# Wikipedia:Wikipedia Signpost/2020-03-01/Recent research

< Wikipedia:Wikipedia Signpost | 2020-03-01

Recent research

## Wikipedia generates $50 billion/year consumer surplus in the US alone

By Goran S. Milovanović, Tilman Bayer, and Miriam Redi

A monthly overview of recent academic research about Wikipedia and other Wikimedia projects, also published as the Wikimedia Research Newsletter.

### How much would one need to pay readers to give up Wikipedia? $50 billion/year in the US alone.

*Reviewed by Goran S. Milovanović*

From the paper:[1]

> "In Brynjolfsson, Eggers, and Gannamaneni (2017), we propose a new way of measuring consumer welfare using massive online choice experiments. This brief paper motivates the need for such an approach and introduces the method. [...] In some cases, GDP and welfare are correlated, but in many other situations this need not be the case, and even the signs of the changes in GDP and welfare can go in opposite directions. [...] Because it has zero price, Wikipedia is excluded from GDP measures. As a result, the contribution of encyclopedias to GDP decreased because people shifted from paying for Encyclopedia Britannica to consuming Wikipedia for free. However, consumers are clearly better off because they now have access to a much larger quantity of encyclopedic reference for free."

The main motivation in this paper is to illustrate the development of new, behavioral measures of consumer welfare. These measures are motivated by the fact that information from real welfare is not necessarily encompassed by the data used to infer the GDP of an economy. Wikipedia makes a good example of a free "digital good" which creates welfare and at the same time does not contribute to the estimation of GDP anywhere. The paper summarizes the application of massive online behavioral choice experiments (MOBC) in the measurement of consumer surplus (as a proxy measure of consumer welfare), focusing on single binary discrete choice experiments (SBDC) for simplicity. Asked to estimate their willingness to accept (WTA) to give up their access to Wikipedia in exchange for a monetary payment, the participants in the MOBC under the SBDC estimation generated a distribution of monetary equivalents with the median of $150, 95% C.I. = [$124, $182]). When translated to consumer surplus per year in the US alone created by Wikipedia that turns out be around $50 billion.

## Briefly

- See the page of the monthly **Wikimedia Research Showcase** for videos and slides of past presentations.
- The Wikimedia Foundation's Analytics Engineering team has released (https://dumps.wikimedia.org/other/mediawiki_history/readme.html) a complete dataset to analyze the metadata of Wikimedia content and contributors, containing an enhanced and historified version of user, page and revision metadata.

## Other recent publications

*Other recent publications that could not be covered in time for this issue include the items listed below. Contributions, whether reviewing or summarizing newly published research, are always welcome.*

*Compiled by Tilman Bayer and Miriam Redi*

### "Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems"

This preprint[2] describes a "Value-Sensitive Algorithm Design" approach to understanding ORES - a quality prediction system used in Wikipedia. From the abstract:

> "Five major values converged across stakeholder groups that ORES (and its dependent applications) should: (1) reduce the effort of community maintenance, (2) maintain human judgement as the final authority, (3) support differing peoples' differing workflows, (4) encourage positive engagement with diverse editor groups, and (5) establish trustworthiness of people and algorithms within the community. We reveal tensions between these values and discuss implications for future research to improve algorithms like ORES."

*See also research project page on Meta-wiki*

Page 5 of the paper, listing the stakeholder participants in the study, and some of the 24 resulting recommendations

### Despite content saturation, "the activities of editors are still improving with time"

From the abstract of a book chapter titled "Investigating Saturation in Collaboration and Cohesiveness of Wikipedia Using Motifs Analysis":[3]

> "Initially, [Wikipedia's] contents such as articles, editors and edits grow exponentially. Further growth analysis of Wikipedia shows slowdown or saturation in its contents. In this paper, we investigate whether two essential characteristics of Wikipedia, collaboration and cohesiveness also encounter the phenomenon of slowdown or saturation with time. Collaboration in Wikipedia is the process where two or more editors edit together to complete a common article. Cohesiveness is the extent to which a group of editors stays together for mutual interest. [...] We observe saturation in collaboration while the linear or sudden rise in cohesiveness in most of the [top 22] languages of Wikipedia. We therefore notice, although the contents of Wikipedia encounter natural limits of growth, the activities of editors are still improving with time."

### "Individual and collaborative information behaviour of Wikipedians in the context of their involvement with Hebrew Wikipedia"

From the abstract and conclusions:[4]

> "The qualitative study consisted of in-depth semi-structured interviews with Israeli Wikipedians and a content analysis of posts published on talk pages and sandboxes, subpages and drafts, while the quantitative study's data were obtained through structured questionnaires. [...]
>
> A content analysis was performed on ten random posts published on talk pages and sandboxes, subpages and drafts of forty Wikipedians. The quantitative study's data were obtained through structured questionnaires delivered to eighty Wikipedians. [...] Overall, Wikipedians are able to overcome the difficulties that might occur when writing or updating Wikipedia entries on which they have no formal education or expertise. This implies that Wikipedians' individual and collaborative information behaviour supports them in their attempt to fulfil various tasks intended to help construct an important knowledge repository, Wikipedia."

### "Knowledge curation work in Wikidata WikiProject discussions"

This study[5] of how editors participate in Wikidata, identified 6 main activities including: conceptualizing curation, appraising objects, and welcoming newcomers.

### "Building Knowledge Graphs: Processing Infrastructure and Named Entity Linking"

A PhD thesis[6] about extracting knowledge from text via entities, presenting multilingual entity linkers and Docria, a document representation specific for Wikipedia

### "A deep learning-based quality assessment model of collaboratively edited documents: A case study of Wikipedia"

From the abstract:[7]

> "The existing approaches assess Wikipedia quality by statistical models or traditional machine learning algorithms. However, their performance is not satisfactory. Moreover, most existing models fail to extract complete information from articles, which degrades the model's performance. In this article, we first survey related works and summarise a comprehensive feature framework. Then, state-of-the-art deep learning models are introduced and applied to assess Wikipedia quality. Finally, a comparison among deep learning models and traditional machine learning models is conducted to validate the effectiveness of the proposed model. The models are compared extensively in terms of their training and classification performance. Moreover, the importance of each feature and the importance of different feature sets are analysed separately."

*See also related earlier coverage: "Improved article quality predictions with deep learning"*

### "Wikipedia: Why is the common knowledge resource still neglected by academics?"

From the abstract (the author is a Wikipedia editor and current member of the Wikimedia Foundation's Board of Trustees):[8]

> "... the academic world is still treating [Wikipedia] with great skepticism because of the types of inaccuracies present there, the widespread plagiarism from Wikipedia, and historic biases, as well as jealousy regarding the loss of the knowledge dissemination monopoly. This article argues that it is high time not only to acknowledge Wikipedia's quality but also to start actively promoting its use and development in academia."

### "Finding Synonymous Attributes in Evolving Wikipedia Infoboxes"

From the abstract:[9]

> "Policies establish for each type of entity represented in Wikipedia the attribute names that the Infobox should contain in the form of a template. However, these requirements change over time and often users choose not to strictly obey them. As a result, it is hard to treat in an integrated way the history of the Wikipedia pages, making it difficult to analyze the temporal evolution of Wikipedia entities through their Infobox and impossible to perform direct comparison of entities of the same type. To address this challenge, we propose an approach to deal with the misalignment of the attribute names and identify clusters of synonymous Infobox attributes. [...] We formalize the problem as a correlation clustering problem over a weighted graph constructed with attributes as nodes and positive and negative evidence as edges. [...] Our experiments over a collection of Infoboxes of the last 13 years shows the potential of our approach."

### "Weakly Supervised Multilingual Causality Extraction from Wikipedia"

From the abstract:[10]

> "We present a method for extracting causality knowledge from Wikipedia, such as Protectionism -> Trade war, where the cause and effect entities correspond to Wikipedia articles. Such causality knowledge is easy to verify by reading corresponding Wikipedia articles, to translate to multiple languages through Wikidata, and to connect to knowledge bases derived from Wikipedia. Our method exploits Wikipedia article sections that describe causality and the redundancy stemming from the multilinguality of Wikipedia. Experiments showed that our method achieved precision and recall above 98% and 64%, respectively."

### "Temporal Analysis of Entity Relatedness and its Evolution using Wikipedia and DBpedia"

From the abstract:[11]

> Many researchers have made use of the Wikipedia network for relatedness and similarity tasks. However, most approaches use only the most recent information and not historical changes in the network. We provide an analysis of entity relatedness using temporal graph-based approaches over different versions of the Wikipedia article link network and DBpedia, which is an open-source knowledge base extracted from Wikipedia. We consider creating the Wikipedia article link network as both a union and intersection of edges over multiple time points and present a novel variation of the Jaccard index to weight edges based on their transience. We evaluate our results against the KORE dataset, which was created in 2010, and show that using the 2010 Wikipedia article link network produces the strongest result, suggesting that semantic similarity is time sensitive.

### Some of the editors contributing information about the circadian sleep cycle don't have one

From the abstract:[12]

> "We traced the changes made to the [English Wikipedia's] articles for 'Circadian clock' and 'Circadian rhythm' and reviewed the debates that informed them over a span of a decade, using Wikipedia’s native and third-party tools. Specifically, we focused on how groundbreaking research pertaining to the function of biological oscillators was integrated into the articles to reflect a wider paradigmatic shift within the field. We also identified the articles’ main editors to detail the dynamic collective editorial process that took place during a time that saw the field undergo a fundamental change. We discuss the different concerns the academic community has with Wikipedia—specifically regarding its content and its contributors—to ask whether the online encyclopedia’s open model is inherently at odds with scientific culture or whether the model could reflect science or even expand on its core values and practices such as peer review and the idea of communicating science."

As summarized (https://twitter.com/omerbenj/status/990550445109186565) by one of the authors on Twitter: "We graphed when the top editors of the article for Circadian Clocks/Rhythms edited. All worked in circadian cycles - except [one of them], who we identified as an American w sleeping disorders living in Norway".

## References

1. Brynjolfsson, Erik; Eggers, Felix; Gannamaneni, Avinash (May 2018). "Measuring Welfare with Massive Online Choice Experiments: A Brief Introduction". *AEA Papers and Proceedings*. **108**: 473–476. doi:10.1257/pandp.20181035 (https://doi.org/10.1257%2Fpandp.20181035). ISSN 2574-0768 (https://www.worldcat.org/issn/2574-0768). 🔒, eprint version: Brynjolfsson, Erik; Eggers, Felix; Collis, Avinash (2018-05-01). *Measuring Welfare with Massive Online Choice Experiments: A Brief Introduction* (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3169662). Rochester, NY: Social Science Research Network.
2. Smith, C. Estelle; Yu, Bowen; Srivastava, Anjali; Halfaker, Aaron; Terveen, Loren; Zhu, Haiyi (2020-01-14). "Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems". arXiv:2001.04879 (https://arxiv.org/abs/2001.04879).
3. Chandra, Anita; Maiti, Abyayananda (2020). "Investigating Saturation in Collaboration and Cohesiveness of Wikipedia Using Motifs Analysis" (https://books.google.com/books?id=Y3jADwAAQBAJ&pg=PA117&source=gbs_toc_r&cad=3#v=onepage&q&f=false). In Hocine Cherifi; Sabrina Gaito; José Fernendo Mendes; Esteban Moro; Luis Mateus Rocha (eds.). *Complex Networks and Their Applications VIII*. Studies in Computational Intelligence. Cham: Springer International Publishing. pp. 117–128. doi:10.1007/978-3-030-36683-4_10 (https://doi.org/10.1007%2F978-3-030-36683-4_10). ISBN 9783030366834.
4. Lieberman, Yehudit Shkolnisky; Bar-Ilan, Judit (2019-12-15). "Individual and collaborative information behaviour of Wikipedians in the context of their involvement with Hebrew Wikipedia" (http://informationr.net/ir/24-4/paper843.html) (text).
5. Kanke, Timothy (2019-01-01). "Knowledge curation work in Wikidata WikiProject discussions". *Library Hi Tech*. ahead-of-print (ahead-of-print). doi:10.1108/LHT-04-2019-0087 (https://doi.org/10.1108%2FLHT-04-2019-0087). ISSN 0737-8831 (https://www.worldcat.org/issn/0737-8831). 🔒
6. Klang, M. (2019). Building Knowledge Graphs: Processing Infrastructure and Named Entity Linking (https://lup.lub.lu.se/search/ws/files/69709434/Marcus_Corrected_PhD_Thesis.pdf). Ole Römers väg 3, Lund: Department of Computer Science, Lund University

7. Wang, Ping; Li, Xiaodan; Wu, Renli (2019-09-30). "A deep learning-based quality assessment model of collaboratively edited documents: A case study of Wikipedia". *Journal of Information Science*: 0165551519877646. doi:10.1177/0165551519877646 (https://doi.org/10.1177%2F0165551519877646). ISSN 0165-5515 (https://www.worldcat.org/issn/0165-5515). 🔒
8. Jemielniak, Dariusz (2019-12-01). "Wikipedia: Why is the common knowledge resource still neglected by academics?". *GigaScience*. **8** (12). doi:10.1093/gigascience/giz139 (https://doi.org/10.1093%2Fgigascience%2Fgiz139).
9. Sottovia, Paolo; Paganelli, Matteo; Guerra, Francesco; Velegrakis, Yannis (2019). "Finding Synonymous Attributes in Evolving Wikipedia Infoboxes". In Tatjana Welzer; Johann Eder; Vili Podgorelec; Aida Kamišalić Latifić (eds.). *Advances in Databases and Information Systems*. Lecture Notes in Computer Science. Cham: Springer International Publishing. pp. 169–185. doi:10.1007/978-3-030-28730-6_11 (https://doi.org/10.1007%2F978-3-030-28730-6_11). ISBN 9783030287306.
10. Hashimoto, Chikara (November 2019). "Weakly Supervised Multilingual Causality Extraction from Wikipedia". *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. EMNLP-IJCNLP 2019. Hong Kong, China: Association for Computational Linguistics. pp. 2979–2990. doi:10.18653/v1/D19-1296 (https://doi.org/10.18653%2Fv1%2FD19-1296).
11. Prangnawarat, Narumol; McCrae, John P.; Hayes, Conor (2018-12-12). "Temporal Analysis of Entity Relatedness and its Evolution using Wikipedia and DBpedia". arXiv:1812.05001 (https://arxiv.org/abs/1812.05001).
12. Benjakob, Omer; Aviram, Rona (2018-06-01). "A Clockwork Wikipedia: From a Broad Perspective to a Case Study". *Journal of Biological Rhythms*. **33** (3): 233–244. doi:10.1177/0748730418768120 (https://doi.org/10.1177%2F0748730418768120). ISSN 0748-7304 (https://www.worldcat.org/issn/0748-7304). PMID 29665713 (https://pubmed.ncbi.nlm.nih.gov/29665713).

*← PREVIOUS "Recent research"*

*NEXT "Recent research" →*

***+ Add a comment (https://en.wikipedia.org/w/index.php?editintro=Wikipedia:Wikipedia_Signpost/Templates/Comment-editnotice&title=Wikipedia_talk:Wikipedia_Signpost/2020-03-01/Recent_research&action=edit&preload=Wikipedia:Signpost/Templates/Signpost-article-comments-end/preload)***

TO FOLLOW COMMENTS, ADD THE PAGE TO YOUR WATCHLIST (HTTPS://EN.WIKIPEDIA.ORG/W/INDEX.PHP?TITLE=WIKIPEDIA_TALK:WIKIPEDIA_SIGNPOST/2020-03-01/RECENT_RESEARCH&ACTION=WATCH). IF YOUR COMMENT HAS NOT APPEARED HERE, YOU CAN TRY .

*No comments yet. Yours could be the first!*

***+ Add a comment (https://en.wikipedia.org/w/index.php?editintro=Wikipedia:Wikipedia_Signpost/Templates/Comment-editnotice&titl***

***e=Wikipedia_talk:Wikipedia_Signpost/2020-03-01/Recent_research&action=edit&preload=Wikipedia:Signpost/Templates/Signpost-article-comments-end/preload)***

## IN THIS ISSUE

**1 MARCH 2020** (ALL COMMENTS)

*THE SIGNPOST* NEEDS YOUR HELP PUTTING TOGETHER THE NEXT ISSUE.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedia_Signpost/2020-03-01/Recent_research&oldid=1193879079"

▪