Wikipédia : avec quels encyclopédistes ?

Léo Joubert

Aix-Marseille Univ., CNRS, LEST, France

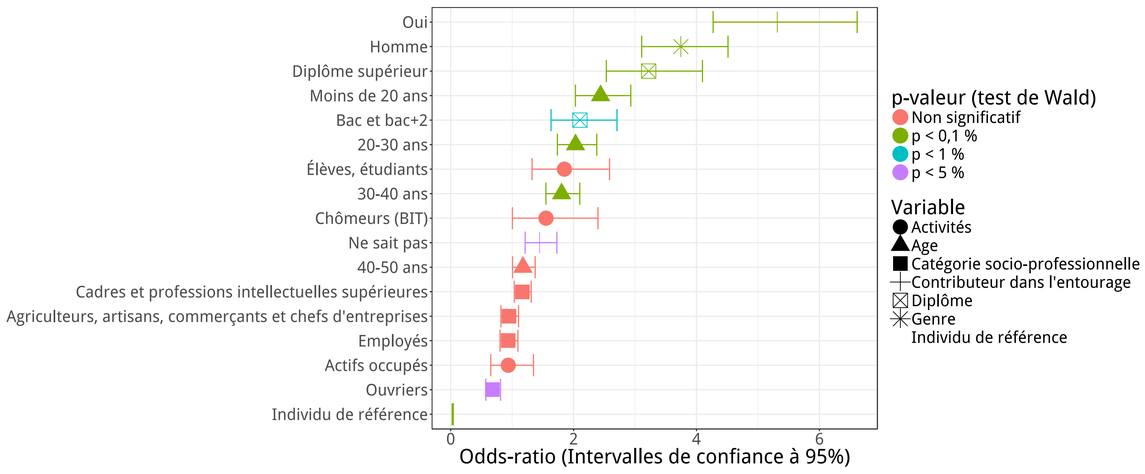

Figure 1: Odds-ratio de la probabilité de commencer à contribuer

Wikipédia ?

Après sa mise en ligne en janvier 2001 pour la version française et en avril 2001 pour la version anglophone, Wikipédia est rapidement devenue un des sites web les plus fréquentés. Fonctionnant sur le principe du wiki, la seule règle vraiment constitutive d'un tel site est la suivante : *tout le monde peut tout modifier n'importe quand et n'importe comment.* Les activités de contributeur sont très diverses : créer des articles, améliorer du texte existant, classer les articles dans des catégories, surveiller les contributions des autres, trancher les conflits entre contributeurs... Pour les deux dernières, certains contributeurs utilisent des outils particuliers : c'est le cas des administrateurs qui peuvent bannir un contributeur ou bloquer une page en écriture. *Les administrateurs sont désignés par le biais d'une élection à laquelle tous les contributeurs peuvent participer. C'est de cette élection dont il est question à plusieurs reprises dans ce qui suit.*

Résumé

Faire une sociographie des contributeurs de Wikipédia pose beaucoup de problèmes méthodologiques à cause de la forte dispersion géographique de la population. Non dépourvue de biais, l'enquête Wikipédia 2015 menée par le GIS Marsouin est néanmoins une des tentatives les plus intéressantes par l'ampleur de son échantillon et la précision des questions qu'elle couvre. Ce travail s'appuie sur les données issues de cette enquête pour offrir des éléments de réponse aux deux questions suivantes : **les caractéristiques sociodémographiques expliquent-elles la probabilité de contribuer au moins une fois ?** et **parmi les contributeurs, ceux qui avaient une forte probabilité de contribuer contribuent-ils différemment des autres ?**.

La probabilité de contribuer au moins une fois dépend (par ordre de significativité) du fait de connaître un autre contributeur, du genre, du diplôme et de l'âge. La catégorie socioprofessionnelle n'est pas significative. L'espace des activités sur Wikipédia est structuré sur au moins deux axes : l'engagement du contributeur qu'une activité donnée suppose et l'influence qu'elle a sur le contenu. *Une probabilité d'entrée forte influence la position d'un contributeur dans cet espace de façon significative, mais trop faiblement pour ne pas ressentir la nécessité de trouver d'autres explications.*

Les répondants de l'enquête Wikipédia 2015

Pour diminuer la complexité de l'analyse ainsi que pour mieux évaluer la représentativité de la population, les analyses présentées ici sont restreintes aux répondants de France Métropolitaine. Cette restriction exploratoire n'exclut bien sûr pas de futurs élargissements. La base de données est issue de l'enquête Wikipédia 2015 menée par le GIS Marsouin. Elle a été construite en diffusant un questionnaire par le biais d'un bandeau affiché sur la page d'accueil de la version francophone de Wikipédia entre fin février et fin mars 2015. Ce mode de passation du questionnaire répond à des contraintes particulières de l'objet de recherche : **il est très difficile de réaliser un échantillonnage représentatif de la population des wikipédiens dans la mesure où l'on ne peut accéder à elle que par le truchement de l'outil qui la fait exister**. Un biais évident est alors celui qui touche toutes les enquêtes déclaratives, mais qui se fait particulièrement sentir ici : **on ne sait pas dans quelle mesure l'outil de mesure - le questionnaire - n'a pas conçu lui-même un artefact en sélectionnant son public cible**.

Des analyses ont déjà été faites par le collectif Marsouin ; elles rapportent que la population des répondants de l'enquête à celle des internautes français en prenant pour source de référence l'enquête ARCEP du CREDO. **Il ressort de cette analyse que les répondants de l'enquête sont plus jeunes. Les cadres et professions intellectuelles supérieures sont nettement surreprésentés alors que les ouvriers, les employés et les professions intermédiaires sont sous-représentés.**

Nous avons conduit une analyse similaire en prenant cette fois pour référence la population française âgée de 15 ans et plus sondée par l'enquête emploi en continu de l'INSEE (2015). Il ressort **une forte surreprésentation des hommes (63 % dans Wikipédia 2015 ; 48 % dans la population française)**. Les répondants sont également **beaucoup plus fortement diplômés (43 % ont un diplôme du supérieur dans Wikipédia 2015 ; 14 % dans la population française)**. Les autres surreprésentations recoupent celles déjà pointées sur la population des internautes avec une amplification notable : **on trouve un *effet questionnaire* - tous les internautes ne répondent pas - accentuant un *effet internaute* - tous les français ne sont pas des internautes**.

Contribuer : une action située socialement ?

Après avoir retiré les valeurs manquantes du tableau, il reste 11 124 individus. La base de données permet de distinguer 5 niveaux de contribution : *Jamais* (67 %), *C'est déjà arrivé* (19 %), *Quelques fois* (8 %), *Vous contribuez régulièrement* (3 %), *Vous vous considérez comme un gros contributeur* (1 %). Comme nous cherchons à mesurer la probabilité de commencer à contribuer, nous regroupons ces niveaux en deux : *Jamais* (67 %) et *Au moins une fois* (33 %). Il s'agit de la variable de réponse du modèle. Les variables explicatives sont listées dans la figure 1 qui récapitule les coefficients du modèle sous la forme d'odd-ratio. Chacun des coefficients doit être interprété comme une hausse plus ou moins significative par rapport à un individu de référence dont les caractéristiques sociodémographiques rendent le moins probable la contribution. **L'individu de référence est une femme, âgée de 50 à 60 ans, appartenant aux professions intermédiaires, retraitée, sans diplôme et qui n'a pas de contributeurs dans son entourage.** Toutes choses égales par ailleurs, la probabilité de contribuer au moins une fois varie de la manière suivante :

- Le fait de **connaître un contributeur** multiplie par au moins 4 la probabilité de contribuer.
- Le fait d'**être un homme** la multiplie par au moins 3.
- Le fait d'**avoir un diplôme du supérieur** la multiplie par au moins 2,5. Celui d'avoir le bac ou un cursus court à bac+2 la multiplie par au moins 1,5, même si la significativité baisse.
- L'effet de l'âge dépend des modalités de la variable : **avoir moins de 20 ans** multiplie la probabilité par au moins 2, **avoir entre 20 et 40 ans** multiplie la probabilité par au moins 1,5.

On ne remarque pas d'effet significatif de la variable activité. Par rapport à l'individu de référence retraité, être au chômage, être un actif occupé ou être étudiant ne change pas significativement la probabilité de contribuer. **Le très faible effet de la CSP est notable. Alors que la catégorie socioprofessionnelle est généralement considérée comme significative de la stratification sociale de la société française, elle ne permet pas d'expliquer la contribution.** Même si l'odd-ratio des cadres est légèrement au-dessus de 1, le test de Wald ne permet pas de conclure à sa significativité.

Il faut avoir une certaine prudence dans l'analyse de ces résultats : les caractéristiques qui permettent de franchir la barrière à l'entrée sont les mêmes que celles qui sont surreprésentées dans la base de données. Une pondération des individus est ici nécessaire, et des travaux ultérieurs seront nécessaires pour réfléchir au sens sociologique et statistique de cette modification des données : une pondération avec pour référence la population française ou la population des internautes n'a pas les mêmes conséquences analytiques.

Ceux qui avaient une forte probabilité de contribuer contribuent-ils différemment des autres ?

Une entrée plus ou moins probable dans la communauté influence-t-elle ce que l'on fait dans cette communauté ? Pour essayer de répondre à cette question, nous utiliserons notre modèle pour identifier des contributeurs dont le recrutement était prévisible sur la base de critères sociodémographiques, et ceux dont il était improbable. À l'aide d'une analyse de variance, nous testons ensuite la significativité des variations de cette probabilité en fonction de la réponse à la question *en quoi consiste votre contribution à Wikipédia ?* Les items de réponse à cette question étaient les suivants : *1) contribution à l'orthographe, la syntaxe, la typographie, 2) ajout d'une référence, d'un lien externe ou interne, 3) reformulation, clarification d'un contenu existant, 4) contribution à la rédaction des articles, traduction d'un article d'une langue à une autre, 5) création de nouveaux articles, 6) participation aux discussions sur les articles, 7) présentation à une élection des administrateurs du site (administrateur, bureaucrate...).*

Pour cette analyse, nous sélectionnons les contributeurs qui ont répondus *Vous contribuez régulièrement* et *Vous vous considérez comme un gros contributeur* à la question *Avez-vous déjà fait une modification, une contribution, dans Wikipédia ?*. Après suppression des valeurs manquantes, il reste 506 individus.

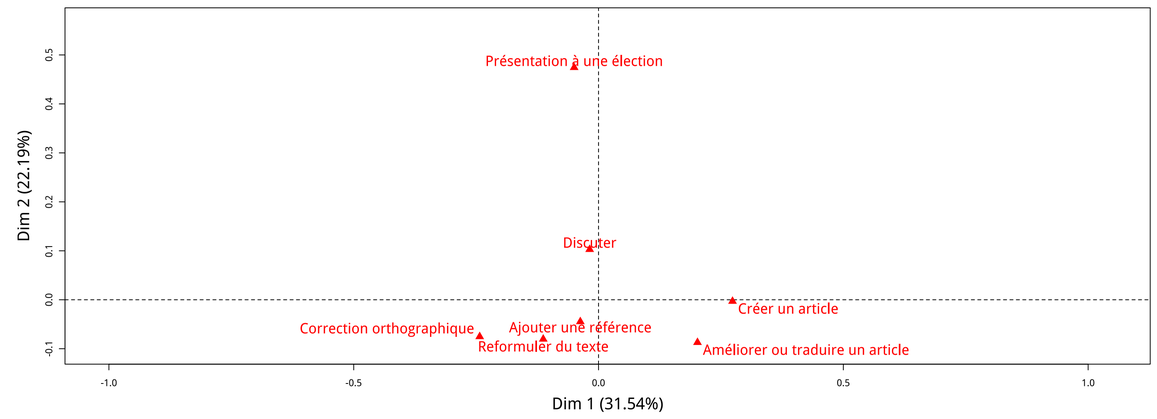

Figure 2: Espace des activités

L'espace des activités wikipédiennes

La figure 2 présente les résultats d'une analyse en composantes principales (ACP) réalisée sur les différents items de réponse.

- **Le premier axe (32 % de la variance)** semble opposer des activités ayant une forte influence sur le contenu de Wikipédia comme la création d'article et l'amélioration d'articles existants, à des activités qui ont une faible influence sur le contenu comme la correction orthographique. Il est notable que la présentation à des élections ne contribue que très peu à la construction de cet axe. En effet, on pourrait s'attendre à trouver parmi les contributeurs pratiquant cette activité des acteurs très influents. C'est un élément qui plaide en faveur d'un modèle de la division du travail wikipédien : il y a ceux qui écrivent, il y a ceux qui régulent, et ce ne sont pas les mêmes.
- **Le deuxième axe (22 % de la variance)** semble opposer des activités qui demandent un fort engagement dans Wikipédia comme la présentation à des élections ou la discussion, à des activités qui ne demandent qu'un engagement léger et ponctuel. Par exemple, l'action de créer un article n'engage pas le contributeur : il peut tout à fait s'en dégager en travaillant sur d'autres articles ou même en proposant au vote la suppression de l'article. En revanche, la discussion ou la présentation à des élections demandent que le contributeur assume publiquement une identité de wikipédien. Les activités situées en haut de cet axe sont pratiquées par peu de contributeurs. Cette petite population pratique également les activités situées en bas de l'axe. Par exemple, 78 % des contributeurs qui se présentent souvent à des élections corrigent des fautes d'orthographe. L'inverse est en revanche complètement fausse : les contributeurs pratiquant les activités en bas de l'axe ne pratiquent que très peu celles du haut. Seuls 8 % des contributeurs qui corrigent souvent des fautes d'orthographes se présentent aux élections.

L'espace des activités est structuré autour de deux dimensions : l'influence de l'activité sur le contenu du wiki, et l'engagement dans Wikipédia nécessaire pour pratiquer cette activité. Cela ouvre la voie pour penser des carrières de contributeur comme l'on pourrait imaginer des flèches sur le plan factoriel. Par exemple, un contributeur pourrait cesser de créer des articles pour se préparer à l'élection d'administrateur. Dans notre modèle, cela revient à penser que son influence directe sur le contenu va baisser, mais que son engagement dans Wikipédia va augmenter.

	Jamais	*Rarement*	*De temps en temps*	*Souvent*	*p-valeur (test F)*
Discuter	0.408	0.466	0.511	0.509	0.002
Ajouter une référence	0.402	0.459	0.484	0.497	0.003
Améliorer ou traduire un article	0.408	0.492	0.496	0.473	0.027
Présentation à une élection	0.463	0.525	0.546	0.567	0.033
Reformuler du texte	0.382	0.481	0.486	0.475	0.077
Créer un article	0.410	0.495	0.494	0.473	0.161
Correction orthographique	0.461	0.469	0.487	0.472	0.719

Tableau 1: Probabilité d'entrée moyenne ventilée par réponse à chaque item. Lecture : les individus qui ont répondu ne jamais discuter avaient une probabilité de 0,408 de contribuer au moins une fois, selon notre modèle.

Type d'activité

Le tableau 1 présente les probabilités moyennes d'entrée dans Wikipédia ventilées par réponse à chacun des items. **Ces probabilités sont calculées à l'aide de la régression présentée à la figure 1**. La dernière colonne donne la probabilité critique de l'analyse de variance. Même quand la différence de probabilité est significative, il faut en souligner la faiblesse. Par exemple, le fait de discuter semble réservé à des contributeurs "bien entrés" : on s'attendrait à ne trouver comme contributeurs bavards sur Wikipédia que des hommes de moins de 20 ans fortement diplômés. Or, lorsque l'on regarde plus précisément les données, on voit bien que l'amplitude n'est que de 10 %.

La probabilité d'entrée n'est pas un critère suffisant pour expliquer le positionnement d'un contributeur dans l'espace des activités wikipédiennes. Par exemple, on ne peut pas dire qu'un contributeur discute d'autant plus qu'il entre dans Wikipédia avec une probabilité plus forte, même si les contributeurs qui discutent le plus souvent sont entrés de façon un peu plus probable dans Wikipédia que ceux qui discutent rarement ou jamais.

La construction sociale des hiérarchies numériques

Ces premiers résultats sont évidemment exploratoires et ne sauraient offrir de conclusions définitives quand aux hiérarchies entre contributeurs. En revanche, il est possible avec eux d'ouvrir un programme de recherche sur la construction sociale de ces hiérarchies. L'hypothèse fondamentale de notre programme est la suivante : *si un contributeur A qui produit des contributions plus durables qu'un contributeur B, alors A occupe une position hiérarchique plus forte que B.* Des travaux en *data science* ont montré qu'une très petite minorité (de l'ordre de 2 %) produit la majorité du contenu durable sur Wikipédia. **Les choses se compliquent lorsque d'autres travaux montrent que les contributions les plus durables ne sont pas nécessairement le fait des contributeurs les plus engagés.** Nous venons quant à nous de montrer que les caractéristiques sociodémographiques n'expliquent pas à elles seules la position d'un contributeur dans l'espace des activités. Mais alors, **comment expliquer que les contributions de certains contributeurs durent beaucoup plus longtemps que celles des autres ?**

Le programme de recherche que nous proposons s'appuie sur des entretiens en cours de traitement. Il vise à comprendre comment les catégories structurantes de la pratique wikipédienne prennent sens pour les contributeurs au fil du temps. **Il nous faut pour cela articuler *l'objectivité d'un dispositif wikipédien de contrôle* et *la subjectivité d'un acteur-contributeur* dont les catégories personnelles (*identité pour soi*) ne sont pas d'emblée congruentes avec les institutions wikipédiennes (*identité pour autrui*).** Au vu de la force de la barrière à l'entrée, nous serions même tentés de conclure que cette congruence est une exception plutôt qu'une règle. Mais surtout, c'est le processus de construction de cette congruence qui est le moteur de la carrière d'un wikipédien. **On devient contributeur lorsque les règles prennent du sens.** C'est à l'intersection du dispositif et du sujet qu'il faut se situer pour comprendre la construction sociale des hiérarchies. **Le fait que Wikipédia ne soient hiérarchisée ni par la force de l'engagement des contributeurs ni par leurs caractéristiques sociodémographiques ne signifie pas qu'il n'existe pas sur Wikipédia une hiérarchie forte. Cela appelle en revanche à *une analyse critique de l'autonomie des hiérarchies wikipédiennes par rapport aux hiérarchies de l'espace social où les contributeurs se recrutent.***

Prolongement

Bien que sommaire et pour l'instant très exploratoire, cette étude nous sera doublement utile dans la conduite de la campagne d'entretiens biographiques. D'une part, il est important de comprendre que les caractéristiques sociodémographiques de l'acteur qui est en train de nous parler ne jouent qu'un rôle très limité (bien que significatif) dans les contours objectifs de sa pratique de contribution. Ainsi, une femme contribue à peu de choses près comme un homme, même s'il est plus fréquent d'interroger des hommes sur leur pratique de contribution dans la mesure où ils sont plus nombreux. D'autre part, l'analyse en composantes principales présentée dans la figure 2 permet de construire une grille d'analyse pour les entretiens : nous sommes par exemple en mesure de comprendre ce qu'implique *en terme de position sociale* le fait de rejeter une activité au profit d'une autre.

Dans des travaux ultérieurs, il faudra mesurer l'influence des caractéristiques sociodémographiques non plus seulement sur l'activité mesurée à l'aide d'items discontinus, mais surtout sur la durée de vie des contributions individuelles.

Remerciements

Ces analyses sont conditionnées par la générosité scientifique du GIS Marsouin qui met à disposition toutes les données issues d'un travail de longue haleine. Notre analyse contient des erreurs inévitables que nous assumons en notre nom propre.