Production de data et gouvernance de la qualité dans Wikipédia

Marie-Noëlle Doutreix

Marie-Noëlle Doutreix, « Production de data et gouvernance de la qualité dans Wikipédia ». *H2PTM'19 De l'hypertexte aux humanités numériques*, Ioan Roxin et *Al*., Oct 2019, Montbéliard, France.

Production de data et gouvernance de la qualité dans Wikipédia

RÉSUMÉ. La communauté wikipédienne a créé et mobilise de nombreux outils pour quantifier les activités de ses contributeurs et décrire de manière chiffrée l'évolution de ses articles, créant ainsi une masse de données pouvant servir tant au contributeur qu'au chercheur. Parallèlement, des principes et procédés encadrent les contributions de manière à tenter d'homogénéiser la qualité des articles. Cette recherche voudrait questionner le rôle des data dans le projet collaboratif et éditorial Wikipédia. Il s'agit de comprendre ce que la production massive de données à propos des contributions est censée faire – et fait effectivement. Dans quelle mesure ces données massives servent-elles au pilotage de Wikipédia ? Pour cela, nous nous intéresserons à l'outil d'analyse de données XTools dont la fonction est résumée dans la formule « Feeding your data hunger ».

ABTRACT. The Wikipedian community has created and mobilizes numerous tools to quantify the activities of its contributors and to describe in a numerical way the evolution of its articles. At the same time, principles and procedures frame the contributions in order to try to homogenize the quality of the articles. This research work on determine the role of data in the collaborative and editorial project Wikipedia. It's about understanding what massive data production about contributions is supposed to do - and actually does. Are these massive data used to manage Wikipedia? For this, we will focus on the XTools data analysis tool whose function is summarized in the formula « Feeding your data hunger ».

MOTS-CLES : Données – Management – Qualité – Quantification – Wikipédia – Gouvernance

KEYWORDS: Data – Management – Quality – Quantification – Wikipedia – Governance

1. Introduction

La Fondation Wikimedia, dans une logique d'*open data*, met à la disposition de tous une large base de données concernant les différents wikis qu'elle gère. Certains contributeurs ont ainsi créé, parfois sous l'impulsion de la Fondation, parfois de leur propre initiative, des outils d'analyse de données concernant de nombreux aspects de l'activité éditoriale de Wikipédia. Ces outils puisent directement dans la base de données de la Fondation Wikimedia pour rendre accessibles de très nombreuses informations. L'interface web du plus central de ces outils, XTools, est utilisée plusieurs milliers de fois par jour. Néanmoins, il apparaît que les usages faits à partir de ces données ne sont pas bien identifiés par les concepteurs de l'outil. Quels rôles jouent les données produites dans le fonctionnement de Wikipédia ? Nous nous interrogerons sur trois usages, deux internes à l'encyclopédie – la lutte contre le vandalisme et le management des contributeurs – et un autre externe – les recherches effectuées par l'entreprise Google à partir de la base de données de Wikipédia. Nous appuierons cette recherche par l'exploration et l'analyse des fonctionnalités et des discours d'accompagnement de l'outil XTools. Nous nous fonderons également sur l'entretien réalisé en avril 2019 avec son développeur actuel, contributeur prolifique et administrateur de la Wikipédia anglophone devenu, depuis 2016, un employé de la Fondation Wikimedia intégré à l'équipe *Community Tech*. Cet article s'inscrit dans les études sur l'organisation de Wikipédia et son mode de gouvernance (Langlais, 2014), (Fallery et *al*., 2013), (Pentzold, 2010), (Auray et *al*., 2009), (Konieczny, 2009), (Levrel, 2006) et les prolonge par son attention portée au développement d'outils peu étudiés dans le cadre d'une réflexion sur le fonctionnement du travail dans Wikipédia.

2. Des outils d'analyse de données pour l'identification de *fake news* ?

Au sein de Wikipédia la tâche d'identification de *fake news* est comprise dans une activité plus vaste de détection du vandalisme. Parmi les différents types de vandalisme possibles (suppression du contenu d'un article, ajout d'insultes ou de propos familiers…) l'identification d'ajout d'informations volontairement fausses nécessite une attention particulière car elle est plus difficile à effectuer de manière semi-automatique. En effet, si les *bots* wikipédiens suivent avec attention les modifications et repèrent et corrigent en quelques minutes les plus suspectes, interpréter un ajout de texte comme étant une *fake news* nécessite de croiser plusieurs paramètres. Selon le développeur de

l'outil d'analyse de données massives XTools, celui-ci peut servir à identifier les comptes dissimulés, c'est-à-dire les comptes d'utilisateurs ayant déjà un compte et utilisant leur second compte à des fins de mainmise sur certains articles. Croiser les données de contributions permet alors de reconnaître les contributeurs initiés derrière les nouveaux comptes. Des contributeurs humains mobilisent donc ces données *via* l'interface web pour identifier des contributions ou des contributeurs déviants.

Les interfaces de programmation des outils d'analyse de données massives sont quant à elles intensivement utilisées par les *bots*, plus de 130 000 fois juste pour une seule journée. L'interface de programmation d'XTools est à elle seule visitée plusieurs milliers de fois par jour par des *bots* et des programmes automatisés. Les outils d'analyse de données massives comme XTools permettent de nourrir les *bots* d'informations leur permettant de fonctionner. Le développement des outils vise ainsi à fournir toujours plus de données aux *bots* qui semblent particulièrement visés par la devise « *Feeding your data hunger* ».

Ainsi, même si XTools n'est pas présenté comme un outil officiel de lutte contre le vandalisme il est utilisé dans ce but par la communauté, de manière directe ou à travers ses *bots*. Néanmoins, il semble que cet aspect soit secondaire et que l'outil produise des données sans que leurs usages et leurs usagers soient clairement identifiés. Ainsi, l'employé de la Fondation Wikimedia qui a réécrit l'outil XTools avec d'autres employés et de nombreux bénévoles pour fournir des données beaucoup plus complètes conclut notre entretien sur la phrase suivante « We have no knowledge of who is using XTools, so we would need to do a survey to find out exactly why they're using it ».

Le développeur de l'outil XTools assure toutefois que ce dernier est utilisé par certaines équipes de la Fondation Wikimedia, or la Fondation n'a pas d'activité éditoriale officielle quant à Wikipédia qui est censée être gérée par les contributeurs eux-mêmes (Jacquemin, 2011). Dans ce contexte, quelles peuvent être les finalités à la fois de la production de données massives et du financement d'outils d'analyse de données très complets pour un site sans annonceur ni activité de revente des données personnelles ? En effet, la collecte et le traitement de données constituent une activité à très forte valeur marchande, voire la valeur économique principale des plateformes commerciales telles qu'Amazon ou des réseaux sociaux tels que Facebook. Les traces volontaires et involontaires des internautes y sont ainsi scrutées par des humains et des algorithmes afin à la fois de proposer de la publicité ciblée programmatique et d'entraîner des algorithmes à améliorer leur performance et leur compétence pour adresser des contenus ou suggérer des produits supposément adaptés aux goûts et aux préférences détectés des utilisateurs. Dans le cas de Wikipédia, quelle valeur les données produites prennent-elles ?

3. Les données comme indicateurs de performance

À travers cet article, nous souhaitons interroger le lien entre données et qualité. Néanmoins, ce lien, s'il existe, n'est pas toujours aussi direct que la simple consultation des données dans le but d'identifier des manquements ou des dégradations à rectifier. Le lien entre données et qualité repose aussi sur un processus de réputation et de reconnaissance basé, entre autres, sur la quantification des performances des contributeurs. Ainsi, si la version originale d'XTools, créée en 2008, ne permettait pas de compter le nombre d'éditions par éditeur, le principal développeur actuel de XTools présente ce dernier comme un outil permettant de fournir des indicateurs sur la productivité des contributeurs. Ce point est ainsi mis en avant tout au long de notre entretien :

> The Edit Counter is the most popular tool, and is used by editors who want to see how much of an impact they've made, for example. Another example is users who want to attain extra privileges on the wiki. Administrators may use the Edit Counter or the Pages Created tool to see if they are suitable for the position.

Ces outils d'analyse de données produisent ainsi à leur tour des données. Cette production de données ne sert pas seulement à documenter la productivité mais à la stimuler, à la fois car les grands contributeurs cherchent à être dans le « top editors » (contributeurs ayant enregistré le plus de modifications) des articles pour lesquels ils s'investissent, mais aussi car cette preuve quantitative de leur travail peut leur permettre d'évoluer au sein de l'encyclopédie. Ces outils mettent ainsi en œuvre une évaluation des contributeurs par la quantification de leur activité.

Ce management de la performance via les données chiffrées fonctionne d'autant mieux dans Wikipédia que le système de reconnaissance par la profession ou par l'institution des contributeurs ne s'y applique pas en théorie (Willaime, 2015). La communauté crée donc un système de reconnaissance interne pour susciter un engagement plus fort de la part de ses membres. Certains dispositifs de reconnaissance sont qualitatifs et interpersonnels : messages d'encouragement sur les pages de discussions des utilisateurs, récompenses ou prix symboliques pour le travail effectué sur un article. Ces dispositifs, bien qu'informels, comptent pour les contributeurs qui les valorisent sur leur page et les aident à accéder à des responsabilités (administrateur,

arbitre…). Ainsi, les trois membres actuels du comité d'arbitrage[1] de la Wikipédia francophone affichaient déjà, avant d'être élus, des récompenses ou des données chiffrées concernant leur activité. La mise en œuvre de procédés de valorisation et de gestion de la réputation s'illustre par leurs pages d'utilisateurs. L'un d'eux, sous une rubrique qu'il a intitulée « Vanitas » expose les trois wikiconcours qu'il a gagnés. Un autre présente sur sa page la récompense « The #100wikidays Barnstar » décerné par un autre utilisateur qui le félicite d'avoir réussi ce challenge qui se présente comme « un défi personnel dans lequel une personne vise à créer (au moins) un article par jour pendant 100 jours consécutifs »[2]. Là encore, les exploits et déboires des wikipédiens participants sont renseignés par des données sur leur progression et leurs résultats dans les pages consacrées à ce challenge. Le troisième arbitre n'a pas de prix affiché sur sa page d'utilisateur mais a, lui aussi, participé à des wikiconcours. Il affiche sur sa page le nombre de contributions qu'il a effectuées depuis le début de sa participation à Wikipédia, soit plus de 34 000. Il présente également la liste des très nombreux articles qu'il a créés et les articles qu'il a améliorés et qui ont reçu le label Article de qualité ou le label Bon article.

Du fait du caractère ouvert de Wikipédia et des nombreuses critiques émises envers son mode de fonctionnement, en particulier durant les premières années d'existence de l'encyclopédie, de nombreux outils et dispositifs ont été mis en place pour améliorer la qualité des articles (Vandendorpe, 2008 : 21). Cette attention portée à la qualité s'est traduite par la création de niveaux d'avancement, ceux-ci traduisant une progression dans la qualité (ébauche ; bon début ; bien construit ; avancé) et de labels (Bon article ; Article de qualité). Les niveaux d'avancement témoignent d'une volonté de stabiliser les contenus. Les critères d'attribution du label Article de qualité révèlent, pour leur part, les normes visées en matière de qualité.

Trois dimensions doivent être évaluées par les contributeurs pour décerner le label : la qualité encyclopédique pour laquelle sont citées comme composantes la clarté, l'exhaustivité, la neutralité, la pertinence et la citation des sources ; la qualité de la finition qui comprend l'orthographe, la syntaxe, la typographie, la mise en page et les illustrations ; et le respect des licences de droits d'auteurs pour le texte et les images. D'autres critères généraux sont proposés dans une page dédiée à cette question qui

[1] « Le comité d'arbitrage est un groupe de wikipédiens élus par la communauté, qui est chargé de régler les conflits entre les participants. Il a le devoir de s'assurer que toutes les possibilités de médiation sont mises en œuvre, et le pouvoir de décider, si nécessaire, des mesures à prendre, qui vont de l'avertissement au blocage temporaire ou définitif des utilisateurs en conflit. » Voir la page « Wikipédia : Comité d'arbitrage ».

[2] Voir la page « 100wikidays » https://meta.wikimedia.org/wiki/100wikidays.

précise que la taille d'un article ne constitue pas un critère de qualité en soi mais doit être adéquate au sujet. Certains critères y sont aussi spécifiés : L'article doit respecter les conventions de style concernant son introduction et sa structuration, être bien écrit (maîtrise de la langue, homogénéité de la typographie), complet (le sujet est traité en totalité), argumenté (justification des faits énoncés par des preuves et des références) et neutre (l'article ne prête pas à controverse).

On observe ainsi un couplage entre les données quantitatives et l'usage de labels qui, dans le cas de Wikipédia, reposent en théorie sur des critères qualitatifs. Les labels auraient alors la triple fonction d'orienter le lecteur vers des articles reconnus par la communauté, de motiver les contributeurs à effectuer un intense travail de relecture des articles et de les inciter à respecter les critères de qualité afin d'obtenir une qualité homogène et normée. Les données quantitatives constituent quant à elles un repère pour les wikipédiens et pour la Fondation Wikimedia pour situer les contributeurs au sein de Wikipédia et une preuve de performance mobilisable par les contributeurs pour monter les échelons du projet encyclopédique et accéder à une place (rémunérée pour un très petit nombre d'entre eux) au sein de la Fondation. On constate ainsi que deux régimes d'évaluation coexistent, l'un qualitatif estimant la valeur des articles de l'encyclopédie, l'autre quantitatif, estimant la valeur des contributeurs.

L'hypothèse que les données de l'outil XTools servent à des fins de management de la qualité se renforce en observant les dix outils qu'il recouvre et met à disposition. Si l'on retrouve les outils liés à la quantification des activités des contributeurs et à la visualisation de leur place dans le Top éditeur des articles, d'autres fonctions semblent encore davantage servir un management par les données. Ainsi, l'outil « Admin Stats » produit des statistiques concernant les actions des administrateurs et l'« Admin Score » indique si un contributeur est apte à devenir administrateur. Le score est ainsi produit par un algorithme basé sur treize indicateurs pondérés par un multiplicateur : l'ancienneté du compte, le nombre d'éditions effectuées, la taille de la page d'utilisateur, le nombre d'articles suivis, le nombre de blocages administrés, le nombre de modifications sur des articles risquant d'être supprimés, le nombre d'éditions récentes (dans les deux dernières années), le nombre d'interventions sur les pages contre le vandalisme, le nombre d'éditions dans l'espace encyclopédique contenant un résumé de la modification effectuée, le nombre d'éditions dans l'espace encyclopédique, le nombre de pages créées encore effectives, le nombre de pages créées qui ont ensuite été supprimées et le nombre de requêtes pour la protection de pages. Cet outil est présenté comme « donnant un aperçu concis de la valeur administrative d'un contributeur ». La valeur et la carrière des contributeurs se fondent donc sur les données analysées par

l'outil XTools à partir de la base de données massives de la Fondation Wikimedia et transformées en indicateurs de performances.

L'usage des données dans Wikipédia se rapproche en fait de la description qu'Antonio Casilli fait du management des plateformes numériques :

> C'est par exemple au travers de métriques de performances (les likes, les scores, les évaluations, les étoiles, mais aussi le nombre de followers, de partages, de contacts) que l'effort productif des utilisateurs est quantifié. Ces indicateurs sont souvent associés à des mécanismes de ludification (gagner des badges, des goodies, déclencher des animations en échange de données personnelles et de contributions), de compétition (comparaison des scores entre différents utilisateurs d'une plateforme permettant de les classer en fonction de leur rendement) ou d'autoévaluation (bilan d'activité, analyse des plages horaires les plus productives, etc). (Casilli, 2019 : 260)

On retrouve en effet les métriques de performances – à travers les outils d'analyse de données Top Edits, Admin Stats et Admin Score – les mécanismes de ludification – prix symboliques et humoristiques à collectionner, Wikithon – de compétition – Wikiconcours –, et d'autoévaluation – à travers les outils Edit Counter, Pages Created, Simple Counter. Or, pour Antonio Casilli, ces types de métriques constituent des instruments de contrôle. Si les sanctions dans Wikipédia ne peuvent être liées à une baisse de productivité, les outils de mesure décrits participent bien de la stimulation de celle-ci. Les données des outils servent de plus à la surveillance mutuelle des contributeurs ainsi qu'à fonder les accusations de mauvaise foi ou de vandalisme (Cardon, Levrel, 2009).

4. Des data au service des algorithmes des plateformes numériques

Les données des outils d'analyse de Wikipédia sont présentées par le responsable de XTools comme servant la recherche. Si de nombreuses recherches portent sur Wikipédia, beaucoup se nourrissent d'elle dans une optique différente. Au-delà d'XTools, qui traite des données de la database pour les rendre directement accessibles, l'ouverture de la base de données de la Fondation Wikimedia sert également à des recherches à très grande échelle. Parmi les plus importantes figurent celles de l'entreprise Google. Outre, depuis 2014, la récupération et l'affichage dans le Google

Knowledge Graph du contenu de Wikipédia sur son moteur de recherche – qui a d'ailleurs fait perdre à celle-ci une part importante de son trafic (McMahon et *al.*, 2017) – l'encyclopédie est aussi convoitée pour ses données. Ainsi, Google utilise les données de Wikipédia via l'interface data DBpedia pour améliorer son moteur de recherche : « Les contributeurs de Wikipédia se retrouvent alors à fournir des contenus et des données pour les grandes plateformes et à alimenter leurs intelligences artificielles » (Casilli, 2019 : 182). En effet, Wikipédia se trouve aussi sujette à des formes de captation secondaire de la part du réseau social Facebook qui se sert d'elle à des fins d'automation et de vérification des sources d'information.

L'instrumentalisation de Wikipédia par Google à travers l'exploitation de ses données semble compensée par les généreuses donations de l'entreprise. L'encyclopédie produit ainsi des externalités positives pour la plateforme Google qui se trouve aussi être l'un des principaux donateurs de Wikipédia. Enfin, en plus de puiser dans les données de Wikipédia, Google fait créer de nouvelles données aux contributeurs de celle-ci. Outre les trois millions de dollars versés en janvier 2019 à la Fondation, Google a également offert deux outils de *Machine Learning*[3] qui au-delà de leur fonction éditoriale, produiront également de nombreuses données observées dans la perspective là encore de renforcer les intelligences artificielles.

5. Conclusion

Notre recherche s'interrogeait sur les objectifs de gouvernance auxquels la production de data et la récente réécriture d'outils d'analyse de données participaient. Si ces données sont utilisées dans de très divers cas, il semble que l'un de ses principaux impacts se situe dans l'incitation à la performance via un système de récompenses plus ou moins informelles, dont l'horizon de consécration est incarné par la perspective d'une activité officielle, voire professionnelle, au sein de la Fondation Wikimedia. Par ailleurs, la Fondation semble consciente de la valeur des données qu'elle produit et, si elle ne les vend pas, ces données n'en constituent pas moins un levier efficace pour obtenir des « dons » de la part des grandes plateformes du numérique cherchant à nourrir leurs intelligences artificielles du travail des contributeurs.

[3] https://www.lebigdata.fr/google-wikipedia-machine-learning.

6. Bibliographie

Cardon D., Levrel J., « La Vigilance participative. Une interprétation de la gouvernance de Wikipédia », *Réseaux*, 2, n°154, 2009, pp. 51-89.

Casilli A., *En attendant les robots*, Seuil, 2019.

Fallery B., Rodhain F., « Gouvernance d'Internet, gouvernance de Wikipédia : l'apport des analyses d'Elinor Ostrom sur l'action collective auto-organisée », *Management & Avenir*, 2013/7 (N° 65), pp. 169-188.

Jacquemin B., « Autorégulation de rapports sociaux et dispositif dans Wikipédia », *Document numérique*, (Vol. 14), 2011, pp. 57-79.

Konieczny P., « Wikipedia : Community or Social Movement ? », *Interface : a Journal for and about Social Movements*, 2009, pp. 212-232.

Langlais P.-C., « La Négociation contre la démocratie : le cas Wikipédia », *Négociations*, 2014/1 (n° 21), pp. 21-34.

Levrel J., « Wikipédia, un dispositif médiatique de publics participants », *Réseaux*, 2006/4 (n°138), pp. 185-218.

McMahon C., Johnson I., Hecht B., « The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies », Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017), Association for the Advancement of Artificial Intelligence, 2017.

Pentzold C., « Imagining the Wikipedia community: What do Wikipedia authors mean when they write about their 'community'? », *New Media & Society*, 2010, pp. 704-721.

Vandendorpe C., « Le Phénomène Wikipédia : une utopie en marche », *Le Débat*, 2008/1 (n° 148), pp. 17-30.

Willaime Pierre, « Une Analyse épistémologique de l'expertise dans Wikipédia », in Barbe L., Merzeau L., Schafer V., (sous la direction de), *Wikipédia, objet scientifique non identifié*, Nanterre, Presses Universitaires de Paris X, 2015, pp. 105-120.