Wiki Workshop (11th edition) – June 20, 2024

Wikipedia's ordinary readers

Léo Joubert
Univ. of Rouen Normandy
DySoLab
leo.joubert@univ-rouen.fr

Nicolas Jullien
IMT Atlantique
Marsouin-LEGO
nicolas.jullien@imt-atlantique.fr

Laurent Mell
IMT Atlantique
Marsouin-CREAD
laurent.mell@imt-atlantique.fr

Abstract

Are Wikipedia readers ordinary readers? We propose to answer this question by presenting a typology of Wikipedia readers, based on the results of an online questionnaire published in 8 languages on the Wikipedia websites.

Keywords: readers; information access; information trust; cultural activities ; typology

Introduction

The important role of Wikipedia in the search for information and the construction of knowledge has been well documented in the academic literature. The crucial role of reading in the construction of socio-cultural inequalities even more so (Sullivan, 2008). The research presented in this article deepens and precises previous description of the profile of the Wikipedia readers (Lemmerich et al., 2019). It focuses on readers: how do they use Wikipedia? How do they read? In what contexts (school, university, news, etc.)? How much trust do they have in the information published? It provides a better understanding of the determinants of reading practices and characterizes different types of Wikipedia readers.

Methods

This survey of Wikipedia readers (and sometimes contributors) was carried out by a European team of university researchers, managed by the research center Marsouin.org. The method for collecting the data is described in a data paper (Cruciani et al., 2023), and on the research project's Wikimedia page.

The questionnaire included nearly 200 questions about: what people were doing on Wikipedia before clicking on the link to the questionnaire; how they use Wikipedia as readers ("professional" and "personal" uses); their opinion about the quality, the thematic coverage, the importance of the encyclopedia; the making of Wikipedia (how they think it is made, if they have ever contributed and how); their social, sport, artistic and cultural activities; their socio-economic characteristics including political beliefs, and their trust propensities.

Different sources were used to build the questions:

- The questions regarding the reader characteristics (reading habits) were taken from the questionnaire done in 2019 by the WMF;
- The questions regarding reading behaviors and the assessment of the articles' readability benefited from the WMF's "Research:Understanding perception of readability in Wikipedia";
- The questions about the cultural dimensions were taken from the European survey's individuals ICT;
- The socio-demographic data, aimed at capturing the reading/contributing gaps identified by the WMF, are compatible with it and with the European's and Marsouin's surveys; the trust/values questions were taken from the World Value Survey (wave 7).

The data was collected using a questionnaire available online between June and July 2023, via a banner shown to those who opened Wikipedia in the different participating languages, asking to answer a questionnaire (see the English version here). The banner would redirect to a LimeSurvey page managed by the research center Marsouin.org.

Filling out the questionnaire was thus voluntary. More than 200,000 people opened the questionnaire, 100,332 started to answer, and constitute our dataset, and 10,704 finished it.

We created a typology using a mixed method combining factorial analysis and classification to identify the elements that differentiate Wikipedia reading practices. After processing the responses, we constructed synthetic indices corresponding to the dimensions we wished to measure. To do this, we took the answers to the relevant questions from the questionnaire and then, by carrying out a Multiple Correspondence Analysis (MCA) for each group of variables, we selected the most relevant axis in terms of what we were trying to measure.

We identified the following dimensions:

- Proximity to the topic of interest
- Method of searching for information on Wikipedia
- Satisfaction with content accessed
- (Subjective) quality of the last article read

- Frequency of using Wikipedia in a professional and personal context
- Use of Wikipedia as a source for professional and personal work
- Intensity of use of digital devices to access Wikipedia

Based on this selection of relevant axes, we performed a Hierarchical Agglomerative Clustering (HAC) using Ward's method. The classes resulting from the HAC were then consolidated using the K-means partitioning method (See Figure 1 in Appendix).

Results

This classification allowed us to identify five clusters/profiles of Wikipedia "reading practices", whose importance in our dataset is specified in Table 1, and sociodemographic characteristics are specified in Table 2:

- ***Low readers*** have a reduced consumption of information. They have limited trust in Wikipedia. They have more doubts about the reliability and neutrality of published content. They see Wikipedia as just another website rather than an encyclopedic reference. Their use of Wikipedia is close to that of the working class (Pasquier, 2018).
- ***Satisfied*** users see Wikipedia as an encyclopedic knowledge institution. However, they read Wikipedia selectively (one article at a time, in specific contexts). They are very satisfied with the content, but also use other sites for information or education. It is one of the most feminized classes and people belonging to it have the most free time.
- ***Shameful readers*** are the most paradoxical. They are critical of the information they find on Wikipedia, but allow themselves to use it for study and work (more or less citing the source and bibliographical references). They seem uncomfortable with the encyclopedic format. They spend time searching for 'useful' information.
- ***Contributors*** are characterized by the highest proportion of contributors and a high level of trust in Wikipedia. They have a sophisticated approach to searching for information and feel that they dominate the content (very good knowledge of the topic of the page they are visiting).
- For ***Fans***, Wikipedia is an important source of personal (and professional) development. These are the readers with the most extensive and varied use of Wikipedia. This group is characterized by young, active connected people with little free time and a high level of trust in Wikipedia.

Discussion/Conclusions

The typology of Wikipedia readers that we have developed seems to largely reflect an 'ordinary' organization of reading practices. On Wikipedia, we find the same logic of social differentiation in reading practices that we find outside of digital spaces, with distinctions between individuals according to gender, age, educational level, professional and financial situation (Robinson et al., 2015).

There are also inequalities in Wikipedia reading practices: prior knowledge of the subject varies, reading styles differ, content is compared to other information sites to a greater or lesser extent, and confidence in the quality of the information varies. Overall, the typology of Wikipedia readers does not seem to differ from the diversity of readers found outside the Internet.

However, there do seem to be some specific features, particularly in terms of investment in the Wikipedia platform (by editing a Wikipedia page, consulting or writing a message on a "talk page", etc.) The *low readers* include a significant proportion of high contributors (5,000 contributions or more). Reduced reading habits seem to be less discriminating when it comes to investing in Wikipedia. These initial results will be further investigated and refined in the near future with a second survey of the same respondents. For example, the cultural (language) variations would deserve deeper investigation, as the links between reading and contributing practices.

References

[Cruciani et al.2023] Caterina Cruciani, Léo Joubert, Nicolas Jullien, Laurent Mell, Sasha Piccione, and Jeanne Vermeirsche. 2023. Surveying wikipedians: a dataset of users and contributors' practices on Wikipedia in 8 languages. *arXiv preprint arXiv:2311.07964*.

[Lemmerich et al.2019] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the world reads Wikipedia: Beyond English speakers. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 618–626.

[Pasquier2018] Dominique Pasquier. 2018. *L'Internet des familles modestes. Enquête dans la France rurale*. Presses des Mines.

[Robinson et al.2015] Laura Robinson, Shelia R. Cotten, Hiroshi Ono, Anabel Quan-Haase, Gustavo Mesch, Wenhong Chen, Jeremy Schulz, Timothy M. Hale, and Michael J. Stern. 2015. Digital inequalities and why they matter. *Information, Communication & Society*, 18(5):569–582.

[Sullivan2008] Alice Sullivan. 2008. Cultural capital, cultural knowledge and ability. *Sociological Research Online*, 12(6):91–104.

Appendix

Cluster	Percentage
Low readers	12,8%
Satisfied	21,0%
Shameful Wikipedians	18,9%
Contributors	12,1%
Fans	35,3%

Table 1: Readers' classification table

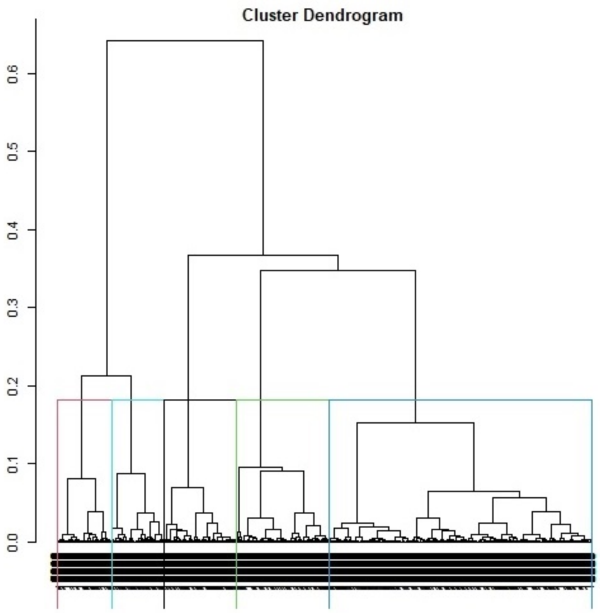

Figure 1: Reader classification dendrogram

	Low readers	Satisfied	Shameful	Contributors	Fans
Gender					
Man	72%	61%	65%	75%	62%
Woman	28%	39%	35%	25%	38%
Age					
Under 18	10%	5%	11%	9%	10%
Between 18 and 34	20%	13%	22%	23%	24%
Between 35 and 44	10%	9%	13%	10%	11%
Between 45 and 54	13%	13%	15%	13%	14%
Between 55 and 64	17%	24%	20%	20%	21%
Over 64	30%	36%	19%	25%	19%
Matrimonial situation					
Single	47%	45%	49%	44%	45%
In couple, not same roof	14%	7%	11%	14%	11%
In couple, same roof	39%	48%	41%	42%	44%
Diploma					
Secondary, primary education or no diploma	33%	25%	27%	22%	23%
Higher education (2 years or less of college education)	15%	18%	16%	15%	17%
Bachelor's degree or equivalent (3 years of College education)	19%	25%	21%	22%	25%
Master's degree and more	33%	33%	35%	41%	36%
Professionnal situation					
Student (middle school to college)	13%	7%	21%	19%	20%
Out of work	50%	57%	33%	36%	31%
Employed	37%	36%	47%	45%	49%
SPC					
SPC-	22%	18%	16%	15%	16%
SPC=	28%	33%	30%	25%	28%
SPC+	50%	49%	53%	60%	55%
Financial resource level					
Life difficult or more	17%	12%	12%	13%	11%
Getting by	35%	31%	33%	31%	30%
Comfortable life	35%	42%	44%	42%	45%
Very comfortable life	13%	16%	10%	15%	14%
Time avaibility					
No free time at all	9%	2%	4%	4%	4%
A little free time	26%	24%	35%	34%	35%
Some free time	38%	38%	42%	41%	41%
A lot of free time	26%	37%	19%	22%	20%

Table 2: socio-demographic characteristics of Readers' cluters