed to address individual bad actors working alone, rather than targeted and coordinated campaigns to spread disinformation.⁶² That said, in recent years the Foundation has taken several steps to try to proactively mitigate risks.

To address harmful content, the Foundation created the Knowledge Integrity program in 2018. The program is intended to “*help our communities represent, curate and understand information provenance in Wikimedia projects more efficiently.*”⁶³

These efforts include:⁶⁴

- ◆ Research on why editors source certain information and how readers access sources;
- ◆ The development of an algorithm to identify statements in need of sources and gaps in information provenance;
- ◆ The design of data structures to represent, annotate and analyze source metadata; and
- ◆ The development of tools to monitor in real time changes made to references across the Wikimedia ecosystem.

Wikimedia architecture also provides additional tools to preserve information quality, including:

- ◆ Watchlists for registered users that enable monitoring select pages for vandalism;
- ◆ Locking articles so only established users, or in some cases, administrators can edit them; and
- ◆ Blocking and banning those who have repeatedly committed acts of vandalism.

In addition, Wikimedia volunteers have developed and deployed AI tools to counter vandalism, including Clue Bot and VoxelBot as well as systems such as ORES which provide tools to scale up the ability for curators to monitor content quality in real time.⁶⁵

Under the Foundation’s Terms of Use, any contributor who is paid to edit on behalf of an individual, corporation, or government should be required to disclose their affiliation. However, these requirements are enforced more strictly in some projects than others, and are less applicable to other Wikimedia projects, such as Wikimedia Commons where volunteers upload freely-licensed or public domain media.

In 2020, many staff raised concerns about the Foundation’s reactive stance towards the risk of widespread disinformation and project capture. One reason for this was a lack of effective grievance mechanisms for volunteers or targeted individuals to elevate concerns to the Foundation. The bias toward community moderation of content can be effective where there is a diversity of voices and distributed power, but in cases where a diversity of voices is pushed out or where targeted individuals lack knowledge and insights into the moderation process, there is a risk that concerns go unreported.

Despite the lack of effective grievance mechanisms, as of 2020 initial efforts were already underway to better anticipate risks. For example, the Foundation launched a project in 2020 to assess alignment of Wikipedia texts across languages to determine whether they cover the same information, however, resources and volunteer support remain limited.⁶⁶ This work was complemented by experimental work to scrape talk pages where there may be signals of potential disinformation to trigger human review. In addition, as of 2020 the Foundation was in the process of hiring experts to research and combat disinformation on its projects.

■ Recommendations

Issues	Human Rights	Recommendation
Profile Attacks	<ul style="list-style-type: none"> → UDHR 1 → ICCPR 17 	<ul style="list-style-type: none"> ◆ Explore solutions to limit the ability to upload pornographic content and personally identifiable information on the profile pages of living people

⁶²⁻⁶⁴ Wikimedia Foundation: “Knowledge Integrity” (2019)

⁶⁵ Wikimedia Foundation: “Knowledge Integrity” (2019), Wikimedia: “Bots/Status”, MediaWiki: “ORES”

⁶⁶ Article One interview with Foundation staff member in June 2020.

Issues	Human Rights	Recommendation
<i>Historical Misrepresentation</i>	<ul style="list-style-type: none"> → UDHR 3 → UDHR 19 → UDHR 21 → UDHR 25 → Resolution 2005/17 	<ul style="list-style-type: none"> ◆ Continue efforts outlined in the Knowledge Integrity white paper to develop: a) a machine-readable representation of knowledge that exists within Wikimedia projects along with its provenance; b) models to assess the quality of information provenance; and c) models to assess content neutrality and bias. Ensure that all AI/ML tools are designed to detect content and action that would be considered illegal under international human rights law and that the response aligns with the three-part ICCPR test requiring that any restriction on the right to free expression be legal, proportional and necessary ◆ Expand efforts to build out a threat intelligence program through partnerships with NGOs and develop tools and strategies to equip admins with necessary knowledge and tools to effectively moderate content
<i>Project Capture</i>	<ul style="list-style-type: none"> → UDHR 3 → UDHR 19 → UDHR 21 → UDHR 25 → ICCPR 17 → Resolution 2005/17 	<ul style="list-style-type: none"> ◆ Develop an audit protocol to assess projects that are at high risk of capture or government-sponsored disinformation ◆ Develop a network of trusted advisors to surface key risks across Wikimedia's free knowledge projects globally ◆ Develop a Content Oversight Committee (COC) in line with the recommendations from Wikimedia Research on the capture of Croatian Wikipedia. The COC should review content with a focus on bias and have the ability to make binding editorial decisions in line with ICCPR 19
<i>Dangerous Content</i>	<ul style="list-style-type: none"> → UDHR 1 	<ul style="list-style-type: none"> ◆ Provide access to a geotargeted suicide prevention hotline at the top of the Suicide Methods ◆ Develop guidance for editors on how to write responsibly on suicide within the Wikimedia context ◆ Encourage and support edit-a-thons to add new content and improve existing content across projects ◆ Survey other pages that could be considered to provide guidance on how users can be harmful to themselves and others and develop similar mitigation tactics to the ones proposed for suicide
<i>All</i>	<ul style="list-style-type: none"> → UDHR 19 	<ul style="list-style-type: none"> ◆ Include a standard notice to readers outlining the limitations on the veracity of content on the knowledge platforms and linking to information about the Foundation's efforts to mitigate against these risks ◆ Partner with English-language Wikimedia volunteers to develop best practice guidance for compelling individuals paid to contribute to disclose their affiliation as outlined in the Terms of Service



Risk
Harassment

- Overview
- Analysis
- Mitigation Measures
- Recommendations

Harassment on Wikimedia platforms can take many forms. It can include annoying or rude comments made by volunteers; gendered attacks on volunteers who identify publicly as female, transgender, or non-binary, or edits on specific verticals (e.g., biographies of women); doxing⁶⁷ of personal information, as well as threats of violence. At its most fundamental level, online harassment can impact on the right to be treated with dignity – though the degree of harm depends both on the type and scale of harassment.

At the same time, many forms of comments, including those that shock and offend, are protected under Article 19 of the UDHR and ICCPR. However, as outlined in the ICCPR, the exercise of the right to free expression:

carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:

- a. *For respect of the rights or reputations of others;*
- b. *For the protection of national security or of public order (ordre public), or of public health or morals.*⁶⁸

Known as the three-part test, Article 19 of the ICCPR requires that any restriction on the right to free expression be legal, proportional, and necessary. According to the Australian Human Rights Commission which has focused significant attention on the interplay between technology and human rights, an appropriate reading of Article 19 may “*present possible justification for limitations on freedom of expression through the internet*” when speech infringes on the right: to non-discrimination; to be free from cruel, inhuman, or degrading treatment; the right of children to special protection; and privacy.⁶⁹

Amnesty International’s research on the experience of women on Twitter, for example, suggests that “*many forms of violence and abuse against women, such as direct threats of physical or sexual violence, are widely considered to be illegal in many domestic systems, and this is generally consistent with the right to freedom of expression.*”⁷⁰

With this understanding, Article One found that harassing actions by Wikimedia volunteers may have infringed on the following rights:

Human Rights	Justification for Inclusion
<i>Dignity (UDHR 1)</i>	Harassment, whether online or offline, infringes on the right to be treated with dignity –though the degree of harm depends both on the type and scale of harassment.
<i>Non-discrimination (UDHR 2)</i>	Online harassment impacts the right to be free from discrimination especially as harassment is often targeting vulnerable groups, including women, racial and ethnic minorities, political dissidents, and Human Rights Defenders (HRDs).
<i>To right to security of persons (UDHR 3)</i>	Online harassment that results in offline harm or action against targeted individuals can impact on the right to security of person.
<i>Privacy (UDHR 12)</i>	Certain forms of harassment, including doxing, are designed to share private information about people without their consent.

⁶⁷ Doxing is defined as: “search for and publish private or identifying information about (a particular individual) on the Internet, typically with malicious intent.”

⁶⁸ UN: [International Covenant on Civil and Political Rights](#)

⁶⁹ Australian Human Rights Commission: “[Permissible limitations of the ICCPR right to freedom of expression](#)”

⁷⁰ Amnesty International: “[Toxic Twitter: A Toxic Place for Women](#)”

Human Rights	Justification for Inclusion
<i>Expression & Assembly (UDHR 19 and 20)</i> <i>Participation in Cultural Life (UDHR 27)</i>	Online harassment can have a chilling effect on speech and assembly. If individuals feel unsafe engaging on Wikimedia’s projects the right to free expression, assembly and to participate in cultural life may be infringed.
<i>Free from unlawful attacks on one’s honor and reputation (ICCPR 17)</i>	Online harassment is often designed to attack the reputation of the targeted individual. This can occur both through direct statements as well as edits to free knowledge projects to defame individuals.

■ Overview

In 2017 Pew Research Center found that 41% of Americans have been personally subjected to harassing behavior online, and that 66% have witnessed these behaviors directed at others.⁷¹ While specific to the US market, anecdotal evidence and the results of country-level human rights impact assessments commissioned by Facebook suggest that these risks are global in nature and are likely to impact all types of online platforms, including Wikimedia’s free knowledge projects.⁷²

English-language Wikipedia defines harassment as a “*pattern of repeated offensive behavior that appears to a reasonable observer to intentionally target a specific person or persons. Usually (but not always), the purpose is to make the target feel threatened or intimidated, and the outcome may be to make editing Wikipedia unpleasant for the target, to undermine, frighten, or discourage them from editing.*”⁷³ The impact of harassment, however, may go well beyond the individual harms suffered to potential bias in content, limitations on the type of content contributed by volunteers and impacts related to the diversity of voices contributing to the projects. As outlined by the Berkman Klein Center:

*Harmful speech is antithetical to maintaining a high-quality encyclopedia. Reducing the level of abuse among Wikipedians is also of vital operational importance. Maintaining an active community of editors while attracting new participants is essential to the survival of Wikipedia. If only those individuals with the thickest of skins continue to participate, the future of the platform is less promising.*⁷⁴

In addition to the harassment of volunteers, Wikimedia Foundation staff have been the recipients of online and offline harassment by the volunteers who may disagree with Foundation programming or policy.

Below, we outline potential human rights risks associated with harassment within the Wikimedia community and against Foundation staff.

■ Analysis

A. Harassment within the Volunteer Community

Harassment of minority voices on Wikimedia projects was the most cited concern by staff during the assessment process. Research on English Wikipedia from the Berkman Klein Center found that “*there is a growing appreciation that minority*

and vulnerable communities tend to bear the brunt of online harassment and attacks and that this constitutes a major obstacle to their fully participating in economic, social, and cultural life online.”⁷⁵

In 2015, the Wikimedia Foundation conducted a survey to understand the impact of harassment on volunteers who have been the victims of the behavior. While not designed

⁷¹ Pew Research Center: “Online Harassment 2017” (2017)

⁷² Article One: “Our Assessment of Facebook’s Human Rights Impacts in Sri Lanka and Indonesia” (2020)

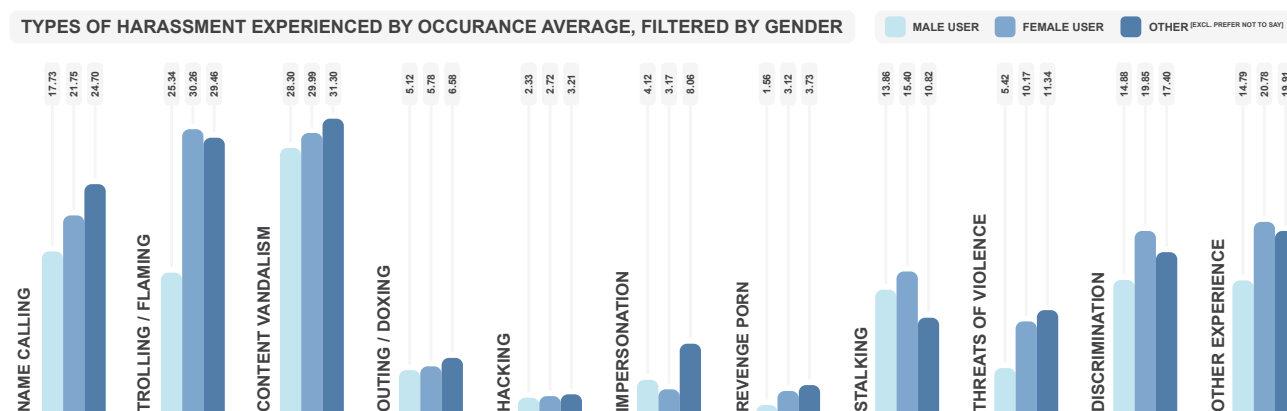
⁷³ Wikimedia: “Harassment”

⁷⁴ Berkman Klein Center for Internet & Society: “Content and Conduct: How English Wikipedia Moderates Harmful Speech” (2019)

to understand incidence rates, the survey found that 38% of the respondents could confidently recognize that they had been harassed, while 15% were unsure. In addition, 51% of respondents reported witnessing others being harassed.⁷⁶ Importantly, only 11% of survey respondents identified as female.

Research on English-language Wikipedia by the Foundation and Jigsaw found that 30% of attacks come from registered users with over 100 contributions and that an “outsized percentage of attacks” come from a handful of “highly toxic” contributors. Indeed, 9% of attacks in 2015 came from 34 users.⁷⁷

The experience of harassment on Wikimedia projects differed based on gender and cultural background. For example, women and “other genders” who responded to the survey reported higher rates than men when it came to 10 out of 11 forms of harassment.⁷⁸ This includes higher rates of name calling, trolling, doxing, threats of violence and discrimination. In only one instance, stalking, did men report higher rates than other genders, but not more than women. These findings align with the 2019 Community Insights study which found that almost half of female volunteers reported feeling “unsafe or uncomfortable” in Wikimedia spaces.⁷⁹



The 2015 Harassment survey also found that of those who responded:⁸⁰

- ◆ Male contributors are more likely to be targeted for single-time harassment, while 69% of other genders and 57% of female respondents report being targeted multiple times.⁸¹
- ◆ Contributors with similar cultural backgrounds are more likely to be targeted in single-time incidents, while 70% of culturally different editors report being harassed during multiple incidents.
- ◆ Other genders reported experiencing higher rates of long-term harassment (33%), including harassment that lasts more than a year.

Foundation staff reported that volunteers are more likely to receive harassment if they reveal themselves to be from a minority background—either through statements on talk pages or through usernames—or if they are seen to be focused on content deemed to be of greater interest to minority communities such as biographies of women or racial minorities.⁸²

Survey respondents identified differences in point of view (30%), administrative actions or status (26%) and edits or content (21%) as the top three reasons for harassment. Editing mistakes are most likely to impact new volunteers who may not be as familiar with both written and cultural norms. This can result, in the words of one interviewee, in:

Systemic harassment if you are a new editor and not familiar with social norms, even related to minor errors. People will show up and say, “get off our project, your content is crappy.” When this happens, your only experience will be one of significant hostility.⁸³

⁷⁶ Wikimedia: “Harassment Survey 2015 Results Report” (2015)

⁷⁷ Wikimedia Foundation and Jigsaw: “Ex Machina: Personal Attacks Seen at Scale” (2017)

⁷⁸ Wikimedia: “Harassment Survey 2015 Results Report” (2015)

⁷⁹ Wikimedia: “Community Insights” (2019)

⁸⁰ Wikimedia: “Harassment Survey 2015 Results Report” (2015)

⁸¹ Wikimedia: “Harassment Survey 2015 Results Report” (2015)

⁸² Interview with Foundation staff in June 2020

⁸³ Interview with Foundation staff in June 2020

When asked to describe the harassment they experienced, survey respondents detailed the following examples (please note abusive content follows):⁸⁴

- ▶ *"I'm going to kill your grandchildren"*
- ▶ *"All queers will be shot! You fucking faggot, I hope you burn in Hizzell!"*
- ▶ *"What entitles a feminized nebbish like you to delete a book that you haven't even read"*
- ▶ *A harasser "hurtfully mocked me for my gender and an illness."*

In addition, survey respondents reported the following kind of harassing behavior:⁸⁵

- ▶ *"User promised to kill me."*
- ▶ *"A user accused me of working for the KGB ..."*
- ▶ *A harasser "had an explicit pornographic website created based on my username"*
- ▶ *"Anti-Semitic slurs and cartoons, Twitter dog piling, off-wiki threats"*
- ▶ *"Legal threats, on-wiki statements using my real name stating that I acted illegally and corruptly"*
- ▶ *"IP editor attempted to link my name to a sexual criminal in that subject's Wikipedia article"*
- ▶ *"Someone edited Wikipedia articles about criminals and replaced their names with mine."*
- ▶ *"My email was flamed, my personal name posted without permission, many accounts were created to impersonate and embarrass me. [information redacted]. I think that someone paid freelancers to disrupt the article and attack me personally and make me appear unreasonable."*

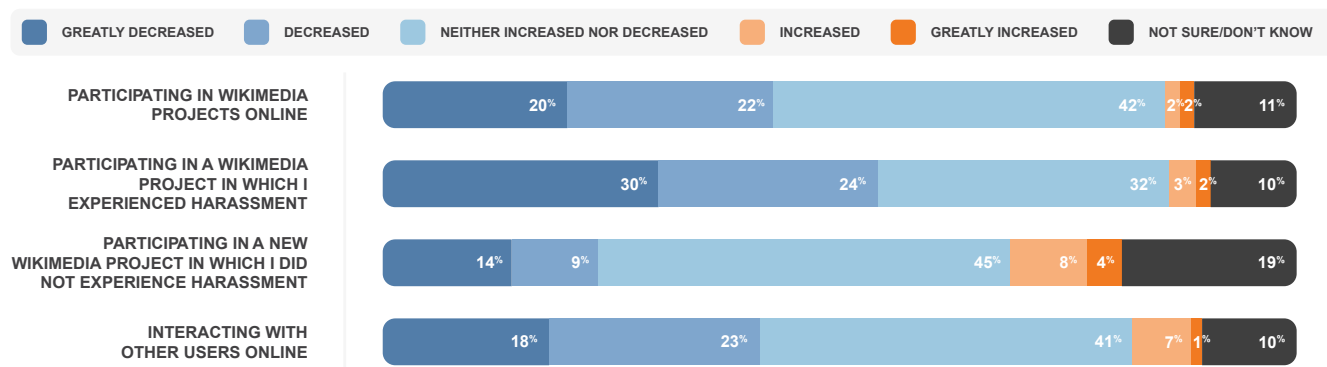
▶ *"Received a phone call on [my] work number from [name redacted], who threatened to phone my employer and try to get me fired."*

In many of these cases, Wikipedia articles were used as the tool for harassment. By editing pages to defame others, harassers directly infringe on the right to be free from unlawful attacks on one's honor and reputation (ICCPR 17). It is important to note that one need not be a Wikimedia volunteer for this type of harassment to occur. Indeed, anyone can be impacted by this type of action: staff, contractors, volunteers, readers, and non-readers. In addition, this type of harassment can lead to offline harm perpetrated by others who may not be aware the information is false. A classic example of this is the 2016 US "pizzagate" conspiracy theory where false information (not on Wikimedia projects) about a restaurant running a trafficking and child exploitation ring resulted in an individual firing a rifle into the restaurant to break up the "ring."⁸⁶

In addition to defamation, Wikimedia staff also reported being aware of cases where volunteers were blackmailed by other volunteers to behave in certain ways.⁸⁷

The majority of harassment survey respondents reported that their harassment was at least somewhat upsetting, with 14% reporting it as being extremely upsetting.⁸⁸ Despite this, 56% of respondents reported ignoring the harassment, while 45% requested the harasser cease their abuse. Importantly, 42% of the respondents felt that their reactions were not effective at all, 17% felt that their reaction was a little effective, 19% felt that it was somewhat effective, and 13% felt that it was mostly effective.⁸⁹

When asked how the harassment affected their engagement on Wikimedia projects, participation levels were unaffected for 23%-45% of the respondents and greatly decreased for 14% - 30% of the respondents.



⁸⁴ Wikimedia: "Harassment Survey 2015 Results Report" (2015)
⁸⁵ Wikimedia: "Harassment Survey 2015 Results Report" (2015)
⁸⁶ Wikipedia: "Pizzagate conspiracy theory"
⁸⁷ Interview with Foundation staff in June 2020
⁸⁸ Wikimedia: "Harassment Survey 2015 Results Report" (2015)
⁸⁹ Wikimedia: "Harassment Survey 2015 Results Report" (2015)

According to research from the Berkman Klein Center, harassing behavior and harmful content is more likely to persist on article talk and user talk pages than on Wikipedia articles. One reason for this is that the Cluebot NG, an AI tool to tackle Wikipedia vandalism, does not operate on talk pages and there are fewer editors focused on policing harmful content on talk pages. According to the Center's research, while administrators may choose to close toxic conversation threads to address harassing behavior, "the policies and guidelines that govern harassment, personal attacks, and incivility are not interpreted and enforced consistently across the community."⁹⁰ Indeed, unless blatantly offensive language is used, volunteers report that harassing comments are less likely to be policed, highlighting the more permissive culture around "borderline" statements. However, given the use of veiled yet highly harassing language, what constitutes a "borderline" statement may not be fully understood by administrators.⁹¹

■ Risk Mitigation Measures

Most efforts to govern harassment behavior by volunteers takes place at the community level. The governance structures differ between each project's public facing content (e.g. Wikipedia articles) and content on article talk pages, user pages, and user talk pages where oversight is more limited.⁹³

English-language Wikipedia is one of the most studied projects from a governance perspective. Research by Berkman Klein found that:⁹⁴

Wikipedia simultaneously operates multiple regulatory regimes that employ different sets of tools and have different objectives. One regime guides the actions of its volunteer editors in the creation and maintenance of the encyclopedia. Another mediates the interpersonal conduct of the editors while they do their work. Both operate based on a detailed set of guidelines and policies that are meant to reflect the social norms that have emerged over the many years of the project but that are implemented in a decentralized way, which offers a lot of flexibility and autonomous judgment in their application.

However, even in the case of English-language Wikipedia with a robust and long-standing approach to self-governance, "governing discourse among Wikipedians continues to be a major challenge for Wikipedia and one not fixed by content

B. Harassment of Foundation Staff

Foundation staff reported instances where employees and contractors were directly targeted and harassed by Wikimedia readers or volunteers – both on and offline. According to one staff member: "We have had staffers who have been mercilessly harassed by users when users didn't like certain policies."⁹² Staff also reported that higher profile employees, including Foundation leadership, are often the targets of online harassment, given their prominent public figures. In some cases, the harassment has resulted in physical danger to employees and contractors and as such required the involvement of local law enforcement.

*removal alone.*⁹⁵ Indeed, research from Cornell University found that content removal of personal attacks and harassing comments are removed at much higher rates than initially expected, suggesting that content removal alone does not address the concerns of volunteers who experience or witness harassment.⁹⁶

To address this, the Foundation has embarked on a Community Health Initiative to help the Wikimedia volunteer community "reduce the level of harassment and disruptive behavior on our projects."⁹⁷ The initiative is focused on:

1. **Policy enforcement and growth** including: a) working with communities to ensure their user conduct policies are clear, effective, and enforceable; and b) providing analysis of how behavioral issues are covered in policy and enforced in the community.
2. **Anti-Harassment tools**, including tools to detect, report, evaluate, and block such as the development of machine learning models to detect personal attacks and aggressive tones in article talk pages.⁹⁸

While the Foundation has a Terms of Use policy describing the rights and responsibilities that guide the Foundation and its users, volunteers are instructed to follow policies and guidelines set by each individual Wikimedia project.⁹⁹

⁹⁰⁻⁹¹ Berkman Klein Center for Internet & Society: "Content and Conduct: How English Wikipedia Moderates Harmful Speech" (2019)

⁹² Interview with Foundation staff in June 2020

⁹³⁻⁹⁵ Berkman Klein Center for Internet & Society: "Content and Conduct: How English Wikipedia Moderates Harmful Speech" (2019)

⁹⁶ Cornell University: "WikiConv: A Corpus of the Complete Conversational History of a Large

Online Collaborative Community" (2018) The research found that nearly 33% of toxic comments are removed within a day and over 82% of severely toxic comments are deleted within a day.

⁹⁷⁻⁹⁸ Wikimedia: "Community Health Initiative"

⁹⁹ Berkman Klein Center for Internet & Society: "Content and Conduct: How English Wikipedia Moderates Harmful Speech" (2019)

On English language Wikipedia, these policies include bans on:

- ✦ Personal attacks, including derogatory phrases directed against another editor or group of editors based on race, sex, sexual orientation, gender identity, age, religious or political beliefs, disabilities, ethnicity, nationality, etc.
- ✦ Linking to external attacks or harassment.
- ✦ Comparing editors to Nazis, communists, terrorists, dictators, or other infamous people.
- ✦ Accusations of inappropriate behavior by an editor without evidence to support the claim.
- ✦ Threats of legal action, threats of violence, threats to reveal personal info about an editor, or threats of actions that may expose editors to political, religious, or other persecution by a government, employer, or others.¹⁰⁰

Despite this policy, as of 2020 harassment continued, and one reason may be a lack of effective grievance channels. As outlined in the conclusion, volunteers have reported concerns about limitations in private grievance channels to raise and resolve concerns. As one volunteer stated: *“Why would anyone ever come forward to say ‘I’m being harassed’ on this site, ever?”*

*The only thing that happens is that people get dragged before the court of public opinion and told that everything they feel and experience is invalid.*¹⁰¹ This perspective was supported by 2017 research from the Wikimedia Foundation and Jigsaw which found that only 17.9% of personal attacks led to a warning or ban against the harasser.¹⁰² Indeed, there have been instances where *“there is a contributor who has been around for a long time with a lot of friends where the self-governance process starts to break down and the Foundation should step in.”*¹⁰³ In a statement to Slate, the Foundation reported that it only takes these steps under *“very particular circumstances, where there is a gap in the community’s ability to successfully address a known challenge, or for legal reasons.”*¹⁰⁴

Given the ongoing concerns around harassment, the Foundation’s Board of Directors voted in 2020 to ratify new trust and safety standards for all Wikimedia projects. These standards are designed to *“address harassment and incivility within the Wikimedia movement and create welcoming, inclusive, harassment-free spaces in which people can contribute productively and debate constructively.”*¹⁰⁵ As part of these standards, the Board instructed the Foundation to develop a Universal Code of Conduct (UCoC) that will be a binding minimum set of standards for conduct across all projects.

Recommendations

Issues	Human Rights	Recommendation
Volunteer Harassment	<ul style="list-style-type: none"> → UDHR 1 → UDHR 2 → UDHR 3 → UDHR 21 → ICCPR 17 	<ul style="list-style-type: none"> ✦ Develop and deploy training programs for admins and volunteers with advanced rights on detecting and responding to harassment claims. ✦ In line with recommendations from the 2015 Harassment survey, explore opportunities to rate the toxicity of user behavior, helping to identify repeat offenders and patterns of harassment. Consider awards for projects with the lowest toxicity levels. ✦ Consider developing admin metrics focused on enforcing civility and applying the forthcoming UCoC. ✦ Ensure that the UCoC and its accompanying governance mechanism is reviewed by human rights experts, including experts on free expression and incitement to violence.

¹⁰⁰ Berkman Klein Center for Internet & Society: [“Content and Conduct: How English Wikipedia Moderates Harmful Speech”](#) (2019)

¹⁰¹ Slate: [“Wikipedia’s ‘Constitutional Crisis’ Pits Community Against Foundation”](#) (2019)

¹⁰² Wikimedia and Jigsaw: [“Ex Machina: Personal Attacks Seen at Scale”](#) (2017)

¹⁰³ Interview with Foundation staff in June 2020

¹⁰⁴ Slate: [“Wikipedia’s ‘Constitutional Crisis’ Pits Community Against Foundation”](#) (2019)

¹⁰⁵ Wikimedia: [“Wikimedia Foundation Board announces Community Culture Statement, enacts new standards to address harassment and promote inclusivity across projects”](#) (2020)

Issues	Human Rights	Recommendation
		<ul style="list-style-type: none"> ◆ Continue efforts to develop AI and ML tools to detect and flag abusive and harassing behavior by volunteers. ◆ Research which content verticals are most likely to trigger harassing and abusive behavior (e.g., science and biographies of women). This can help prioritize limited resources and support a greater ability to detect harassing behavior. ◆ Develop more robust grievance mechanisms as outlined in the Conclusion.
<i>Staff Harassment</i>	<ul style="list-style-type: none"> → UDHR 1 → UDHR 2 → UDHR 3 → UDHR 12 	<ul style="list-style-type: none"> ◆ Explore benefits and drawbacks of allowing staff to opt out of public profiles on the Wikimedia Foundation website and Phabricator. ◆ Ensure staff are appropriately trained on the types of harassment they may receive and resources available to them through the Foundation.
<i>Staff Harassment</i>	<ul style="list-style-type: none"> → UDHR 1 → UDHR 2 → UDHR 3 → UDHR 12 	<ul style="list-style-type: none"> ◆ Commission a “social norms marketing” research project to assess what type of messaging is likely to reduce and prevent harassing comments and actions. Social norms marketing has been found to help address social ills including domestic violence and alcohol abuse.



Risk

Government Surveillance & Censorship

- Overview
- Analysis
- Mitigation Measures
- Recommendations

As outlined in the UDHR, privacy and free expression are fundamental human rights. Online platforms, including Wikimedia’s free knowledge projects, are increasingly being used by governments to track and monitor the activities of political dissidents and human rights defenders (HRDs) and to request the illegitimate removal of certain content. While the right to privacy and free expression can be legally overruled, for example in cases of legitimate national security risks, for infringements to

be compliant with international human rights frameworks they have to be “*prescribed by law, necessary to achieve a legitimate aim, and proportionate to the aim pursued.*”¹⁰⁶

Below, we outline the range of human rights that can be impacted by government surveillance and censorship on Wikimedia’s free knowledge projects.

Human Rights	Justification for Inclusion
<i>Right to security of persons (UDHR 3) and the right to be free from torture (UDHR 5)</i>	In a small number of cases, instances of illegitimate surveillance can lead to offline harm, including torture of HRDs and political dissidents by government forces.
<i>Privacy (UDHR 12)</i>	Illegitimate surveillance infringes directly on the right to privacy – whether online or offline.
<i>Expression & Assembly (UDHR 19 and 20)</i>	Illegitimate surveillance may have a chilling effect whereby contributors self-censor and choose not to add truthful information for fear of reprisal. Government censorship can infringe on the right to free expression and to seek and impart information by limiting the ability for individuals to express and seek information online.
<i>To take part in government (UDHR 21)</i>	For an election to be free and fair, voters need to have accurate information about the parties, candidates, and issues when they vote. Government censorship can limit the ability for voters to secure information to inform their votes.

Overview

Human rights are being challenged around the world, especially in relation to free expression, freedom of the press, internet blackouts, internet content controls, and crackdowns on human rights defenders. Freedom House, which reviewed digital rights in 65 countries in 2019 for its Freedom on the Net report, diagnosed a considerable rise of digital authoritarianism in 39 countries, especially in countries key to the Foundation’s medium-term plan.¹⁰⁷

¹⁰⁶ EFF: “Necessary and Proportionate”

¹⁰⁷ Freedom House: “Freedom on the Net” (2019)

For the Foundation, these infringements on rights manifest in three primary ways:

- A. Online Surveillance of Wikimedia Volunteers and Readers
- B. Requests for User Data
- C. Government Censorship

We describe each of these risks below.

■ Analysis

A. Online Surveillance of Wikimedia Volunteers and Readers

In 2018, an internal Wikimedia Foundation report “*observed a rise in cases of Wikimedia volunteers experiencing pressure by authoritarian governments.*”¹⁰⁸ Interviews with Foundation staff corroborated this finding, highlighting the increased attention volunteers have received – especially in countries with restrictions on internet freedom. In the words of one staff member: “*Wikimedia volunteers do large and small acts of bravery to make sure the world has access to knowledge.*”¹⁰⁹ These acts range from contributing content on topics that are considered taboo in their countries, such as LGBTQ rights, to editing political pages that may anger political elites.

In 2018 the Foundation supported cases against government surveillance and pressure on Wikimedia volunteers:¹¹⁰

- ◆ Members of a Wikimedia organization in one country were doxed after a government allegedly tried to block access to Wikipedia;
- ◆ A Wikimedia volunteer was targeted by a government’s state security after organizing movement events; and
- ◆ A government pressured a local Wikimedia user group to take down content on the local language Wikipedia.

Foundation staff reported being aware of individual contributors being arrested or questioned for their contributions.

One form of government surveillance occurs through online logs involving contributor IP addresses. Any contributor who chooses not to register a Wikimedia account or is not logged into an account will have their IP address publicly and permanently logged as part of a page’s edit history, a functionality that was

incorporated early on into the software that underlies Wikimedia platforms and was originally created to help address and prevent vandalism on the site. According to multiple staff, this approach is counter-intuitive from a privacy perspective given that individuals with an account do not have their IP addresses publicly logged. This can result in privacy-sensitive volunteers potentially selecting the riskier option without full awareness and understanding.¹¹¹ In some cases, volunteers may support each other in navigating risks. However, there is no formal approach to upskilling volunteers or a central hub with relevant information to help contributors protect themselves.

B. Requests for User Data

Human rights standards outline the right to be “*free from illegal or arbitrary interference with the right to privacy*” and to “*have the right to the protection of the law against such interference or attacks.*”¹¹² For this right to be respected, requests for user data must go through legal channels, be proportional to the potential risk, and necessary for the protection of other rights. When these factors are in place, there may be legitimate rights-compatible reasons for governments to request user data.

According to the Foundation’s transparency reports, it received 60 requests for user data in 2019. The Foundation granted two of these requests. The 60 requests included informal non-government requests and informal government requests as well as requests that were made through legal avenues, including: civil, criminal, and administrative subpoenas; search warrants; court orders and national security requests.¹¹³ These requests only include those made directly to the Foundation, giving the Foundation visibility into such requests. It does not include requests made to Wikimedia chapters or individual administrators, unless those were escalated to the Foundation.

¹⁰⁸ Wikimedia Foundation: “Voices Under Threat Protocol”

¹⁰⁹ Interview with Wikimedia Foundation staff in June 2020)

¹¹⁰ Wikimedia Foundation: “Voices Under Threat Protocol”

¹¹¹ Interview with Wikimedia Foundation staff in June 2020

¹¹² Global Network Initiatives: “Implementation Guidelines”

¹¹³ Wikimedia Foundation: [Transparency Report](#) (2019)

Wikimedia volunteers handling non-public data have increasingly heightened risk profiles. There are also certain groups of users (such as Checkusers) who are entrusted by the community to have access to non-public data. These users may create an additional risk that data is disclosed.

C. Government Censorship

Wikimedia, like all large content platforms, receives government requests for content removal and alteration. In some cases, these restrictions may be appropriate under human rights standards, for example requesting the removal of content designed to promote terrorist activity.

In 2019, the Foundation received 17 government requests of which zero were actioned by the Foundation.¹¹⁴ However, government actors have other avenues to influence Wikimedia content. Research by the Berkman Klein Center for Internet and Society found efforts from governments around the world have sought to censor various Wikipedia language projects.

In some cases, censorship is focused on specific pages. For example, Turkey blocked a selection of articles related to reproductive biology, as well as at least one political article prior to the Foundation implementing HTTPS. In the UK, several ISPs blocked access to a page about a music album that included album art of a naked child.¹¹⁵

In other cases, government concerns about Wikimedia content have resulted in wholesale blocks of the website. Russia, for example, has intermittently blocked access to Wikipedia reportedly due to concerns about pages related to marijuana.¹¹⁶ In Turkey, the government banned all language versions

of Wikipedia in April 2017, citing two articles that were well-sourced and factual: Foreign involvement in the Syrian Civil War and State-sponsored terrorism.¹¹⁷ The sites remained blocked until the Turkish Constitutional Court ruled in favor of Wikimedia at the end of 2019, and the ban was lifted in early 2020.

Iran has intermittently blocked access to the HTTPS version of Wikipedia since it was introduced in 2011 and the English and Kurdish versions of the site have also been temporarily blocked. The government has also filtered over 1,000 articles, 400 of which contained political content. Berkman Klein's study found that Iran's filtering of Wikipedia is in part keyword-based and is triggered when users request URLs that match a blacklist of terms. Importantly, the transition to HTTPS-only access in 2015 is likely to have "substantially affected the Iranian government's ability to censor Wikipedia articles."¹¹⁸

When it comes to China, the Berkman Klein Center's research found that "China was likely censoring the Chinese language Wikipedia project" and that Chinese censors have a "long and contentious history with Wikipedia."¹¹⁹ Indeed, the Chinese government temporarily blocked Chinese Wikipedia during the anniversary of the Tiananmen Square protests and massacre and the government was found to do article-level filtering of sensitive content dating back to 2006. Importantly, the introduction of an HTTPS version of Wikimedia in 2011 temporarily gave users in China full access to the project.¹²⁰ However, access to Chinese Wikipedia was later blocked and since April 2019 all versions of Wikipedia have been blocked in China.¹²¹

In each of these cases, the actions of governments directly infringed on the right to free expression and to information (UDHR 19).

■ Risk Mitigation Measures

The Wikimedia Foundation and its volunteers are guided by the Foundation's Privacy Policy, which was created in close consultation with volunteer editors and designed to be easily read and understood.¹²² The policy allows users to edit Wikimedia projects without creating an account or providing either an email address or their real name. The policy further outlines that the Foundation collects certain types of data when volunteers make public contributions, register an account, or

when project users access the Wikimedia sites. Minimal user data is collected and maintained on Foundation servers in the US, most of which is anonymized and deleted after 90 days. Additionally, the Foundation never sells user data.¹²³

As of 2020, the Foundation was exploring opportunities to strengthen privacy protections through an IP Masking project. This project seeks to secure a balance between the need to

¹¹⁴ Wikimedia Foundation: [Transparency Report](#) (2019). Importantly, the vast majority of content removal or alteration requests come from non-state actors.

¹¹⁵⁻¹¹⁶ Berkman Klein Center for Internet & Society: ["Analyzing Accessibility of Wikipedia Projects Around the World"](#) (2017)

¹¹⁷ The Guardian: ["Turkey's Wikipedia block violates human rights, high court rules"](#) (2019)

¹¹⁸⁻¹²⁰ Berkman Klein Center for Internet & Society: ["Analyzing Accessibility of Wikipedia Proj-](#)

[ects Around the World"](#) (2017)

¹²¹ Wikipedia: ["Censorship of Wikipedia"](#)

¹²² Wikimedia: ["Privacy Policy."](#)

¹²³ Additional caching servers are located in the Netherlands and Singapore. Wikimedia: ["Wikimedia Data Centers"](#)

keep vandalism and harassment at bay and a recognition that restricting access to IP addresses advances the fundamental right to privacy and in turn diminishes offline threats from government interference in high human rights risk countries.¹²⁴

As for censorship risks, the Foundation has implemented HTTPS technology and taken steps to fight content removal

and alteration requests and advocate against filtering and blocking content.

Finally, in 2020 the Foundation has started the process to hire a Human Rights Lead and clarified roles with Wikimedia volunteers to increase accountability.¹²⁵

■ Recommendations

Based on the analysis above, we would recommend the following for the Foundation's consideration:

Issues	Human Rights	Recommendation
<i>Privacy</i>	→ UDHR 12	<ul style="list-style-type: none"> ◆ Continue efforts underway as part of the IP Masking project to further protect users from public identification.
<i>Volunteer Awareness Raising</i>	<ul style="list-style-type: none"> → UDHR 3 → UDHR 5 → UDHR 12 → UDHR 19 	<ul style="list-style-type: none"> ◆ Develop awareness raising tools and programs for all volunteers to understand and mitigate risks of engaging on Wikimedia's free knowledge projects. Tools should be made publicly available and should be translated into languages spoken by volunteers in higher risk regions. Higher risk regions can be determined based on historical knowledge from the Foundation combined with country rankings on human rights and internet freedom, including for example Freedom House's Freedom on the Net report. ◆ Develop a prioritized list of countries (for regular review) where additional capacity building is required; invest additional resources to educate and empower volunteers in the region, including through partnership with local and regional digital rights organizations.
<i>Free Expression & Privacy</i>	<ul style="list-style-type: none"> → UDHR 12 → UDHR 19 	<ul style="list-style-type: none"> ◆ Continue efforts underway as part of the IP Masking project to further protect users from public identification. ◆ In line with the Foundation's new GNI membership, incorporate the GNI Principles into the Foundation's response to government requests.
<i>Staff Capacity Building</i>	<ul style="list-style-type: none"> → UDHR 3 → UDHR 5 → UDHR 12 → UDHR 19 	<ul style="list-style-type: none"> ◆ Build the capacity of staff to identify veiled and coded language that could suggest additional risks and ensure all staff are advised how to escalate these reports.
<i>Grievance Channel</i>	<ul style="list-style-type: none"> → UDHR 3 → UDHR 5 	<ul style="list-style-type: none"> ◆ Develop a private and confidential channel for volunteers to report concerns. Given limited resources, this channel should be narrowly communicated to high-risk groups as well as administrators with additional powers and responsibility who can funnel concerns to the Foundation.

¹²⁴ Wikimedia: "Privacy Enhancement and Abuse Mitigation"

¹²⁵ Article One email communication with Foundation staff in June 2020



Risk
Risks to Child Rights

- Overview
- Analysis
- Mitigation Measures
- Recommendations

The Convention on the Rights of the Child (CRC) states that “the child shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of the child’s choice.”¹²⁶ In addition, the Optional Protocol to the CRC outlines the urgent need to eliminate both “child prostitution” and “child pornography.”¹²⁷

The internet has increased the ability for children to participate in civic engagement while also increasing the risk of certain forms of sexual exploitation of children. Online platforms can increase the speed of the grooming process, “partly because offenders can pretend to be children initially.”¹²⁸

The following child rights may be impacted on Wikimedia projects:

Human Rights	Justification for Inclusion
Dignity (UDHR 1)	Child sexual exploitation and harmful content attack the fundamental right to dignity.
Privacy (UDHR 12)	Child pornography impacts on the right to privacy.
Free Expression (UDHR 19) Education (UDHR 26)	If Wikimedia does not create a safe and welcoming space for children, it can limit their free expression and access to information which can in turn impact their right to education.
Protection from Harmful Content (CRC 17)	Harmful content on Wikipedia may impact on children’s rights to access information.
Protection from sexual exploitation and abuse (CRC 34)	Human rights law provides for the protection of children from sexual exploitation and abuse, including “child pornography”.
Be free from unlawful attacks on one’s honor and reputation (ICCPR 17)	Child and adult victims of sexual exploitation may have their reputation tarnished by the release of intimate images.

Overview

According to guidance from UNICEF:

*Companies can fulfill their respect for children’s civil and political rights by ensuring that technology, legislation, and policies developed to protect children from online harm do not have the unintended consequences of suppressing their right to participation and expression or preventing them from accessing information that is important to their well-being.*¹²⁹

¹²⁶ UN: “Convention on the Rights of the Child”

¹²⁷ UN: “Optional Protocol to the Convention on the Rights of the Child on the sale of children, child prostitution and child pornography” (2002)

¹²⁸ UN: “Optional Protocol to the Convention on the Rights of the Child on the sale of children, child prostitution and child pornography” (2002)

¹²⁹ UNICEF: “Children’s Rights and the Internet From Guidelines to Practice” (2016)

This requires balancing the risk of online platforms being misused by bad actors to harass, groom and abuse children with the recognition that platforms can be powerful tools to advance the realization of child rights.

Editors of all ages are welcome to contribute to Wikimedia projects. Indeed, some people who have served in important roles in the Wikimedia communities of editors later disclosed that they had been minors at the time.¹³⁰ However, risks to children remain, including:

- A. Privacy and reputation risks
- B. Exposure to harmful content
- C. Child sexual exploitation material
- D. Harmful contact

Below we explore each of these risk areas.

■ Analysis

A. Privacy and Reputation Risks

Research from UNICEF has found that there are three primary risks associated with children's privacy and reputational rights in the online world. These include:

1. Unauthorized use of children's images, including in cases where children have voluntarily uploaded a photograph without understanding its potential applications on other platforms;
2. Bullying and harassment, including sharing images and information about a child in attempt to dishonor them; and
3. Permanence of information shared on the internet, whether by the child or others, that "*creates public online representations of children's lives about which they may neither know nor feel comfortable.*"¹³¹

When it comes to risks on Wikimedia, a primary risk relates to children being the focus of content, for example biographies on child stars. Given their young age, children deserve additional privacy protections and protections against adverse impacts on their reputation. According to the Wikimedia Foundation, editors have vandalized articles about living people and have made edits "*designed to smear others.*"¹³² The degree to which these smear campaigns have targeted children is unclear, though email communication with Foundation staff suggests there have not been many complaints.¹³³

B. Exposure to Harmful Content

One way Wikimedia projects can adversely impact on child rights is by exposing children to harmful content. As with all online spaces children may either intentionally or inadvertently access harmful content while browsing the web. Examples of content that may be inappropriate include content that "*promotes substance abuse, racial hatred, risk-taking behavior or suicide, anorexia or violence.*"¹³⁴ The suicide methods article on Wikipedia, for example, may not be seen as an infringement on human rights for adult readers but may for child readers.¹³⁵

Child contributors to Wikimedia projects may also be exposed to harassing and harmful behavior in talk pages and other forums where content is policed at lower levels.

C. Child Sexual Exploitation Material

Child sexual exploitation material directly infringes on the right to be treated with dignity and to be protected from child exploitation. Online platforms have long been used by predators to seek and share exploitative content. The scale of this challenge has only increased under COVID-19 lockdown. Indeed, BBC reports that demand for abuse imagery has drastically increased in 2020. In the Philippines alone, reports of online abuse material increased from 59,000 in February 2020 to more than 101,000 in March of the same year – the same month the lockdown began.¹³⁶

¹³⁰ For example, one Wikipedia administrator began serving that role at the age of 11.

¹³¹ UNICEF: "[Children's Rights and Business in a Digital World: Privacy, Protection of Personal Information and Reputation](#)" (2017)

¹³² Wikimedia: "[Biographies of Living People](#)"

¹³³ Email communication with Wikimedia staff (July 2020)

¹³⁴ UNICEF: "[Children's Rights and the Internet From Guidelines to Practice](#)" (2016)

¹³⁵ Wikipedia: "[Suicide Methods](#)"

¹³⁶ BBC: "[Online child abuse rising during lockdown warn police](#)" (2020)

While most Wikimedia projects have strong guidelines against this type of content, there have been instances where child exploitation material has been found on Wikimedia sites.

D. Harmful Contact

UNICEF research outlines the challenge of harmful contact between adults and children online. Harmful contact includes

grooming children to perform sexual acts online or offline and access them as potential customers for illegal products such as drugs. As one country study suggests, online platforms are increasingly becoming the predominant channel for grooming. Police in England and Wales recorded more than 10,000 online grooming offenses on social media from 2017 to 2019.¹³⁷ When it comes to Wikimedia projects, the greatest risk lies with talk pages which may be used by predators to identify and build relationships with minor editors.

■ Risk Mitigation Measures

The Foundation and the Wikimedia volunteer community have taken several steps to mitigate risks to children on its projects.

These include:¹³⁸

- ◆ Project level commitments to block and ban any contributor who identifies themselves as a pedophile;
- ◆ Project level efforts to protect the privacy of children for whom there is limited encyclopedic need to cover, such as the children of celebrities; and

- ◆ Immediately removing any child sexual exploitation content that is identified and developing hash technology to remove known content across projects.

Despite these steps, as of 2020 there remained limited options for children and their guardians to report concerns directly to the Foundation or to access emergency resources in cases of grooming.

■ Recommendations

Issues	Human Rights	Recommendation
<i>Assessing Risks to Children</i>	<ul style="list-style-type: none"> → UDHR 1, 3 → CRC 6 → CRC 34 	<ul style="list-style-type: none"> ◆ Conduct a child rights impact assessment of Wikimedia projects, including conducting interviews and focus groups with child contributors across the globe.
<i>Harmful Content</i>	<ul style="list-style-type: none"> → CRC 17 → UDHR 1 → UDHR 3 	<ul style="list-style-type: none"> ◆ Explore options to limit child access to harmful content. For example, for pages such as Suicide Methods, there could be age gating blocks to increase friction. ◆ Consider developing and piloting a Kids' Wikipedia project in one market. Content would be specifically curated for children. Special consideration and mitigation would be needed to prevent the use of the project by child abusers. ◆ Deploy hashing technology to detect known exploitation content globally and across projects.

¹³⁷ BBC: "Facebook dominates cases of recorded social media grooming" (2020)

¹³⁸ Interview with Wikimedia Foundation staff in June 2020

Issues	Human Rights	Recommendation
<i>Harmful Contact</i>	<ul style="list-style-type: none"> → UDHR 1 → UDHR 3 → CRC 6 → CRC 34 	<ul style="list-style-type: none"> ◆ Explore options to develop emergency help buttons on talk pages for verticals children are most likely to edit. This should be paired with a robust governance process to support the Foundation in responding to these alerts and identifying false positives. ◆ Develop an “if you see something, say something” campaign regarding online grooming to help raise awareness of the risks and provide avenues to identify concerns.
<i>Digital Literacy</i>	<ul style="list-style-type: none"> → UDHR 3 	<ul style="list-style-type: none"> ◆ Create child safeguarding tools, including child-friendly guidance on privacy settings, data collection, reporting of grooming attempts, the forthcoming UCoC as well a “Child’s Guide to Editing Wikimedia Project” to help advance the right of children to be civically engaged. ◆ Develop parental controls to help parents keep children safe on Wikimedia projects.



Risk
Limitations on Knowledge Equity

- Overview
- Analysis
- Mitigation Measures
- Recommendations

In 2019, UNESCO put forth a recommendation to member states which supports the creation, use and adaptation of inclusive and quality Open Educational Resources (OER), such as the Wikimedia Foundation’s free knowledge projects, as a strategic approach for implementing SDG 4.¹³⁹ UNESCO’s third area of action is focused on knowledge equity, and seeks to encourage:

*effective, inclusive and equitable access to quality OER, for all stakeholders, including: learners in formal and non-formal education contexts irrespective of, inter alia, age, gender, physical ability, and socio-economic status, as well as those in vulnerable situations, indigenous peoples, those in remote rural areas (including nomadic populations), people residing in areas affected by conflicts and natural disasters, ethnic minorities, migrants, refugees, and displaced persons.*¹⁴⁰

In its comment on the draft recommendation, the Foundation emphasized the need for the recommendation to “*both represent and respect the knowledge of marginalized peoples.*”¹⁴¹ The Foundation further noted that the draft recommendation fell short of addressing “*ICT as a barrier to quality education for marginalized groups*” while recognizing that “*there are many people for whom access to ICT is limited whether because of a lack of infrastructure or digital literacy.*”¹⁴² Indeed, the elimination of barriers offered by OER solutions, does not ensure their equitable access or use, a problem the Foundation faces within its own free knowledge projects.

With this in mind, the following rights may be impacted by inequitable access to Wikimedia’s free knowledge projects:

Human Rights	Justification for Inclusion
<i>Freedom from Discrimination (UDHR 2)</i>	Perpetuating social exclusion of contributors based on their racial, ethnic, cultural, national origin, linguistic background or gender identity infringes on the human right to non-discrimination.
<i>Freedom of Expression and Right to Seek and Impart Information (UDHR 19)</i>	Perpetuating social exclusion of potential volunteers on the basis of race, ethnicity, culture, national origin, linguistic background or gender identity, infringes on their rights to seek, receive and impart information and ideas through any media and regardless of frontiers.
<i>Right to Cultural Participation (UDHR 27)</i>	Perpetuating social exclusion of potential volunteers in free knowledge projects infringes on their rights to participate in the documentation of cultural life.

Overview

The Wikimedia Foundation is committed to making information more accessible and bringing forward knowledge left out by systems of privilege and power.¹⁴³ In pursuit of these ambitions, the Foundation recognized that Wikimedia projects may disproportionately amplify some voices, while being inequitably accessible or useful to others. Some interviewees reflected on these challenges, sharing: “*What we have noticed in the last few years is there is a deep representation problem for women, and people across the globe. We don’t have the diversity we’d like to have, that will support our mission.*”¹⁴⁴ One interviewee cautioned, “*If it becomes the case where only people wealthy enough to have stable connectivity [can access the projects] then their speech gets amplified. This can privilege certain viewpoints. We want to make sure we are not exacerbating some of those issues.*”¹⁴⁵

¹³⁹ UNESCO, “Open Educational Resources” (2019)

¹⁴⁰ UNESCO, “Recommendation on Open Educational Resources” (2019)

¹⁴¹ Wikimedia Foundation, “Comment on the Draft OER Recommendation Text” (2019)

¹⁴² Wikimedia Foundation, “Comment on the Draft OER Recommendation Text” (2019)

¹⁴³ Berkman Klein Center for Internet & Society: “Will Wikimedia Exist in 20 Year?” (2017)

¹⁴⁴ Interview with Foundation staff in June 2020

¹⁴⁵ Interview with Foundation staff in June 2020

The Wikimedia Foundation’s ambition to provide for knowledge equity both reflects its responsibility to respect human rights and its commitment to promote rights. As one interviewee put it, “we’re inviting people to build knowledge. That invitation comes with a great deal of responsibility.”¹⁴⁶ The challenges that the Foundation is encountering in achieving their goals can be attributed to multidimensional causes, many of which impact both contributors and readers far before they interact with the Foundation or its projects. These challenges include:

- A. Gender Equity
- B. Racial & Ethnic Diversity
- C. Accessibility
- D. Knowledge Equity of the Global South

In the following section we outline each of these potential challenges. However, it is important to highlight where the Foundation’s responsibility under the UNGPs starts and ends. The Foundation is not expected or required to ensure every person’s right to access information is realized. Rather, there is a responsibility to promote equal access and a safe and welcoming environment for a diversity of voices. Efforts above and beyond these, for example to promote infrastructure investments in the global south, are considered promotion of human rights, and not a core expectation of the UNGPs.

Analysis

A. Gender Equity

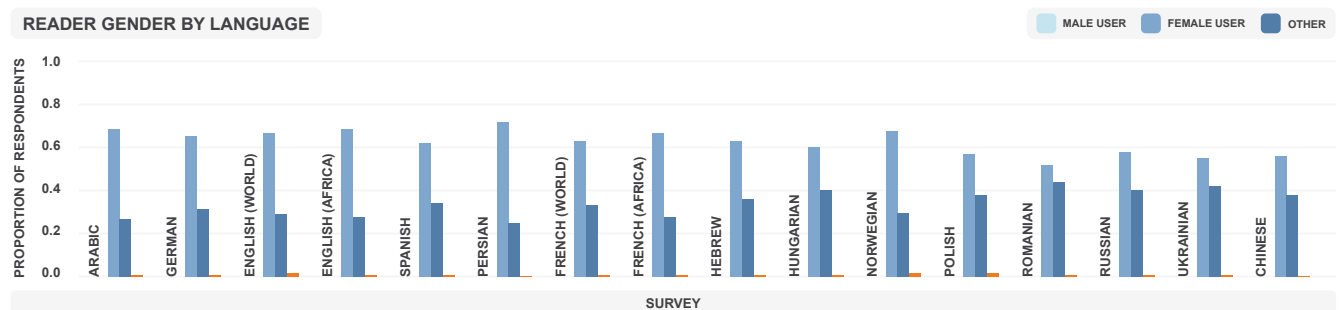
Gender bias remains a pressing concern among Wikimedia projects. This bias manifests as a) a lack of female contributors, b) limited coverage of women on the projects, and c) limited female readers.

Women are underrepresented as contributors to knowledge platforms. Research from 2015 on English-language Wikipedia found that less than 15% of contributors were women.¹⁴⁷ This may be due to:

- ◆ A lack of confidence in their knowledge or expertise as contributors¹⁴⁸
- ◆ An editing interface that is deemed not to be user friendly¹⁴⁹

- ◆ Discomfort editing other people’s work¹⁵⁰
- ◆ Fear of being criticized¹⁵¹ and harassed¹⁵²
- ◆ A feeling that their edits are likely to be reverted or deleted¹⁵³
- ◆ Lack of free time¹⁵⁴

In addition, gender bias on the platform also manifests through fewer and less extensive articles about women or topics important to women, and perhaps as a result, fewer women using Wikipedia as a resource. One 2011 study found that Wikipedia’s topical coverage of women was 16%, whereas representation of men was at 82%¹⁵⁵ and a survey on Wikipedia reader demographics shows significant differences between men and women.



Wikimedia Foundation: Characterizing Wikipedia Reader Demographics

¹⁴⁶ Interview with Foundation staff in June 2020

¹⁴⁷ Wikipedia, “Gender Bias on Wikipedia” (2020)

¹⁴⁸ Sex Roles, “Where are the Women in Wikipedia? Understanding Different Psychological Experiences of Men and Women” (2016)

¹⁴⁹ Sue Gardner, “Nine Reasons Why Women Don’t Edit Wikipedia (in their own words)” (2011)

¹⁵⁰ Sex Roles, “Where are the Women in Wikipedia? Understanding Different Psychological

Experiences of Men and Women” (2016)

¹⁵¹ Harvard Business Review, “Why Do So Few Women Edit Wikipedia?”

¹⁵² Interview with Foundation staff in June 2020

¹⁵³⁻¹⁵⁴ Harvard Business Review, “Why Do So Few Women Edit Wikipedia?” (2016) and Sue Gardner, “Nine Reasons Why Women Don’t Edit Wikipedia (in their own words)” (2011)

¹⁵⁵ International Journal of Communication, “Gender Bias on Wikipedia and Britannica” (2011)

Cultural norms in editing communities that regard gender as binary have manifested in harm done not only to contributors, but also in edit wars over certain content. The New York Times reported on a French Wikipedia article titled “Femme,” the French word for woman, in which there is a controversy over whether the first paragraph should refer to gender in addition to biological sex, and whether transgender women should be included in the definition of woman.¹⁵⁶ Likewise, when public figures have come out as transgender, non-binary, gender fluid, or gender non-conforming, volunteers have extensively debated whether the individual’s self-declared pronouns should be used.¹⁵⁷ Despite Wikipedia’s guidelines, articles about transgender or nonbinary individuals are often subject to vandals who revert their pronouns back to their gender assigned at birth.¹⁵⁸

Finally, Foundation staff raised concerns about page view data being “heavily male focused.” This can lead to volunteers prioritizing which pages to translate and in turn, potentially perpetuating biases on a global scale.¹⁵⁹

B. Racial & Ethnic Diversity

According to the New York Public Library’s Center for Research in Black Culture, “*there is a gap that exists when it comes to people of color on Wikipedia, both as subjects of articles and as contributors... The gap in entries related to black people gets worse when you look beyond the U.S. to the rest of the globe.*”¹⁶⁰ Similar concerns were raised regarding a lack of American Latino and Asian American contributors.¹⁶¹ Also less prominent are pages about Indigenous peoples, communities, and cultures. As of August 2018, there were 3,468 articles within the scope of the Indigenous Peoples of the Americas WikiProject, just 0.06% of the articles on English-language Wikipedia at the time.¹⁶² The limited number of editors from diverse racial and ethnic backgrounds can result in the potential that “*many topics may remain uncovered, or at the least these topics will not be given the attention they deserve... White males don’t always accurately portray topics that relate to minorities.*”¹⁶³

The lack of racial and ethnic diversity globally may be attributed to:

- ◆ An under-representation of Black, Indigenous, and People of Color within Wikipedia’s editor base¹⁶⁴
- ◆ A digital divide in the U.S. and globally¹⁶⁵

- ◆ A lack of secondary sources, which historically have been favorable towards and focused on white men¹⁶⁶
- ◆ The “notability” requirement, which relies on a degree of published documentation which “there is simply less of for women and minorities”¹⁶⁷
- ◆ The project’s founding which initially attracted lots of editors who were “tech-oriented” men¹⁶⁸

As for retention, the challenge of bringing on new editors to projects is exacerbated when recruiting writers of content that is considered racialized and systemically marginalized. One key challenge, according to the New York Public Library’s Center for Research in Black Culture, is maintaining engagement of new editors because those “*who predominate as Wikipedia editors aren’t always warm and nurturing to new editors.*”¹⁶⁹ For example, when new editors add content on Black history, their content may be deleted by established editors, and talk pages about changes shared by new editors can be overwhelming to navigate.¹⁷⁰ Harassment also increases for those who are culturally very or moderately different from their peer contributors.¹⁷¹

A final challenge arises from deliberate disruption on the platform, which occurs in articles and talk pages, and “*often comes to a flash-point in user space, when a user openly displays iconography from racist groups on their user page or signature.*”¹⁷²

C. Accessibility

The Wikimedia Foundation is committed to ensuring digital accessibility for people with disabilities.¹⁷³ Inclusive design for visual disabilities and impairments encompasses, but is not limited to, adjusting a project’s font size and readability, enabling fonts that reduce the unintentional mental movement of typographical characters, providing image and video content audio captions for blind users, and so on.

Another way Wikimedia projects promote accessibility is by contributing to and sharing content that improves access to medical information and other topics relevant to those with different accessibility levels. For example, articles on Wikipedia indexing different sign languages or on neurodiversity can provide people with relevant information to learn more about their disabilities, identify communities, and understand histories of how certain disabilities have been diagnosed

¹⁵⁶⁻¹⁵⁸ New York Times, “Wikipedia Isn’t Officially a Social Network. But Harassment Can Get Ugly.” (2019)

¹⁵⁹ Interview with Foundation staff in June 2020

¹⁶⁰ Fast Company, “Black History Matters, So Why is Wikipedia Missing So Much of It?” (2015)

¹⁶¹ El Telecote, “Why Don’t More Latinos Contribute to Wikipedia?” (2015)

¹⁶² Medium, “Doing the work: Editing Wikipedia as an act of reconciliation” (2018)

¹⁶³ El Telecote, “Why Don’t More Latinos Contribute to Wikipedia?” (2015)

¹⁶⁴ Wikipedia, “Racial Bias on Wikipedia” (2020)

¹⁶⁵ Fast Company, “Black History Matters, So Why is Wikipedia Missing So Much of It?” (2015)

¹⁶⁶⁻¹⁶⁷ Wikipedia, “Racial Bias on Wikipedia” (2020)

¹⁶⁸ New York Times, “Wikipedia Isn’t Officially a Social Network. But Harassment Can Get Ugly.” (2019)

¹⁶⁹⁻¹⁷⁰ Wikipedia, “Racial Bias on Wikipedia” (2020)

¹⁷¹ The survey further found that 9% of respondents attributed the grounds of their harassment as their ethnicity, and an additional 4% identified race as grounds for harassment. Wikimedia: “Harassment Survey 2015 Results Report” (2015)

¹⁷² Wikipedia, “Wikipedia: No Nazis” (2020)

¹⁷³ Wikimedia Foundation, “Accessibility Statement” (2020)

or socially understood.¹⁷⁴ However, as noted in the global inequity sub-section, information is not yet globally distributed. For instance, English Wikipedia’s article on Ableism helpfully provides the notice that “*the examples and perspective in this article deal primarily with the English-speaking world and do not represent a worldwide view of the subject.*”¹⁷⁵ In addition, the Foundation has recognized that to meet its accessibility and equity ambitions will require Wikipedia to go beyond written knowledge and begin to think about audio and visual approaches to sharing and contributing information.¹⁷⁶

D. Knowledge Equity of the Global South

The Foundation’s knowledge projects contribute to knowledge equity through its platforms every day. While the Foundation does not have a human rights responsibility to ensure knowledge equity, the projects play an important role in the knowledge ecosystem, helping to advance access to information globally. Wikimedia volunteers, and especially those from Africa and South-East Asia, largely agree that making knowledge accessible should be a top public policy priority for the movement.¹⁷⁷

At the same time, the Foundation’s “New Voices Synthesis” report found multiple barriers to consumption of Wikipedia, including:¹⁷⁸

- ✦ **Access and Affordability:** In order to access Wikipedia, users need the infrastructure to do so, which has financial costs including the cost of a device to access the site, and affordable data and internet to use it. Notably, many of those first coming online

are doing so via a mobile phone, which may make contributing knowledge difficult, especially given that the cost of mobile data is still a connectivity barrier for low-income users.¹⁷⁹

- ✦ **Awareness and Literacy:** Users must be aware of the website itself and have the free time to read the article. For users seeking to contribute, they must fulfill the former expectations, and acquire the digital literacy to understand that they can edit the site, given they have the time to learn the appropriate didactic citation, notability, and style requirements of Wikipedia.¹⁸⁰

- ✦ **Language:** Language representation and relevance remain barriers to the platform’s accessibility.¹⁸¹ Only 500 of the world’s 7000+ languages are represented online, with English and Chinese dominating.¹⁸² On Wikipedia, English-language Wikipedia is both informed by, and contributes to this systemic imbalance, and has been characterized by uneven and clustered geographies of coverage, predominately in Anglophone regions in the global north.¹⁸³ However, English Wikipedia is only one of more than 300 language projects and the Foundation is committed to supporting nascent language projects.

According to Whose Knowledge?, “*when marginalized communities cannot create in their own languages on the internet, this reinforces and deepens inequalities that already exist offline.*”¹⁸⁴ Several research studies have documented that stratification in internet experiences, awareness, and skills compound and reinforce each other, contributing to a positive feedback loop of amplification and exclusion.¹⁸⁵

Risk Mitigation Measures

Staff members acknowledged that “*the Foundation does a good job of creating a culture within the Foundation where equity is valued, and people are thinking about it in their work.*”¹⁸⁶ The Foundation has taken several steps to improve the equitable access of its projects. These steps include:¹⁸⁷

- ✦ Improving diverse recruitment during the 2018–19 fiscal year, in which 53% of new hires in the U.S. were women and 30% of new hires were Black/ African American, Hispanic/Latino, Asian, or Native American;¹⁸⁸
- ✦ Pivoting to experiences in geographic markets that lack content, or where the volunteer pool is limited;
- ✦ Setting goals around creating reliable software that meets contemporary expectations around user

- experience with emphasis on new users;
- ✦ Developing trainings for onboarding new contributors to certain projects;
- ✦ Offering scholarships to fund community members to attend Wikimania;
- ✦ Investing in research to identify challenges with knowledge equity projects;¹⁸⁹ and
- ✦ Developing the WebContent Accessibility Guidelines 2.0 to “*make sure the experience is as accessible as possible.*”¹⁹⁰

In sum, the Foundation’s strategic goal to achieve knowledge equity is a worthy challenge to undertake, one that will have to navigate socially and developmentally systemic factors that may impact its efforts.

¹⁷⁴ Wikipedia, “[Sign Language](#)” (2020) and Wikipedia, “[Neurodiversity](#)” (2020)

¹⁷⁵ Wikipedia, “[Ableism](#)” (2020)

¹⁷⁶ Wikimedia Foundation, “[New Voices Synthesis Report](#)” (2017)

¹⁷⁷ Wikimedia Foundation, “[Copy of Community Insights Report](#)” (2019)

¹⁷⁸ Wikimedia Foundation, “[New Voices Synthesis Report](#)” (2017)

¹⁷⁹ Wikimedia Foundation “[Strategy | Wikimedia Movement | New Voices Synthesis Report](#)” (2017)

¹⁸⁰ The Atlantic, “[The Lopsided Geography of Wikipedia](#)” (2016) and Journal of Communication, “[The Pipeline of Online Participation Inequalities: The Case of Wikipedia Editing](#)” (2018)

¹⁸¹ Wikimedia Foundation, “[New Voices Synthesis Report](#)” (2017) and Whose Knowledge?, “[Decolonizing the Internet’s Languages](#)” (2020)

¹⁸² Digital Culture & Society, “[Mapping Wikipedia’s Geolinguistic Contours](#)” (2019)

¹⁸³ Annals of the Association of American Geographers, “[Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty](#)” (2014)

¹⁸⁴ Whose Knowledge?, “[Decolonizing the Internet’s Languages](#)” (2020)

¹⁸⁵ Journal of Communication, “[The Pipeline of Online Participation Inequalities: The Case of Wikipedia Editing](#)” (2018)

¹⁸⁶⁻¹⁸⁷ Interview with Wikimedia Foundation staff in June 2020

¹⁸⁸ Wikimedia Foundation, “[Wikimedia Foundation diversity and inclusion information about our workers—2019 by the numbers](#)” (2019)

¹⁸⁹ Wikimedia Foundation “[Strategy | Wikimedia Movement | New Voices Synthesis Report](#)” (2017)

¹⁹⁰ Interview with Wikimedia Foundation staff in June 2020

■ Recommendations

Issues	Human Rights	Recommendation
<i>Gender Equity</i>	<ul style="list-style-type: none"> → UDHR 1 → UDHR 2 → UDHR 26 → UDHR 27 	<ul style="list-style-type: none"> ◆ Encourage intersectional efforts towards equity that considers gender as a component to other systems of power and privilege such as sexual orientation, ability, race, and ethnicity. This could include sponsoring edit-a-thons, providing infographics on how different editors can contribute, and coordinating awareness campaigns.
<i>Racial & Ethnic Diversity</i>	<ul style="list-style-type: none"> → UDHR 1 → UDHR 2 → UDHR 26 → UDHR 27 	<ul style="list-style-type: none"> ◆ Engage stakeholders on how the “notability” requirement may be shifted to be more inclusive of oral histories, and to identify what definitions resonate with under-represented communities. ◆ Consider creating a color-coding system for citations of sources with broader notability requirements, or hosting articles operating with these definitions in an addendum or separate project.¹⁹¹
<i>Global South</i>	<ul style="list-style-type: none"> → UDHR 1 → UDHR 2 → UDHR 26 → UDHR 27 	<ul style="list-style-type: none"> ◆ Adapt Wikimedia projects to be more accessible via mobile phones. ◆ Continue investing in machine translation solutions to improve capacity for human editors to translate Wikipedia articles; consider formalizing experimentation with an open source translation engine.
<i>Accessibility</i>	<ul style="list-style-type: none"> → UDHR 1 → UDHR 2 → UDHR 19 → UDHR 26 → UDHR 27 	<ul style="list-style-type: none"> ◆ Expand the efforts of the design team to create more accessible and inclusive projects, extending knowledge produced by volunteers with disabilities. ◆ Explore opportunities to provide captioning for videos hosted on Wikimedia projects.
<i>Strategies for the Foundation</i>	<ul style="list-style-type: none"> → UDHR 1 → UDHR 2 → UDHR 19 → UDHR 26 → UDHR 27 	<ul style="list-style-type: none"> ◆ Design and test technology (such as recommender systems and machine classifiers) to assist contributors in identifying and filling knowledge gaps. ◆ Host Knowledge Mapping workshops in partnership with Whose Knowledge? to develop stakeholder driven lists of content relevant to under-represented communities and incorporate findings into equity resource distribution.¹⁹² ◆ Assess the feasibility of targets for gender, LGBTQ+, racial and ethnic, and disabled individuals’ parity representation in Wikipedia articles. Indicators should focus on quantity, range of subjects, depth, and other relevant factors. ◆ Explore new methods of representing the value and credibility of contributors beyond their “edit-counts,” for example, expanding recognition to organizing events, or building relationships with local partners.¹⁹³ ◆ Provide a certificate editors can secure when contributing to Wikimedia projects. ◆ Support retention by developing peer support and mentoring for under-represented contributors. Engage established editors in advocating for the retention and improvement of content to promote the Foundation’s equity goals.¹⁹⁴ ◆ Consider leveraging machine learning tools to scan articles for language patterns indicating gender and racial bias for human review, including as it related to the use of pronouns for transgender and gender non-conforming individuals.

¹⁹¹ Whose Knowledge?, “Towards a Wikipedia for and From us All” (2019)

¹⁹² Wikimedia Foundation, “Grants: Projects/Whose Knowledge?/Final” (2018)

¹⁹³ Whose Knowledge?, “Towards a Wikipedia for and From us All” (2019)

¹⁹⁴ Whose Knowledge?, “Towards a Wikipedia for and From us All” (2019)

*Assessing the Human Rights Impacts
of Wikimedia Free Knowledge Projects*



Conclusion



V. Conclusion

The Wikimedia Foundation and its community of global volunteers are working towards a bold and transformative vision to make Wikimedia “the essential infrastructure of the ecosystem of free knowledge,” for all people in all places.¹⁹⁵ This vision is essential to advancing multiple rights, including the right to free expression and the right to access information.

In many ways, the Foundation’s approach to managing actual and potential risks related to its projects is one that is aligned with human rights. Indeed, the human rights framework places a strong emphasis on rightsholders (all those potentially impacted by a product or service) helping to design and develop mitigation tactics.

At the same time, there has been a recognition that in some cases the systems developed and implemented by volunteers may fail to appropriately protect all members of the Wikimedia community – from readers and contributors to Foundation staff. A clear example of this is the Foundation’s Board statement on harassment and toxic behavior which reads:¹⁹⁶

Harassment, toxic behavior, and incivility in the Wikimedia movement are contrary to our shared values and detrimental to our vision and mission. They negatively impact our ability to collect, share, and disseminate free knowledge, harm the immediate well-being of individual Wikimedians, and threaten the long-term health and success of the Wikimedia projects. The Board does not believe we have made enough progress toward creating welcoming, inclusive, harassment-free spaces in which people can contribute productively and debate constructively.

In these cases, the Foundation, as the host of the free knowledge projects, has a responsibility under the UNGPs to conduct adequate and ongoing human rights due diligence to surface and respond to risks on its projects.

The development of a Universal Code of Conduct is an important and powerful step to meeting the Foundation’s responsibility to protect human rights. The Code should be grounded in human rights norms and complemented with a robust grievance mechanism and response protocol to support volunteers who believe they have been the victim of Code violations. Indeed, a key limitation of the Foundation’s existing management of human rights-relevant risks is the lack of a human-rights compatible grievance process.

According to the UNGPs, “to make it possible for grievances to be addressed early and remediated directly, business enterprises should establish or participate in effective operational-level grievance mechanisms for individuals and communities who may be adversely impacted.”¹⁹⁷ As it stands, most grievances on Wikimedia projects are dealt with at the community level, often requiring victims to publicly out themselves, potentially impacting on their right to privacy.

In addition, for the few cases that had been escalated to the Foundation as of 2020, the level of satisfaction remains low. For example, the 2015 Harassment survey found that out of the 62 respondents who reached out to the Foundation for assistance, 56% indicated that they were overall dissatisfied by the support that they received, suggesting that existing mechanisms at both the community and Foundation level may be failing to effectively respond to concerns.

Effective grievance mechanisms are an essential tool in the human rights framework and while resourcing to date suggests the Foundation would not be well-equipped to handle a deluge in additional grievances, efforts should be made to create and resource a governance process that would be better equipped to handle an uptick in human rights-related concerns.

The combination of standard setting, including through the Universal Code of Conduct; product tools such as a potential help button for child-relevant project verticals; and greater accountability mechanisms, including more robust grievance channels will help position the Foundation to not only respect human rights but advance its mission of being the essential infrastructure of the ecosystem of free knowledge.

¹⁹⁵ Wikimedia: “2030 Strategy”

¹⁹⁶ Wikimedia: “Wikimedia Foundation Board announces Community Culture Statement, enacts new standards to address harassment and promote inclusivity across projects” (2020)

¹⁹⁷ UN: “UN Guiding Principles on Business and Human Rights”

■ Recommendations

Issues	Human Rights	Recommendation
<i>Foundation Accountability</i>	→ All	<ul style="list-style-type: none"> ◆ Develop a standalone Human Rights Policy that commits to respecting all internationally recognized human rights by referencing the International Bill of Human Rights. ◆ Consider adding an independent board member responsible for human rights and/or establishing a network of human rights advisors to inform ongoing human rights due diligence efforts.
<i>Due Diligence</i>	→ All	<ul style="list-style-type: none"> ◆ Establish a human rights lead within the Foundation to manage human rights risks across key functions within the Foundation, including Trust & Safety, Legal, Public Policy, Engineering, and Community Development. ◆ Conduct ongoing human rights due diligence to continually assess risks to rightsholders. A Foundation level-HRIA should be conducted every three years or whenever significant changes could have an effect on human rights.
<i>Grievance Mechanisms</i>	→ All	<ul style="list-style-type: none"> ◆ Develop rights-compatible channels to deal with human rights concerns, including private channels, and ensure awareness of the mechanism among relevant volunteers. ◆ Develop and implement training for community administrators, including Stewards, to support them in responding and escalating potential human rights violations. ◆ Provide additional resources to Trust and Safety to effectively respond to human rights related grievances. ◆ Ensure grievance processes are aligned with the UNGPs Effectiveness Criteria including that they are rights-compatible by offering, for example, private channels for human rights-related grievances.

*Assessing the Human Rights Impacts
of Wikimedia Free Knowledge Projects*



Appendix



◆ Appendix I: UN Guiding Principles on Business & Human Rights

In 2011, the UN Human Rights Council unanimously endorsed the UNGPs. The UNGPs recognize the state's ultimate duty to protect, and business' responsibility to respect human rights. These principles include guidance for both states and companies related to three core pillars:



Pillar 1, the State Duty to Protect, recognizes the State's duty to protect its citizens against corporate human rights abuses. Protection is best accomplished through robust laws that align with international human rights standards and a strong rule of law that ensures their enforcement.

Pillar 2 calls on companies and other organizations operating in similar capacities to publish a policy commitment in support of human rights and to "know and show" their respect for human rights by acting with due diligence. This includes:

1. Assessing actual and potential impacts, including through human rights impact assessments;
2. Integrating the findings of the assessment across the entire business and taking appropriate action to address adverse impacts; and
3. Tracking and communicating performance.

As part of the due diligence expectation, the UNGPs recognize that companies may need to prioritize which actual and potential impacts to address. However, these impacts should not be prioritized based on the company's relationship to an impact, but rather on its saliency, specifically on the degree of risk to rightsholders. Indeed, a key differentiator of the UNGPs is the focus on risks to rightsholders, rather than on risks to the business or organization.

Pillar 3 outlines the obligations of both states and companies to provide access to effective remedies in cases of human rights infringements. If the company is found to have caused or contributed to an impact, the company may be obligated to provide or facilitate access to a remedy. If the company is directly linked to an impact through a business relationship, there is no obligation to provide or facilitate access though the company may use its leverage to help ensure a remedy is provided.

ARTICLE ONE

