Preserving Endangered Articles on Wikipedia

Tianwa Chen, Gianluca Demartini
The University of Queensland, Australia

Abstract

Wikipedia serves as a crucial, crowd-sourced knowledge repository; however, persistent challenges, including underrepresentation, neglect, and deletion, disproportionately affect articles related to gender and marginalized communities. Such articles are "endangered", as they are particularly prone to being overlooked, undervalued, or deleted, posing a significant threat to the comprehensiveness and equity of the public knowledge landscape. Preserving endangered Wikipedia articles is critical to ensuring the sustainability, diversity, and inclusivity of the Wikipedia platform's content. This project proposes a novel, multifaceted approach that combines advanced data analytics with community-driven engagement to identify, prioritize, and safeguard at-risk articles effectively. By leveraging algorithms with process mining and process discovery approaches that analyze patterns in edit histories, view frequencies, and deletion records, the project seeks to identify, predict, and prioritize articles at high risk and facilitate the early detection of endangered content. In parallel, it fosters active collaboration with Wikipedia editors to implement targeted interventions and strategies that improve article quality, visibility, and resilience. By identifying and addressing at-risk articles, this research aims to ensure the long-term preservation, quality improvement, and resilience of Wikipedia content — with a particular focus on gender and marginalized topics that are especially vulnerable to neglect, bias, underrepresentation, and deletion.

Introduction

Wikipedia is among the most prominent platforms for disseminating free knowledge globally, yet it faces persistent challenges in maintaining the diversity, inclusivity, and comprehensiveness of its content. A significant issue lies in the underrepresentation and vulnerability of certain articles, particularly those related to gender and other marginalized topics. These articles frequently encounter threats such as low visibility, insufficient editorial activity, and a heightened risk of deletion. Evidence shows that content pertaining to gender, minority groups, or controversial social issues is disproportionately affected, thus perpetuating existing knowledge gaps and biases within the public information ecosystem.

Addressing this problem is critically important for Wikimedia projects, as one of the core missions revolves around providing equitable and comprehensive knowledge to users worldwide. Ensuring the preservation and enhancement of endangered articles aligns directly with Wikimedia's commitment to diversity and inclusivity, essential for achieving genuine representation across all areas of human knowledge. Furthermore, the loss of these articles diminishes Wikipedia's value as a diverse repository and undermines public trust in its neutrality and completeness.

This research aims to tackle this challenge through a multifaceted approach combining advanced data analytics and community engagement. Specifically, the research will address the following research questions:

(1) How can we effectively identify Wikipedia articles that are endangered, particularly in gender-related and marginalized topics?
(2) What are the key characteristics and patterns that distinguish endangered articles over time?
(3) How can we evaluate the content quality of at-risk articles to determine specific areas needing improvement, such as notability, citations, and comprehensiveness?
(4) How can community-driven approaches and strategies be effectively designed and implemented to increase the resilience and quality of endangered articles?

Through exploring and answering these research questions, the project seeks to develop practical strategies for Wikimedia communities to preserve endangered Wikipedia articles, ensuring the resilience, inclusivity, and reliability of the knowledge presented on Wikipedia.

Related work

Our previous project — "Measuring the Gender Gap: Attribute-Based Class Completeness Estimation"— showed that integrating statistical completeness metrics with insights from gender-focused Wikipedia editors can effectively surface gender gaps in articles (Demartini, 2023). By applying statistical models to quantify representation across demographic groups and conducting in-depth interviews with editor community dedicated to closing gender gaps, we uncovered two key insights: (1) Quantified Vulnerability. Articles about gender and other marginalized groups consistently were more prone to be flagged for deletion or deemed non-notable, confirming a measurable risk of content erosion. (2) Community Strategies & Pain Points. Editors fighting these gaps rely on labor-intensive monitoring, ad-hoc watchlists, and edit-a-thons — approaches that are invaluable but reactive and difficult to scale. They repeatedly stressed the need for early-warning signals and data-driven prioritization to safeguard vulnerable topics before they disappear. These findings directly motivate our new work on preserving endangered articles.

Upon reviewing the literature, previous studies have consistently highlighted disparities in representation and systemic biases within Wikipedia's content, particularly regarding gender and marginalized communities. Over the past decade, researchers have found significant gender gaps in Wikipedia editing, leading to imbalanced coverage of topics traditionally associated with women or marginalized communities (Lam et al., 2011; Patel et al., 2024; Wagner et al., 2015; Yunus et al., 2025). Findings reveal that articles about female figures and gender-related issues tend to be shorter, less detailed, and receive significantly fewer edits compared to those on male counterparts or mainstream topics (Wagner et al., 2015). These discrepancies reflect biases in editorial practices and also underline the vulnerabilities of such articles to neglect and eventual deletion. Particularly, researchers revealed that certain topical areas, particularly those addressing gender-related or minority subjects, experience higher deletion rates due to perceptions of insufficient notability or low community interest (Schneider et al., 2012; Taraborelli & Ciampaglia, 2010; Tripodi, 2023; Worku et al., 2020). These challenges disproportionately disadvantage

marginalized topics, widening content gaps and undermining equitable representation.

Recent research advocates for data-driven approaches to address the challenges of underrepresentation and deletion. To name a few, algorithmic methods for evaluating article quality were introduced, helping editors prioritize content improvement efforts (Warncke-Wang et al., 2013). Furthermore, researchers developed the automated topic model based on a labeling strategy that detects Wikipedia articles at risk of low-quality ratings (Asthana & Halfaker, 2018). However, these methods were not specifically tailored to gender or marginalized community topics, leaving a significant gap for improvement.

The application of process mining and process discovery to Wikipedia has been limited but shows promising potential for understanding and improving editorial workflows. Process mining has been proven as an effective approach for analyzing complex, event-driven workflows in crowd-sourced environments, highlighting its ability to pinpoint bottlenecks, frequent pathways, and workflow inefficiencies (W. van der Aalst, 2016). While process mining and process discovery methodologies have been effectively applied in business process management and organizational research contexts (Chen et al., 2020, 2023; W. M. P. van der Aalst et al., 2007), their application in Wikimedia editorial workflows remains underexplored.

Drawing on the existing literature, data-driven techniques, and the workflow insights afforded by process mining and process discovery, we therefore propose a multifaceted preservation framework that combines predictive analytics with co-designed community interventions to safeguard and strengthen Wikipedia content — especially articles on gender and other marginalized topics. By combining the quantitative insight from process mining and process discovery models with the qualitative knowledge of editor practices, the new research project targets the root causes — neglect, bias, underrepresentation, and deletion — with the focus from measurement to preservation and sustained resilience of gender and marginalized-topic articles.

Methods

This study employs a multifaceted mixed-methods approach, integrating quantitative data analytics, process mining, and qualitative community-driven approaches. The methodology is structured into three interconnected phases.

Phase 1: Collection of Data about Endangered Articles

To systematically identify, gather, and prepare datasets for analyzing Wikipedia articles at risk, the study collects a diverse set of data from edit history data, viewership statistics, deletion logs, and article history from English Wikipedia articles with a focus on gender and marginalized community topics.

Tasks 1.1. Data Collection. The study will first access from Wikimedia APIs to track and collect edit history data, monthly and daily views, and deletion logs, including editor interactions, timestamps, and types of edits, covering a longitudinal timeframe. We plan to deploy and collect data from English Wikipedia articles on gender and marginalized community topics. In addition, the study will access Wikipedia's deletion archives, tracking discussions, nominated articles, reasons cited for deletion, and outcomes to identify common themes and outcomes influencing article survival.

Tasks 1.2. **Data Integration and Data Quality.** During data integration and data quality management, the study will consolidate data

into structured formats and detect and address data quality issues such as missing values, duplicates, and incompleteness issues using Python data libraries (such as Pandas and Numpy).

Phase 2: Identification of Endangered Articles
Phase 2 focuses on accurately identifying Wikipedia articles at risk through two analytical methods: (1) Process Mining and Discovery to analyze editing and deletion workflows, identifying patterns linked to article neglect or deletion; and (2) Predictive Analytics, leveraging insights from process mining to develop machine learning models through a public API for early detection of endangered articles, especially those related to gender and marginalized communities.

Tasks 2.1. Process Mining and Process Discovery. To preserve endangered articles effectively, the first step is identifying them accurately and efficiently. The research will first utilize process mining to analyze workflows and patterns from viewing, editing, and deletion data, enabling deeper insights into behavioral patterns that result in articles at risk of endangerment. Articles can be considered endangered and at risk of deletion due to several factors, including low engagement and attention, frequent nomination for deletion, bias, and underrepresentation. Hence, the research will evaluate performance metrics such as article survival duration, frequency of edits and views, and response times from the community. This analysis approach can help to identify critical activities or decision points correlated with article resilience or vulnerability. Through the analysis, we would like to reveal how view and edit patterns correlate with editorial actions and public interest over time. In addition, the research will apply process discovery algorithms (e.g., Alpha Miner, Inductive Miner, or Heuristic Miner using ProM and Python-based libraries like PM4Py) to generate process models representing article editing, viewing, and deletion workflows and patterns. The analysis will visualize frequent paths revealing article improvement, stagnation, or deletion processes. Moreover, the findings can identify recurring patterns or bottlenecks associated with endangered content (e.g., prolonged periods without edits or views, escalating controversies, or rapid nomination-to-deletion transitions). Based on the discovered patterns, the research will engage with the Wikipedia editor's community to analyze deviations between discovered process models and idealized editorial workflows (e.g., standard best practices documented in Wikipedia guidelines). This process will gen

Tasks 2.2. Predictive Analytics and Identification of Endangered Articles.
Leveraging insights derived from process mining and process discovery, this research aims to develop robust predictive analytics and an associated identification algorithm, which will be made accessible through a public open-source API and dataset. The predictive analytics will incorporate features identified through detailed process mining analysis, such as editorial activity frequency, edit sequences, and timing, article view counts, shifts in viewership patterns, and common patterns in deletion processes. Utilizing these process-driven features, the project will implement statistical methods and machine learning models to assess article risk levels. Risk assessment criteria will include consistently low views, sudden drops in readership, frequent nomination for deletion, and outcomes from prior community discussions. The models will undergo rigorous validation through cross-validation procedures and comparisons with historical deletion records to measure their accuracy in early detection of endangered articles. Additionally, model predictions will be systematically tested against a manually curated dataset of previously

deleted or endangered articles to further refine predictive accuracy and reliability. The publicly accessible dataset contains article-level risk scores, completeness metrics, and recommended improvement actions. The open-source API will promote community collaboration and ensure the preservation tooling evolves with community needs.

Phase 3: Editor's Community-Driven Engagement Strategies

This phase focuses on translating data-driven insights into practical, community-based strategies to safeguard at-risk Wikipedia articles, particularly those concerning gender and marginalized topics. In collaboration with the Wikipedia editors' community, this phase will adopt a co-design approach to gain insights from the editors' community to ensure that the proposed strategies to enhance the resilience and quality of endangered articles are technically feasible, acceptable, and effectively integrated into existing editing practices.

Tasks 3.1. Actionable Insights: Delivering Endangered Article Lists and Recommendations. Building on the process mining insights from Phase 2, this phase will generate curated lists of at-risk articles along with tailored recommendations for improvement aimed at enhancing article resilience. These recommendations will include specific actions such as adding reliable sources, expanding content, improving structure, and addressing common deletion concerns identified through workflow analysis.

In collaboration with the editor community, the project will explore and co-design the most effective ways to present and integrate these endangered article lists into existing editorial workflows. This will include usability testing and feedback sessions to determine how the information can be best surfaced to support timely and meaningful preservation efforts. This approach can ensure that focused attention is given to topics related to gender and marginalized communities, ensuring that preservation strategies align with the values and priorities of editors working in these knowledge domains.

Tasks 3.2. Co-Design of Notification System for Article Risk Awareness. We will conduct co-design interviews, focus groups, and feedback sessions with Wikipedia editors to explore how the public API (developed in Phase 2) can best support their editing workflows. Through the insights from the editors' community, we would like to propose community-informed guidelines and best practices for using the API to prioritize editing tasks, identify endangered articles in real-time, and guide community-led preservation efforts. In addition, collaborating actively with the Wikipedia editor community, this research will explore and develop a notification system that automatically alerts editors when articles they have contributed to or follow are flagged as at risk, based on predictive analytics and past interaction data discovered and developed in Phase 2. Notifications could be integrated into user talk pages, and watchlists, as per editor preferences gathered during the co-design phase.

Expected Output

We envision the following outputs:

- Two scientific publications (e.g., at the International AAAI Conference On Web And Social Media, ICWSM, WWW, CHI; and a journal publication, such as the Academy of Management Journal, Journal of Strategic Information Systems, International Journal of Human-Computer Studies). We will present and discuss the novel methods and an experimental evaluation of the effectiveness co-authored by the principal investigators and the research assistant at the research conferences and Wikimania. The primary intended audience for

this output is the academic research community and Wikimedia community.

● Open Dataset. A publicly accessible dataset will be provided containing the lists of articles at risk, with metrics of viewing, editing and completeness, and recommended improvement actions. This dataset will enable researchers, tool builders, and the wider Wikimedia community to track progress, replicate analyses, and innovate new preservation strategies.

● Public API. An open-source public API will be developed to enable technical Wikimedia contributors and Wikipedia community members to systematically assess and monitor the completeness and risk status of Wikipedia articles over time. This API will provide predictive insights generated from the research, empowering users to proactively identify and address endangered articles, particularly those related to gender and marginalized communities. The open-source approach encourages community pull requests and peer review of algorithms, fostering trust, inviting collaboration, and ensuring the preservation evolves with community needs in long-term availability.

● Insights to inform decision-making by the Wikipedia user community on which parts of the project to focus on. The primary intended audience for this output is the community of editors who would be empowered to make data-driven editorial decisions to preserve endangered articles.

Risks

We consider the possible project risks including skilled staff availability and low community engagement.

Research Staff: We plan to recruit project staff (i.e, a Research Officer/Senior Research Assistant) majoring in Data Science and Computer Science at the University of Queensland (UQ). This will allow us to tap into a large pool of candidates (approx. 100 graduates/year) with the right skills for the project.

Community Engagement: We budgeted for a community engagement role to dedicate the necessary time to seek and listen to any community feedback and to incorporate it on board of our solution design to increase the likelihood of future acceptability and adoption.

	M1-2	M3-4	M5-6	M7-8	M9-10	M11-12
T1.1						
T1.2						
T2.1						
T2.2						
T3.1						
T3.2						

Community impact plan

We will seek to impact audiences beyond researchers and academics with the help of the community engagement role envisioned in the project. We will engage the editor community from our previous project while actively forging new partnerships with a broad spectrum of Wikipedia user groups (e.g., https://meta.wikimedia.org/wiki/WikiWomen%27s_User_Group), which focuses on the gender gap and marginalized topics to seek their feedback on our proposed design and make the expected output more adoptable and impactful. To this end, the community engagement staff member will start at the beginning of the project (i.e., months 1-2, see also Gantt chart above) to seek early feedback on the overall project idea and solution design. Then, once the solution development has

reached a good maturity level, this staff member will engage again with the user groups to understand the potential for adoption and seek to collect any further additional requirements that could increase it.

Evaluation

We will consider the following research success factors and metrics from technical validity, content impact and community engagement, tool adoption and long-term sustainability.

*Technical validity***.** We will train and cross-validate risk-prediction models on historical deletion logs, edit, and view data. We will also track the number of days between a high-risk flag and a formal deletion nomination. These metrics ensure we can reliably identify endangered articles early enough for editors to act.

Content-level impact and community engagement. Through the collaboration with the editor's community, we will evaluate whether editors use and benefit from the proposed approaches to preserving endangered articles. We will track notification open-rates and the proportion of alerts that lead to edits.

*Proposed strategies for adoption and sustainability***.** we monitor API analytics, dataset downloads and scholarly citations, and the volume of external pull-requests or issue reports on our open-source repository. We will collaborate with the editors' community in co-design and focus group methods to evaluate and foster the transparent, community-led, and sustainability of the proposed strategies.

Budget

We budgeted for a total cost of approximately $50,000 USD. This covers a Research Officer/ Senior Research Assistant to work on the research tasks and community engagement. We will seek to fill the role from recent UQ graduates who have skills in data science and software development.

More importantly, we would like to continue the collaboration with the editor's community and would like to invite a community relationship manager who will be responsible for the community engagement aspects of the project and seek community feedback to make sure the project outcome can reach the desired impact.

In addition, we budgeted for the costs to publish journal and attend a scientific conference to present the research conducted in the project (e.g., by means of an accepted paper presentation). Please refer to the linked spreadsheet for a full cost breakdown (https://docs.google.com/spreadsheets/d/1Du6-AfnNy0yz9_mLc2hqdJt-z8awPmm-oyjdZMk0KTs/edit?usp=sharing)

References

Asthana, S., & Halfaker, A. (2018). With Few Eyes, All Hoaxes are Deep. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW), 21:1-21:18. https://doi.org/10.1145/3274290

Chen, T., Demartini, G., Indulska, M., & Sadiq, S. (2023). Exploring Data Workers' Behaviours in Data Quality Discovery. *ACIS 2023 Proceedings*. https://aisel.aisnet.org/acis2023/99

Chen, T., Sadiq, S., & Indulska, M. (2020). Sensemaking in Dual Artefact Tasks – The Case of Business Process Models and Business Rules. In G. Dobbie, U. Frank, G. Kappel, S. W. Liddle, & H. C. Mayr (Eds.), *Conceptual Modeling* (pp. 105–118). Springer International Publishing. https://doi.org/10.1007/978-3-030-62522-1_8

Demartini, G. (2023). *Research:Measuring the Gender Gap: Attribute-based Class Completeness Estimation—Meta*. Wikimedia Research. https://meta.wikimedia.org/wiki/Research:Measuring_the_Gender_Gap:_Attribute-based_Class_Completeness_Estimation

Lam, S. (Tony) K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., & Riedl, J. (2011). WP:clubhouse? An exploration of Wikipedia's gender imbalance. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 1–10. https://doi.org/10.1145/2038558.2038560

Patel, H., Chen, T., Bongiovanni, I., & Demartini, G. (2024). Estimating Gender Completeness in Wikipedia. *ACIS 2024 Proceedings*. https://aisel.aisnet.org/acis2024/99

Schneider, J., Passant, A., & Decker, S. (2012). Deletion discussions in Wikipedia: Decision factors and outcomes. *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, 1–10. https://doi.org/10.1145/2462932.2462955

Taraborelli, D., & Ciampaglia, G. L. (2010). Beyond Notability. Collective Deliberation on Content Inclusion in Wikipedia. *2010 Fourth IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop*, 122–125. https://doi.org/10.1109/SASOW.2010.26

Tripodi, F. (2023). Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, *25*(7), 1687–1707. https://doi.org/10.1177/14614448211023772

van der Aalst, W. (2016). Data Science in Action. In W. van der Aalst (Ed.), *Process Mining: Data Science in Action* (pp. 3–23). Springer. https://doi.org/10.1007/978-3-662-49851-4_1

van der Aalst, W. M. P., Reijers, H. A., Weijters, A. J. M. M., van Dongen, B. F., Alves de Medeiros, A. K., Song, M., & Verbeek, H. M. W. (2007). Business process mining: An industrial application. *Information Systems*, *32*(5), 713–732. https://doi.org/10.1016/j.is.2006.05.003

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(1), Article 1. https://doi.org/10.1609/icwsm.v9i1.14628

Warncke-Wang, M., Cosley, D., & Riedl, J. (2013). Tell me more: An actionable quality model for Wikipedia. *Proceedings of the 9th International Symposium on Open Collaboration*, 1–10. https://doi.org/10.1145/2491055.2491063

Worku, Z., Bipat, T., McDonald, D. W., & Zachry, M. (2020). Exploring Systematic Bias through Article Deletions on Wikipedia from a Behavioral Perspective. *Proceedings of the 16th International Symposium on Open Collaboration*, 1–22. https://doi.org/10.1145/3412569.3412573

Yunus, Y., Chen, T., & Demartini, G. (2025). *Exploring Wikipedia Gender Diversity Over Time—The Wikipedia Gender Dashboard (WGD)*. Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25). https://doi.org/10.48550/arXiv.2501.12610