Analyse des Contributions Genrées sur Wikipédia : Thématiques et Prédictions Futures

Fatoumata DIALLO
Ismael SOW
Tuteur : Marc SPANIOL

Abstract

Cette étude analyse les contributions sur Wikipédia en fonction du genre (masculin, féminin, transgenre) afin d'identifier les thématiques privilégiées et d'anticiper les tendances émergentes. En exploitant des données publiques et en appliquant des modèles d'analyse temporelle et prédictive, nous examinons les disparités de participation et les biais thématiques. Notre approche intègre également des dimensions géographiques et linguistiques afin d'offrir une vision plus complète de ces dynamiques. Les résultats mettent en évidence des déséquilibres persistants et ouvrent des perspectives pour renforcer l'équité et la diversité sur les plateformes collaboratives.

1 Introduction

Wikipédia, en tant qu'encyclopédie collaborative, repose sur les contributions volontaires de milliers d'éditeurs à travers le monde. Cependant, bien qu'elle aspire à représenter l'ensemble des savoirs humains, des déséquilibres notables persistent dans la participation des contributeurs. En particulier, la sous-représentation des femmes et des personnes transgenres parmi les éditeurs actifs influence la diversité et l'accessibilité des connaissances disponibles.

Les études précédentes ont mis en évidence une sur-représentation masculine qui impacte la couverture et la profondeur des sujets abordés. Par ailleurs, ces disparités varient selon les régions et les langues, ce qui complexifie l'analyse des contributions genrées à l'échelle mondiale. Ces biais posent des enjeux majeurs en matière d'objectivité et d'inclusivité des plateformes collaboratives.

Cette étude vise à analyser la répartition des contributions selon le genre en identifiant les thématiques privilégiées et en étudiant leur évolution dans le temps et à travers différentes régions. En intégrant des modèles prédictifs, nous cherchons également à anticiper les tendances émergentes et à proposer des recommandations pour une meilleure représentativité des contenus sur Wikipédia. L'objectif est de mieux comprendre ces dynamiques afin de favoriser une participation plus équilibrée et inclusive sur la plateforme.

2 Problématique et hypothèse de recherche

Wikipédia joue un rôle central dans la diffusion des connaissances à l'échelle mondiale. Toutefois, la plateforme repose sur des contributions bénévoles, ce qui peut entraîner des déséquilibres dans la représentativité des contenus. Parmi ces déséquilibres, les disparités genrées ont été largement étudiées et soulèvent des interrogations quant à l'objectivité et la diversité des savoirs partagés.

2.1 Problématique

Les recherches antérieures ont montré que les contributeurs masculins sont majoritaires sur Wikipédia, influençant ainsi la couverture des thématiques abordées. Cette prédominance entraîne un déséquilibre dans la représentation des sujets, avec une couverture plus détaillée des domaines historiquement associés aux hommes, comme la technologie, l'histoire militaire et les sciences. À l'inverse, les sujets liés aux femmes et aux minorités de genre sont souvent sous-représentés, tant en quantité qu'en profondeur.

Cependant, cette disparité ne se manifeste pas de manière uniforme à travers les différentes éditions linguistiques de Wikipédia. Certaines langues ou régions présentent une plus grande diversité de contributions, tandis que d'autres affichent des biais plus marqués. Par ailleurs, les évolutions sociétales et les changements culturels influencent progressivement les contributions, rendant l'analyse temporelle essentielle pour mieux com-

prendre ces dynamiques. Dans ce contexte, il est crucial d'examiner les disparités genrées sous les dimensions de la distribution des contributions selon le genre des éditeurs, l'évolution de ces tendances sur une période donnée, les variations régionales et linguistiques influençant ces disparités.

2.2 Hypothèse de recherche

À partir de ces observations, nous formulons plusieurs hypothèses. Les préférences thématiques varient significativement selon le genre des contributeurs (masculin, féminin, transgenre), avec des tendances récurrentes dans les catégories éditées. De plus, ces préférences évoluent dans le temps en réponse à des facteurs externes tels que les avancées technologiques, les mouvements sociaux et les débats culturels.

Les biais genrés observés dans certaines éditions linguistiques de Wikipédia semblent être influencés par des facteurs régionaux et culturels. Enfin, l'émergence de nouvelles thématiques pourrait être anticipée grâce à une analyse prédictive des tendances actuelles des contributions.

Ces hypothèses guideront notre étude et permettront de mieux comprendre comment les dynamiques genrées façonnent la production et la diffusion des connaissances sur Wikipédia.

3 État de l'Art

3.1 Ce que disent les recherches

Hargittai et Shaw (2015) analysent les freins à la participation des femmes sur Wikipédia. Leur étude souligne que l'environnement collaboratif peut être perçu comme hostile, ce qui limite l'engagement des contributrices. Le manque de reconnaissance, les conflits d'édition et la nécessité de maîtriser des compétences techniques spécifiques constituent des obstacles majeurs. En conséquence, de nombreuses contributrices abandonnent rapidement la plateforme, renforçant ainsi un déséquilibre de genre persistant. L'étude met en avant la nécessité d'initiatives spécifiques pour améliorer l'intégration et la rétention des éditrices sur Wikipédia.

Hinnosaar (2019) démontre que, bien que les contributions féminines soient souvent plus détaillées et précises, elles restent largement minoritaires. L'auteur met en évidence une asymétrie structurelle dans la production des contenus, où les thématiques liées aux femmes sont moins documentées et moins visibles que celles dominées par des contributeurs masculins. Cette sous-représentation affecte directement la diversité éditoriale et la qualité de l'information disponible. L'étude suggère que l'absence de contributrices en nombre suffisant nuit à l'équilibre des connaissances sur Wikipédia, qui demeure marqué par un prisme masculin.

Arora et al. (2023) révèlent que les biographies de femmes sur Wikipédia sont significativement plus orphelines que celles des hommes. Ces articles possèdent moins de liens internes les connectant à d'autres pages, ce qui réduit leur accessibilité et leur visibilité pour les lecteurs. Cette faible interconnexion nuit à leur référencement et les rend plus vulnérables aux suppressions. En conséquence, les figures féminines bénéficient d'une reconnaissance moindre dans l'encyclopédie numérique, ce qui perpétue leur marginalisation dans la construction du savoir en ligne. L'étude met en avant un biais structurel qui pénalise systématiquement la visibilité des biographies féminines.

Profesional de la Información (2024) examine l'évolution des biais de genre sur Wikipédia malgré les efforts visant à promouvoir la diversité des contributeurs. L'étude démontre que les articles rédigés par des femmes sont sous-représentés et bénéficient d'une moindre exposition dans les algorithmes de classement. Cette distribution inégale de visibilité influence l'accès au savoir et renforce les inégalités dans la production des connaissances numériques. Même lorsque les femmes participent activement à l'édition, leurs contributions sont moins mises en avant, ce qui limite leur impact éditorial. L'étude suggère que la correction de ces biais nécessite des ajustements non seulement dans les pratiques éditoriales, mais aussi dans les systèmes algorithmiques de la plateforme.

3.2 Conclusion et perspectives

L'ensemble des travaux existants convergent vers un même constat : Wikipédia reste marqué par des biais structurels qui influencent à la fois la participation des contributeurs et la diversité des contenus. La sous-représentation des contributrices, la moindre visibilité des articles qu'elles produisent et l'influence des algorithmes sur la hiérarchisation des connaissances sont autant de facteurs qui perpétuent ces déséquilibres.

Toutefois, ces recherches se concentrent princi-

palement sur l'analyse des contributions passées et présentes. Elles ne cherchent pas à anticiper l'évolution des contributions genrées sur Wikipédia dans les années à venir. C'est dans cette optique que notre étude se distingue : en utilisant des modèles de régression et d'analyse temporelle, nous proposons une projection sur les cinq prochaines années (2025-2030).
Cette approche permet d'anticiper les tendances d'édition et d'évaluer si les évolutions actuelles vont se poursuivre, s'accélérer ou ralentir. En intégrant cette dimension prédictive, notre travail apporte un nouvel éclairage sur les dynamiques d'édition et pourrait aider à mieux orienter les stratégies visant à favoriser une participation plus équilibrée sur Wikipédia.

4 Méthodologie

Elle se déroule en trois étapes : la **collecte** des données,le **prétraitement** et l'**analyse** (incluant un module de prédiction).

4.1 Collecte des données

Les données ont été récupérées à partir des catégories de contributeurs suivantes sur Wikipédia : Female Wikipedians, Male Wikipedians, Transgender Wikipedians.
Dans un premier temps, nous avons tenté d'extraire ces informations via du **web scraping** en récupérant les données directement à partir du HTML des pages utilisateur. L'objectif était de collecter les profils des contributeurs ainsi que leurs historiques de modifications. Cependant, cette approche s'est rapidement heurtée à plusieurs limitations. La structure du HTML de Wikipédia est dynamique et sujette à des modifications fréquentes, ce qui complique l'extraction automatisée des données. De plus, certaines informations cruciales, comme la localisation des contributeurs et leurs historiques détaillés d'édition, ne sont pas directement accessibles dans le code source des pages, rendant la récupération partielle et incomplète.
Face à ces contraintes, nous avons opté pour une approche plus efficace et fiable en utilisant l'**API Wikipédia** (*wikipediaapi*). Cette API permet d'accéder directement aux métadonnées des contributeurs et de récupérer des informations structurées, garantissant ainsi une extraction plus exhaustive et reproductible. Grâce à cette méthode, nous avons pu collecter les identifiants des contributeurs, les articles édités, les horodatages des modifications, la langue d'édition et la localisation lorsque celle-ci était disponible.
L'utilisation de l'API a offert plusieurs avantages par rapport au web scraping : elle permet d'obtenir des données mises à jour en temps réel, d'accéder à des informations non visibles dans l'interface utilisateur standard de Wikipédia et d'éviter les contraintes liées aux restrictions d'accès ou aux évolutions du code HTML de la plateforme. Cette approche garantit ainsi une meilleure fiabilité des données tout en facilitant leur traitement et leur analyse dans le cadre de notre étude sur les contributions genrées.

Genre	Nb. Contributeurs	Nb. Contributions
Masculin	19 000	1 097464
Féminin	7 700	582 038
Transgenre	1650	27 3405

Table 1: Volumes de données collectées

4.2 Prétraitement des données

Une fois les données collectées, un travail de prétraitement a été réalisé afin d'assurer leur qualité et leur cohérence. La localisation des contributeurs, bien que parfois renseignée explicitement, n'était pas toujours disponible ou normalisée. Afin d'améliorer la couverture géographique des éditeurs, nous avons adopté une approche mixte combinant plusieurs sources d'informations.
Dans un premier temps, lorsque la localisation était fournie, elle a été normalisée à l'aide de *pycountry* et *GeoNames* afin d'uniformiser les noms de pays et de villes. Pour les cas où aucune localisation explicite n'était mentionnée, une estimation a été effectuée en se basant sur la langue des éditions et le contenu des articles modifiés. Par exemple, un contributeur nommé **Jean Dupont**, éditant principalement des articles en français et ayant apporté plusieurs modifications à la page *"Histoire de la Révolution française"*, présentait une forte probabilité d'être basé en France. Dans ce cas, une correspondance entre la langue d'édition et le contenu modifié a permis d'attribuer une localisation probable en France.
Cette approche a permis d'assigner une localisation plausible à de nombreux contributeurs sans information directe, tout en limitant les biais d'interprétation. Toutefois, lorsque les indices étaient insuffisants ou contradictoires (par exem-

ple, un contributeur éditant en plusieurs langues sans préférence claire), la localisation n'a pas été attribuée afin d'éviter toute erreur de classification.

4.3 Filtrage des genres

Pour garantir une analyse ciblée, seules les contributions associées aux trois catégories étudiées (*Male*, *Female*, *Transgender*) ont été conservées. Les entrées sans indication explicite du genre n'ont pas été utilisées afin de garantir une analyse plus fiable.

Une vérification supplémentaire a été effectuée sur les dates d'édition afin de s'assurer de leur cohérence temporelle. Les valeurs aberrantes, comme des dates antérieures à la création de Wikipédia ou des horodatages incohérents, ont été corrigées lorsque possible en s'appuyant sur l'historique des modifications. Lorsque les anomalies ne permettaient pas une correction fiable, les entrées concernées ont été exclues afin d'assurer la fiabilité des analyses temporelles sur l'évolution des contributions par genre.

4.4 Enrichissement des données

L'enrichissement des données a été essentiel pour structurer notre base et affiner l'analyse des contributions sur Wikipédia. Nous avons ajouté plusieurs variables permettant d'étudier les tendances éditoriales, en regroupant les contributions par période, en catégorisant les articles et en comptabilisant le nombre de modifications.

La variable **YearMonth** a été introduite pour regrouper les contributions par mois et année, facilitant ainsi l'analyse temporelle et l'identification des variations saisonnières. Pour enrichir la description des articles, nous avons extrait des résumés depuis **DBPedia**, une base de connaissances dérivée de Wikipédia. Après récupération des résumés via API, un nettoyage a été effectué pour harmoniser les formats et éliminer les entrées non valides.

L'identification des thématiques a été réalisée à l'aide d'une **vectorisation TF-IDF**, permettant d'extraire les termes représentatifs des titres et de les regrouper en catégories telles que la technologie, la santé ou la culture. La vectorisation TF-IDF a été retenue pour identifier les termes les plus représentatifs des articles, tandis que le **clustering K-Means** a permis de regrouper les contenus édités en fonction de leur similarité sémantique, facilitant ainsi l'analyse des préférences éditoriales par genre.

L'analyse des fréquences d'édition a été intégrée via la variable **Counts**, représentant le nombre total de modifications réalisées sur un article par l'ensemble des contributeurs, offrant un indicateur de l'activité éditoriale. Afin d'éviter les analyses sur des échantillons trop faibles, seuls les articles ayant été modifiés au moins cinq fois ont été pris en compte.

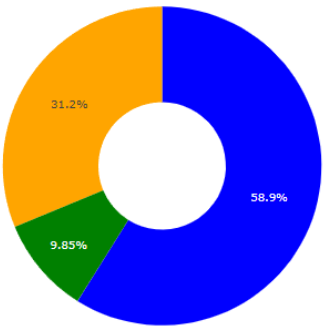
Grâce à ces enrichissements, notre base de données a été optimisée pour une analyse plus fine des dynamiques d'édition sur Wikipédia. Ces transformations ont constitué une étape clé pour les analyses ultérieures, notamment le clustering des articles et la simulation des tendances futures.

5 Résultats et Analyse

Dans cette section, nous présentons les principaux résultats de notre étude, basés sur la répartition des contributions par genre, l'évolution temporelle, la distribution spatiale et la répartition linguistique.

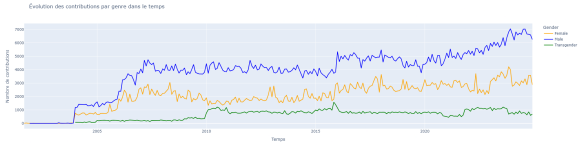
5.1 Répartition par genre :

L'analyse des contributions révèle des différences significatives. Les hommes représentent 58,9% des contributeurs, tandis que les femmes comptent pour 31,2%. Enfin, les contributeurs transgenres constituent 9,85% du total



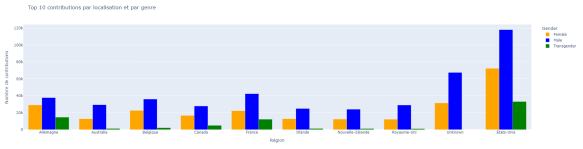
5.2 Évolution temporelle :

Les contributions féminines ont progressé successivement au fil du temps, celles des transgenres ont connu une augmentation constante. Mais la majorité des contributions reste masculine.



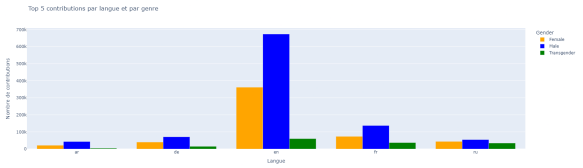
5.3 Distribution spatiale :

Les contributions masculines sont principalement concentrées en **Europe et aux États-Unis**. La participation féminine, quant à elle, est plus marquée en **Asie et aux États-Unis**. Les contributions des personnes transgenres montrent une **distribution plus équilibrée** entre les différentes régions.



5.4 Répartition linguistique :

L'édition anglophone **domine globalement** avec une répartition genrée proche des tendances générales. Les langues (*français, espagnol, italien*) présentent un **équilibre** entre les genres, et les langues asiatiques montrent une **forte prédominance masculine**.



6 Modélisation et Simulation des Prédictions

6.1 Utilisation des modèles de prédiction

Dans un premier temps, nous avons testé plusieurs modèles d'apprentissage automatique pour prédire le nombre de contibutions genrés dans les années à venir. Parmi eux, nous avons exploré des modèles de régression linéaire, ainsi que des algorithmes de forêts aléatoires (*Random Forest*) et de boosting (*XGBoost*).

Après plusieurs expérimentations, nous avons comparé ces modèles en fonction de leur capacité à prédire les tendances des éditions d'articles selon le genre des contributeurs. Pour évaluer leur performance, nous avons utilisé le coefficient de détermination (R^2), qui mesure la proportion de variance expliquée par le modèle.

Les résultats obtenus sont résumés dans le tableau ci-dessous :

Les résultats montrent que la régression linéaire offre les meilleures performances, avec des scores R^2 les plus élevés pour l'ensemble des catégories de genre. XGBoost présente des performances légérement inférieures, mais reste une alternative viable pour des tendances plus complexes. Enfin, le modèle Random Forest est le moins performant dans ce cas spécifique.

Modèle	Scores R^2	Adapté pour ?
Régression Linéaire	0.91 (Male), 0.89 (Female), 0.93 (Trans)	Tendance linéaire
XGBoost	0.96 (Male), 0.97 (Female), 0.91 (Trans)	Données complexes
RandomForest	0.95 (Male), 0.95 (Female), 0.90 (Trans)	Données bruitées

Table 2: Comparaison des performances des modèles de prédiction

6.2 Modélisation et visualisation des tendances

À l'aide des modèles retenus, nous avons prédit l'évolution du nombre d'éditions pour la période **2025-2030**. En utilisant les tendances historiques des contributions et en appliquant des algorithmes de régression, nous avons généré des projections permettant d'anticiper les évolutions futures. Une visualisation a été réalisée pour illustrer ces résultats, mettant en évidence la répartition des éditions par genre et par année.

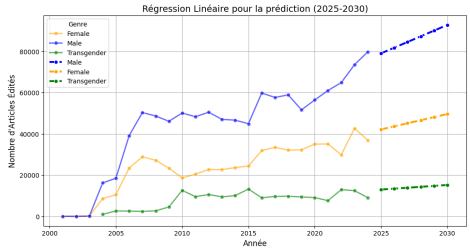


Figure 1: Prédiction du nombre d'articles édités sur Wikipédia (2025-2030) selon le genre des contributeurs

Le graphique ci-dessus illustre la **prédiction du nombre d'articles édités sur Wikipédia entre 2025 et 2030**, en fonction du **genre des contributeurs**. Chaque courbe représente une tendance prédite, avec une couleur associée à chaque groupe (bleu pour les hommes, orange pour les femmes, et vert pour les contributeurs transgenres). Les lignes pleines montrent les données historiques, tandis que les lignes pointillées indiquent les prédictions des modèles.

On observe que certaines tendances d'édition varient selon les profils de contributeurs. Par exemple, les contributions des femmes montrent une croissance plus stable, tandis que celles des hommes affichent une forte augmentation sur certaines périodes. Les contributions des personnes transgenres restent plus modestes en volume, mais

avec une progression continue.

6.3 Discussion des résultats

Les résultats obtenus mettent en évidence des évolutions dans les contributions selon le genre, reflétant des dynamiques d'édition distinctes. L'utilisation de modèles comme la régression linéaire et XGBoost a permis d'obtenir des prédictions cohérentes avec les tendances historiques, bien que certaines limites persistent.

L'analyse comparative des modèles (Table 2) montre que la régression linéaire offre les prédictions les plus robustes, tandis que les modèles non linéaires comme XGBoost et Random Forest s'avèrent moins adaptés aux tendances globales des éditions. Ces résultats suggèrent que la production de contenu sur Wikipédia suit une évolution relativement linéaire, influencée par des facteurs structurels et sociétaux.

Cette étude apporte ainsi un nouvel éclairage sur la dynamique des contributions et ouvre la voie à de futures analyses plus approfondies. En particulier, une meilleure compréhension des biais éditoriaux et des dynamiques communautaires pourrait permettre d'adapter les stratégies d'inclusion et d'équilibre dans la production des savoirs.

7 Conclusion et Perspectives

Les résultats de cette étude ouvrent des perspectives pour renforcer l'équité et la diversité des contributions sur Wikipédia. Une première approche consiste à encourager la participation des femmes et des personnes transgenres dans des thématiques historiquement dominées par des contributeurs masculins, afin d'élargir la représentation des savoirs. Il est également essentiel de promouvoir des initiatives de sensibilisation et de formation pour réduire les barrières à la participation et favoriser un environnement plus inclusif. Enfin, l'intégration d'outils de visualisation et d'analyse permettrait de mieux identifier les biais existants et de suivre l'évolution des contributions dans le temps, offrant ainsi des indicateurs clés pour orienter les efforts vers une plus grande diversité éditoriale.

Bien que cette étude ait permis de combler certaines lacunes dans la littérature, plusieurs axes de recherche méritent d'être approfondis. Une première piste consisterait à étendre l'analyse aux dimensions intersectionnelles, en intégrant des facteurs tels que l'âge, l'origine ethnique ou le niveau socio-économique des contributeurs. L'exploration de ces variables permettrait d'obtenir une vision plus fine des dynamiques d'édition et de mieux comprendre les inégalités structurelles sur la plateforme.

Le développement de modèles prédictifs plus avancés représente également une voie prometteuse. L'intégration de méthodes d'apprentissage profond pourrait améliorer la précision des prévisions et offrir une analyse plus robuste des tendances émergentes. Une approche combinant intelligence artificielle et fouille de texte pourrait notamment permettre d'identifier plus efficacement les évolutions des contributions et les biais persistants. Enfin, il serait pertinent d'étudier l'impact des initiatives communautaires visant à réduire les disparités genrées et à favoriser une meilleure représentativité des contenus. L'évaluation des effets des programmes de sensibilisation et des campagnes de recrutement de nouveaux contributeurs permettrait d'identifier les stratégies les plus efficaces pour rendre Wikipédia plus inclusive.

References

[1] Lam, S. T., et al. WP:Clubhouse? An Exploration of Wikipedia's Gender Imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 2011. https://dl.acm.org/doi/10.1145/2038558.2038560

[2] Hargittai, E., Shaw, A. Mind the Skills Gap: The Role of Internet Know-How and Gender in Differentiated Contributions to Wikipedia. *Information, Communication Society*, 18(4), 424-442, 2015. https://www.tandfonline.com/doi/abs/10.1080/1369118X.2014.957711

[3] Wagner, C., Garcia, D., Jadidi, M., Strohmaier, M. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, 2015. https://arxiv.org/abs/1501.06307

[4] Hecht, B., Gergle, D. The Tower of Babel Meets Web 2.0: User-Generated Content and its Applications in a Multilingual Context. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI)*, 2010. https://dl.acm.org/doi/10.1145/1753326.1753370

[5] Menking, A., Erickson, I. The Heart Work of Wikipedia: Gendered, Emotional Labor in the World's Largest Online Encyclopedia. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work Social Computing*,

2015. https://dl.acm.org/doi/10.1145/2702123.2702514

[6] LADN. Pourquoi Wikipédia est un enfer pour les femmes. 2024. https://www.ladn.eu/tech-a-suivre/pourquoi-wikipedia-est-un-enfer-pour-les-femmes/

[7] Slate. Pourquoi Wikipédia et les femmes ne font pas bon ménage. 2010. https://www.slate.fr/story/34181/wikipedia-femmes-pas-bon-menage

[8] Next. Sur Wikipédia, 20% des biographies concernent désormais des femmes. 2024. https://next.ink/147182/sur-wikipedia-20-des-biographies-concernent-desormais-des-femmes/