Who becomes a Wikipedia contributor?

Analyze Wikipedia practices through time[1]

Léo Joubert[2] et Nicolas Jullien[3]

This survey was supported by the Brittany Region and the PEPR eNSEMBLE (CONGRATS targeted project), which is working on the future of digital collaboration.

Introduction

How does one become a Wikipedia contributor? It has now been established that participation in Wikipedia obeys a series of correlations that define a precise socio-demographic profile. For example, contributors are younger and are more likely to be male than Wikipedia readers as a whole (Hargittai and Shaw, 2015; Bear and Collins, 2016). That only a tiny minority contributes to online projects in general, and to Wikipedia in particular, is not surprising. On the other hand, the fact that this minority is made up solely of men over 40 with at least one licence raises questions about the goal of universality, both in terms of the empowerment of contributors ('that all can edit') and the coverage of the subjects covered (this is illustrated by the 'pageless' movement, which showed how it was mostly women 'without pages', and gave rise to projects to remedy this, such as the Women in red project).

However, despite its usefulness in understanding contribution, the reasoned inventory of these correlations remains trapped in a binary approach of 'non-contributors' versus contributors (often without specifying the level of regularity required to be considered a 'contributor'). This disregards the multitude of stages an individual goes through before eventually becoming a regular contributor. We know, for example, that the way in which the first contribution or contributions are received by the project is essential in determining the

[1] Marsouin project, 2023
[2] LPED, Université de Rouen Normandie
[3] LEGO, IMT Atlantique

likelihood of a person becoming a regular contributor (Warncke-Wang, Morten, et al. 2023). But in order to take the plunge and contribute at least once, you have to feel able to do so, or even entitled to (Lockett, 2020). Finally, and these surveys all too often forget this, just because you use Wikipedia doesn't mean you know how it is built, or even if you do know how, in practical terms, it is possible to contribute.

If, logically, you can't be a regular contributor without first trying, or trying without first learning how to do it, it's not clear whether or not these different stages depend on identical socio-economic circumstances. For example, are regular contributors (mostly) men because men feel more legitimate than women to contribute? Or because the platform, being digital, conjures up an imaginary computer world from which women are socially excluded? Or because they have more time to contribute on a regular basis? Or are these three effects, among others, mutually reinforcing?

As a result of a learning process and trial-and-error, contributing to Wikipedia requires socialisation, as do all activities, especially digital ones (Proulx, 2002). These activities are correlated with socio-demographic characteristics. The under-representation of women among contributors is a case in point. The hypothesis we are making here, and which we want to study, is that if all these stages are correlated with socio-demographic variables, they may not be correlated in the same way. If this proves to be the case, then efforts to increase the diversity of Wikipedia contributors should take into account the population targeted. It is about trying to explain Wikipedia to 'those who don't know' how to contribute, or to convince 'those who do know' to get started? The actions to be taken in either case are different.

To explore this question, we used data collected by means of an online survey carried out between June and July 2023 among Wikipedia readers (and sometimes contributors) (Cruciani et al., 2023 presents the survey and its data, as well as a wikimedia.org wiki page). The link to the questionnaire was distributed via a banner published in 8 languages on the Wikipedia page (the questionnaire in English can be accessed here). The survey included 200 questions on: what people did on Wikipedia before clicking on the link to the questionnaire; how they use Wikipedia as readers (professional and personal uses); their opinion on the quality, thematic coverage and importance of the encyclopaedia; the creation of Wikipedia (how they think it is created, whether they have already contributed to it and how); their social, sporting, artistic and cultural activities, both online and offline; their socio-economic characteristics, including their political beliefs, and their propensity to trust the encyclopaedia. More than 200,000 people opened the questionnaire, 100,332 began answering it and 10,576 completed it.

Of these 10,576 responses, and after processing the missing values, we retain 7,827. The removal of 27% of respondents is justified by the fact that our modelling requires the use of a large number of independent variables and, consequently, the selection of the sub-population of respondents who answered the questionnaire as completely as possible.

It should be pointed out that our sample is what is known as a 'convenience sample', i.e. with no a priori defined sample design. Consequently, it tells us nothing about the actual distribution in the population that uses Wikipedia in the eight languages surveyed. It does not claim to be representative. On the other hand, if we assume that the respondents do indeed represent the full range of profiles and practices for using the encyclopaedia, it becomes possible to analyse the statistical relationships between practices (in this case, contributing) and the socio-demographic, cultural, professional and other characteristics of the people who engage in these practices. In other words, while we can't estimate proportions ('X% of Wikipedia contributors are women'), our sample does allow us to measure correlations (i.e – for illustrative purposes only: 'women who contribute to Wikipedia are more highly educated than men who contribute to Wikipedia', all other things being equal).

Model presentation

Variable to explain

To construct our variable to be explained, we used the following question: 'Have you ever edited a Wikipedia page?' to which our respondents could answer: 'No, and you don't know how it's done' or 'No, but you know how it's done' or 'Yes, rarely' or 'Yes, a few times' or 'Yes, often'.

This question enables us to model our three stages: knowing how to contribute (or not); having done it (or not); doing it regularly, as shown in Table 1 below. By construction, individuals who answered 'No, and you don't know how to do it' have not passed stage 1 and will therefore not be taken into account in the analysis of the factors that explain passing stage 2 (having contributed). Similarly, only those individuals who have completed stage 2 (i.e. who did not answer 'No, but you know how it's done') will be kept for the purposes of studying the factors that explain the completion of stage 3 (contributing regularly)[4].

[4] From a statistical point of view, this means that we need to ensure that the modelling of step N retains the memory of step N-1. This can be done using the 'Heckman procedure', so that each of the models retains the memory of any previous steps taken, in the form of a coefficient assessing the significance of a 'selection bias'.

<table>
<tr><td>Question :</td><td colspan="5">Have you ever edited a Wikipedia page?</td><td rowspan="2">Total</td></tr>
<tr><td>Modalités</td><td>No, and you don't know how it's done</td><td>No, but you know how it's done</td><td>Yes, rarely</td><td>Yes, a few times</td><td>Yes, often</td></tr>
<tr><td>Stage 1: knowing how to contribute</td><td>no</td><td colspan="4">yes</td><td>100%</td></tr>
<tr><td>Stage 2: have contributed at least once</td><td></td><td>no</td><td colspan="3">yes</td><td>100%</td></tr>
<tr><td>Stage 3: contribute a lot</td><td></td><td colspan="2"></td><td>no</td><td>yes</td><td>100%</td></tr>
</table>

Table 1: modelling of the three stages according to the answer to the question

We ran probit regressions at each stage. Because we need to ensure complete responses to be able to model these stages, the population of our model is smaller than the population of all respondents who completed the questionnaire: we kept only 7,827 responses, of which 3,754 respondents indicated that they knew how to contribute to Wikipedia (and therefore 4,073 did not know). These 3,754 respondents correspond to the population we kept for stage 2, and within this population, 1,857 indicated that they had already contributed to Wikipedia (and therefore 1,897 had never contributed). These 1,857 respondents, 487 of whom indicated that they had contributed often, correspond to the population retained for stage 3.

Explanatory variables

The explanatory variables included variables linked to the respondent, firstly socio-demographic variables (gender, age, diploma). Next, a series of variables included perceived digital competences: the fact of using a computer, of producing complex documents “professionally” (in one's work or studies), or in one's private activities. Other explanatory variables concerned available free time, cultural activities, sports and more, or pro-social activities (such as giving blood, having donated to associations, Wikipedia, etc.) Then, variables were linked to the social environment of the respondent: how Wikipedia is perceived (is there at least one contributor in the entourage?)? Thereafter, the survey included variables related to the respondent’s practice of Wikipedia: the fact of using Wikipedia (professionally and personally), or the characteristics of the first contribution (for those who have already

contributed), or the reasons for not contributing more. The latter variables highlight the fact that while some variables are used at every stage (such as socio-demographic variables), others are only used for those who have already contributed, for example, and are used to explain whether or not they contribute regularly (stage 3 of the model).

First stage: knowing how to contribute

This first model aims to account for the existence of a competence: knowing how to contribute to Wikipedia, which the respondent will feel authorised to mention in a questionnaire. The variable used takes the value 'no' if people answered 'no and you don't know how it's done' (4,073 people) and 'yes' otherwise (3,754 people) - see the first line of the table above.

Three sets of effects are visible for this first stage.

The first set of effects refers to a socio-demographic profile. Respondents who said they knew how to contribute were significantly more likely to be men and young adults (all age groups over 35 had a significantly higher probability of knowing less about how to contribute than the 18-34 age group). It is notable here that, all other things being equal, the level of education is not significantly correlated with knowing how to contribute.

The second set of effects shows us that knowing how to contribute goes hand in hand with a series of social practices. Respondents who say they know how to contribute are also those who often use a computer (at work or at home) and those who have strong digital skills.

The third and final set of factors are more interactional. Firstly, knowing at least one contributor is significantly and positively linked to knowing how to contribute, as is the fact of discussing with the Wikipedia environment (especially if this environment's view of the project is controversial).

These three sets of effects draw a sociological portrait of young adult men who belong to a social environment where Wikipedia is a topic of discussion and who are familiar with digital uses. The idea of belonging to a specific social milieu is reinforced by the idea that these respondents regularly use a computer. All this suggests that knowledge of how to contribute is disseminated in a particular social environment, where the encyclopaedia is part of the information landscape, rather than being learned through teaching the population as a whole. In this respect, the negative effect of gender places us here in a long list of studies that have shown a tendency for women to be excluded from the most technical uses of digital technology.

Second stage: have contributed at least once

Among those who say they know how to contribute, who are the respondents who contribute at least once to Wikipedia? It's important to understand that our new population is made up solely of people who answered 'yes' in step 1, i.e. 3,754 people. We construct a new 'have contributed' variable which takes the value 'no' if people answered 'No, but you know how it's done' to the contribution question (1,897 people) and 'yes' otherwise (1,857 people).

Here again, three sets of effects emerge.

The first set of effects shows a socio-demographic profile that is slightly different from that seen when modelling the first stage. Here again, women are significantly less likely to contribute. In contrast to the first stage, the level of education is a positive correlation: the higher the level of education, the higher the probability of having tried contributing. Age, on the other hand, has no significant effect, even though it is generally a highly discriminating variable when it comes to digital usages. This shows the value of our multi-stage model. Young adults know more than other age groups about how to contribute to Wikipedia - this is what we modelled in stage 1. But among those who know, all ages have the same probability of contributing.

A second set of effects relates to the practices that go with contributing to Wikipedia. Among the social activities we evaluated, it was those linked to digital technology that were positively correlated with the fact of contributing at least once: having commented on news articles online and having experience as an administrator or webmaster... Or having donated money to the Wikimedia Foundation.

A third set of effects emerges by grouping together coefficients relating to cultural practices, barriers to contribution and the importance of Wikipedia. Respondents who have contributed at least once are distinguished by less frequent cultural practices than others and a greater importance attached to Wikipedia. If we consider that these are also highly educated individuals, it seems that Wikipedia plays the role of the main cultural activity here.

The second stage (having contributed, among those who know) seems to indicate that the fact of contributing at least once is closely linked to the fact of feeling legitimate, both in terms of (digital) skills and the fact of having knowledge to contribute.

Third stage: contribute a lot

Of those who say they have contributed at least once, who are the respondents who contribute regularly to Wikipedia? Our third population is therefore made up solely of people who answered 'yes' in step 2, i.e. 1,857 people. We construct a new variable, 'have contributed a lot', which takes the value 'no' if people answered 'yes, rarely' or 'yes, sometimes' to the contribution question (1,370 people) and 'yes' if they answered 'yes, often' (487 people).

The third stage distinguishes two sets of effects.

The first set shows that regular contribution is a choice linked to the context in which the new contributor evolves and is based on idiosyncratic characteristics, rather than belonging to the main social habitus.

Firstly, regular contributors do not ask themselves whether their lack of skills or the fact that they have something else to do is a problem. The question itself is irrelevant, as shown by the fact that large contributors very often cite a lack of free time. If we add to this the negative effect of thinking that contributing is not a time-consuming activity, it becomes clear that the probability of being a major contributor depends on the degree to which the practice of contributing is rooted in a social context. It is linked to personal situation rather than demographic characteristics: knowing at least one contributor has a positive effect, while gender, degree or age have no effect.

The second set of effects relates to Wikipedia. Firstly, it is clear that two divergent positive effects can be discerned for the variable 'knowing another contributor'. People are significantly more likely to contribute a lot if they know no contributor or know at least one contributor, compared with the reference condition of not knowing who among their relatives is a contributor. Similarly, not knowing the year in which you first contributed has a significant negative effect on the probability of being a major contributor.

In other words, contributing a lot is linked to individual socialisation (we know whether people we know are contributors or not) and to attachment to key episodes in this socialisation (we remember when it started), rather than access to information conveyed by society or the social groups to which we belong.

Conclusion

Our modelling enables us to measure more precisely the effect of social profile on contribution. One may be surprised, for example, by the disappearance of the importance of gender in the last stage. This must be understood in relation to what we have said about the first two stages.

This survey does not deny the effect of gender on contribution trajectories, but it does show that this effect acts as a strong barrier to entry into the first two 'stages': compared to men, women are less likely to belong to the social circles where the contribution of knowledge is disseminated, and even if they do belong to this milieu, they are less likely to take the contribution step at least once. A similar result can be observed for age, which has a highly concentrated effect on the first step. Age is above all a factor in knowledge of how Wikipedia works, but hardly a factor in contributory practices. Of course, this must be understood in the context where knowing how to contribute is an a priori condition for contributing.

Our modelling strategy has enabled us to break down the well-known effect of socio-demographic variables on contribution practices, and to formulate more precise recommendations for diversifying the profile of Wikipedians. The three types of action depend on the targeted audience: either teach people to contribute, encourage them to contribute at least once, or encourage them to contribute more. For example, to reach the elderly, you first need to explain how to do it. Once they know, they'll probably feel justified in contributing (often). Young people, on the other hand, generally know how to contribute but don't take the plunge, and that's where leverage is needed. An intervention aimed at a female audience will need to combine two levers: showing how to contribute, and providing a supportive discourse - in which activism can play a role, as the experience of the 'Women in red' project shows.

References

Bear JB, Collier B (2016) Where are the Women in Wikipedia? Understanding the Different Psychological Experiences of Men and Women in Wikipedia. Sex Roles 74(5–1): 254–265.

Cruciani C, Joubert L, Jullien N, Mell L, Piccione S et Vermeirsche J (2023) Surveying wikipedians: a dataset of users and contributors' practices on wikipedia in 8 languages. arXiv preprint arXiv:2311.07964 URL https://hal.science/hal-04279803.

Hargittai, E., & Shaw, A. (2015). Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia. Information, communication & society, 18(4), 424-442.

Lockett, A. (2020). Why do I have authority to edit the page? The politics of user agency and participation on Wikipedia. Wikipedia @ 20 Reagle and Koerner (edit.), MIT Press: 205-220.

Proulx, S. (2002). Trajectoires d'usages des technologies de communication : les formes d'appropriation d'une culture numérique comme enjeu d'une «société du savoir». *Ann. des Télécommunications*, *57*(3-4), 180-189.

Warncke-Wang, Morten, et al. 'Increasing Participation in Peer Production Communities with the Newcomer Homepage.' *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW2 (2023): 1-26.