The State of Wikimedia Research: 2022–2023

Tilman Bayer
Benjamin Mako Hill
Miriam Redi

Wikimania 2023, Singapore

I'm BMH and presenting with Tilman Bayer today. Not here today, except in spirit is Miriam, but we owe her enormous credit for her work.

This is not my time presenting a version of this talk. The first version of this talk was at Wikimania 2008 and has happened at Wikimania almost every single year since. If you've seen this before, some of this introduction will be getting quite familiar…

So I'm a professor now but in 2008 I was a graduate school starting to build a career that would involve studying Wikimedia projects and the broader free culture movement.

This talk began as an excuse for me to make sure that I was up to date on Wikimedia Research.

"This talk will try to [provide] a quick tour … of the last year's academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project."

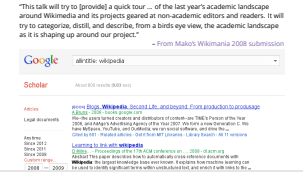
– From Mako's Wikimania 2008 submission

"This talk will try to [provide] a quick tour … of the last year's academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project."
– From Mako's Wikimania 2008 submission

Back at Wikimania 2008, Mako set out to run a session that would provide a comprehensive literature review of articles in Wikipedia published in the last year. Quote Mako:

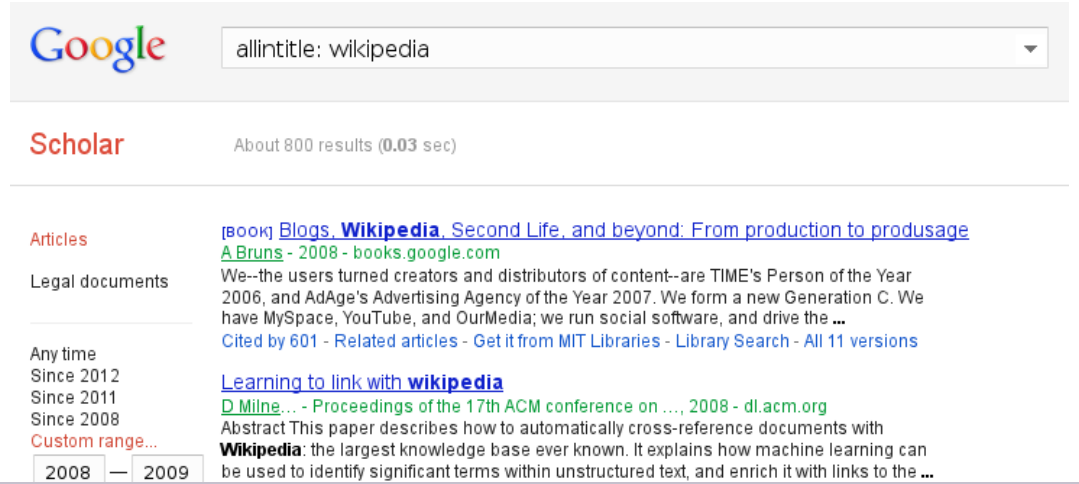
> *"This talk will try to [provide] a quick tour … of the last year's academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project."*
>
> – *From Mako's Wikimania 2008 submission*

"Then, about two weeks before Wikimania, I did the scholar search so I could build the literature."



"This talk will try to [provide] a quick tour ... of the last year's academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project."

– From Mako's Wikimania 2008 submission



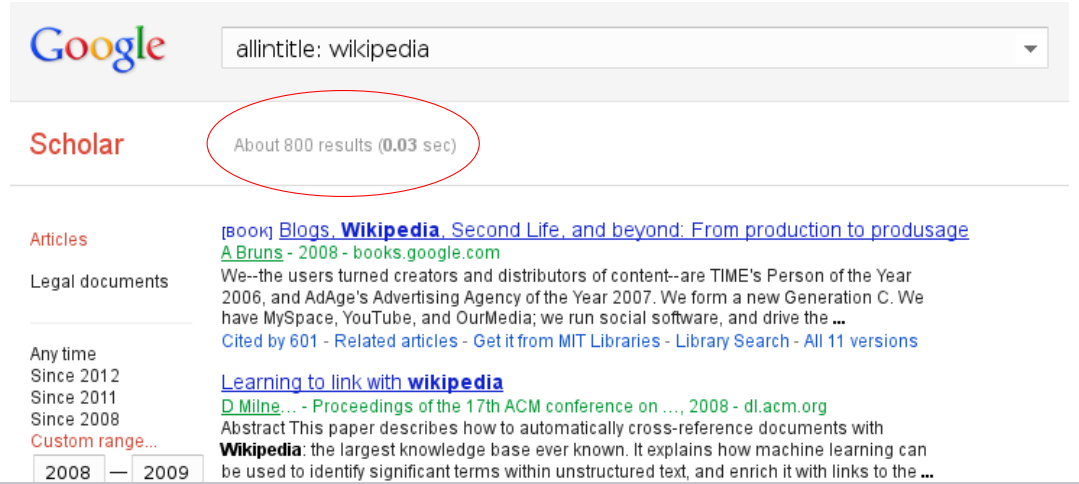
"I tried to import the whole list into Zotero and managed to get banned for abusing Google Scholar because they thought that no human being could realistically consume the amount of material published on Wikipedia that year.

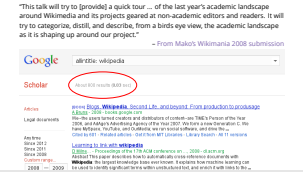
So anyway, I had a 45 minute talk so it worked out to 3.45 seconds to per paper...

And believe it or not, this year is even bigger."

"This talk will try to [provide] a quick tour … of the last year's academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project."

– From Mako's Wikimania 2008 submission

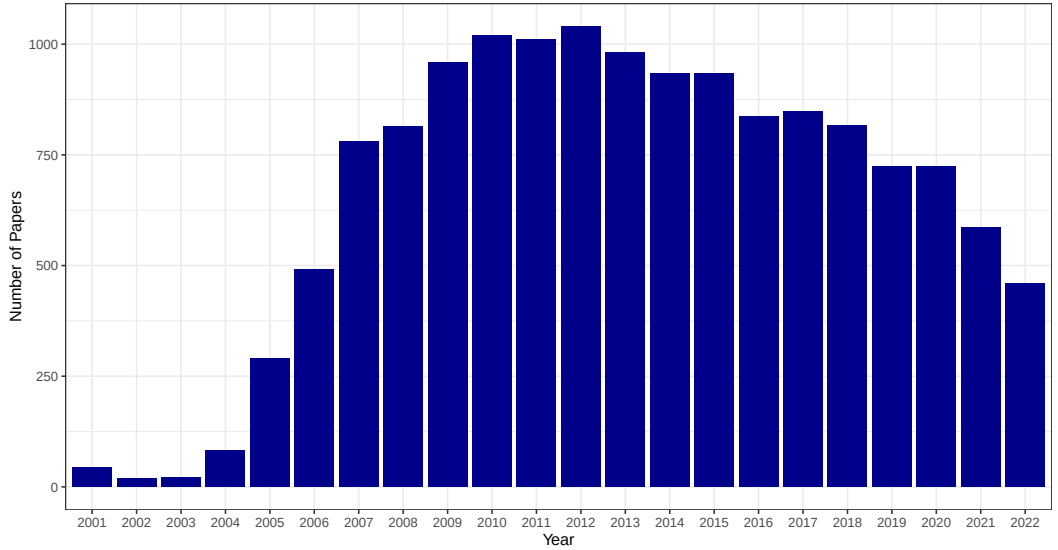




"I tried to import the whole list into Zotero and managed to get banned for abusing Google Scholar because they thought that no human being could realistically consume the amount of material published on Wikipedia that year.

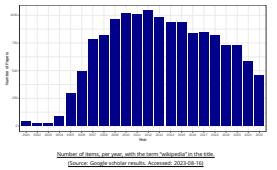
So anyway, I had a 45 minute talk so it worked out to 3.45 seconds to per paper…

And believe it or not, this year is even bigger."



Number of items, per year, with the term “wikipedia” in the title.

(Source: Google scholar results. Accessed: 2023-08-16)



Academics have written a lot of papers about Wikipedia. There are more than 500 papers published about Wikipedia each year and although we’ve reached and moved past a peak it seems, it’s not slowing by much.

- 458 tweets from @WikiResearch account on Twitter/X (covering research papers, events, blog posts etc.)
- 81 recent publications covered in the 13 issues of the Wikimedia Research Newsletter from July 2022 to July 2023 (and hundreds more on our to-do list!)
- 86 extended abstracts presented at the Wiki Workshop 2023 in May 2023

• 458 tweets from @WikiResearch account on Twitter/X (covering research papers, events, blog posts etc.)
• 81 recent publications covered in the 13 issues of the Wikimedia Research Newsletter from July 2022 to July 2023 (and hundreds more on our to-do list!)
• 86 extended abstracts presented at the Wiki Workshop 2023 in May 2023

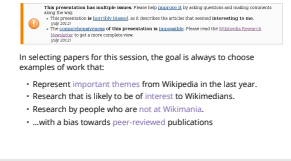
The newsletter aims to be comprehensive, but mostly ignores papers that use Wikipedia as a corpus only (which is popular e.g. in NLP research).

This presentation has multiple issues. Please help improve it by asking questions and making comments along the way.

- This presentation **is horribly biased**, as it describes the articles that seemed **interesting to me**. *(July 2012)*
- The **comprehensiveness of this presentation is impossible**. Please read the Wikimedia Research Newsletter to get a more complete view. *(July 2012)*

In selecting papers for this session, the goal is always to choose examples of work that:

- Represent important themes from Wikipedia in the last year.
- Research that is likely to be of interest to Wikimedians.
- Research by people who are not at Wikimania.
- ...with a bias towards peer-reviewed publications



This is our disclaimer slide...

Within these goals, the selections are incomplete, and wrong.

Themes and Papers

Themes and Papers

So the rest of this talk is going to be structured as 7 research postcards. We're going in about 5 minutes:

- Intro a major theme from research in the last year. Something that's new or important in research that want to highlight.
- We're going to introduce and very quickly summarize the paper
- We're going to quickly reflect on why we think this is important or how we think this could impact the Wikimedia community

And we're going to try to do this 7 times.

Theme 1. Generative AI and large language models

Speaker: Tilman

The most obvious new theme this year in research is also a theme in research in general, and the world in general, and conference program more broadly...

LLM stands for large language models and it's what powers OpenAI's ChatGPT and many other AI applications.

Theme 1. Generative AI and large language models

Semnani, Sina J., Violet Z. Yao, Heidi C. Zhang, and Monica S. Lam. 2023. “WikiChat: A Few-Shot LLM-Based Chatbot Grounded with Wikipedia.” *arXiv*. https://doi.org/10.48550/arXiv.2305.14292

Theme 1. Generative AI and large language models

Semnani, Sina J., Violet Z. Yao, Heidi C. Zhang, and Monica S. Lam. 2023. “WikiChat: A Few-Shot LLM-Based Chatbot Grounded with Wikipedia.” *arXiv*.
https://doi.org/10.48550/arXiv.2305.14292

Goal: ”While LLMs [large language models] tend to hallucinate, our chatbot should be factual.”

Solve this issue by only providing information from a corpus of trusted knowledge - here: English Wikipedia(!)

Goal: ”While LLMs [large language models] tend to hallucinate, our chatbot should be factual.”
Solve this issue by only providing information from a corpus of trusted knowledge - here: English Wikipedia(!)

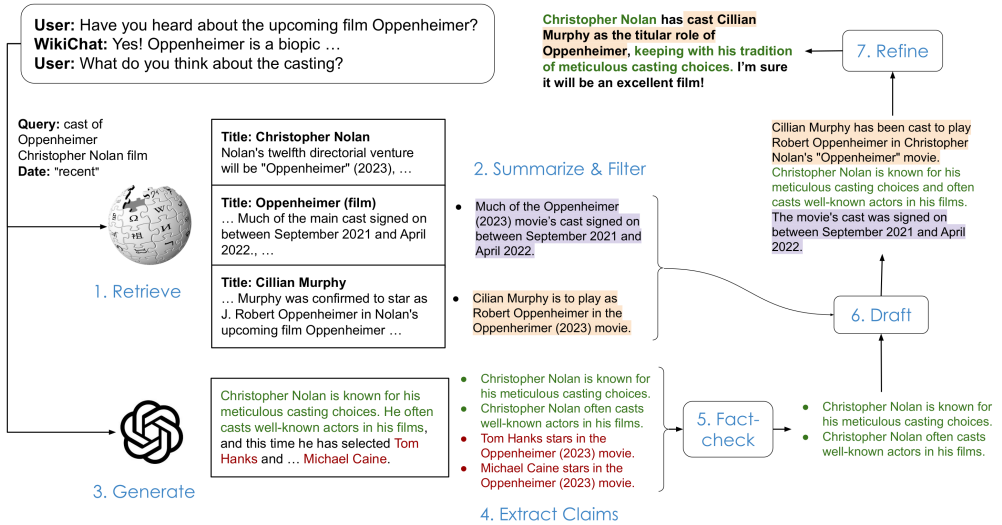
But also: ”some chatbots achieve this by presenting factual but unrelated and repetitive information [...] Therefore, we emphasize that conversationality is also important.”

–> The team needed to use both output from the LLM itself (to continue the chat in a conversational way) and text retrieved from Wikipedia (for fact-checking purposes)

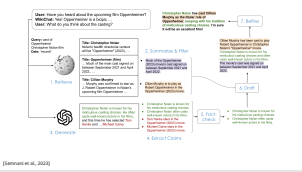
But also: ”some chatbots achieve this by presenting factual but unrelated and repetitive information [...] Therefore, we emphasize that conversationality is also important.”
–> The team needed to use both output from the LLM itself (to continue the chat in a conversational way) and text retrieved from Wikipedia (for fact-checking purposes)

”We prompt the LLM to generate a response to the history of the conversation. The response often contains interesting and relevant knowledge, but is inherently unreliable. We check their correctness against the knowledge corpus.”

Contrast e.g. with the Wikimedia Foundation’s new ChatGPT plugin, which focuses only on the second part (retrieve relevant text from Wikipedia, and have the LLM/ChatGPT answer a specific user’s question based on it)



[Semnani et al., 2023]



"All WikiChat components, and a sample conversation, edited for brevity. The steps taken to generate a response include 1) retrieval from Wikipedia, 2) summarizing and filtering the retrieved passages, 3) generating a response from an LLM, 4) extracting factual claims from the LLM response 5) fact-checking the claims in the LLM response, 6) drafting a response, and 7) refining the response."

The authors also design a new benchmark to evaluate factual accuracy, focused on three kinds of topics:

- familiar topics or ”head topics” (”Examples include Albert Einstein or FC Barcelona”)
- ”tail topics” (occurring at lower frequency in the LLMs pre-training data, e.g. Thomas Percy Hilditch or Hell’s Kitchen Suomi)
- ”recent topics” (which ”are absent from the pre-training corpus of LLMs, even though some background information about them could be present. Examples include Spare (memoir) or 2023 Australian Open”), obtained from a list of most edited Wikipedia articles in early 2023.

They criticize previous LLM accuracy evaluations for focusing too much on the familiar ”head topics”.

The authors also design a new benchmark to evaluate factual accuracy, focused on three kinds of topics:

- familiar topics or ”head topics” (”Examples include Albert Einstein or FC Barcelona”)
- ”tail topics” (occurring at lower frequency in the LLMs pre-training data, e.g. Thomas Percy Hilditch or Hell’s Kitchen Suomi)
- ”recent topics” (which ”are absent from the pre-training corpus of LLMs, even though some background information about them could be present. Examples include Spare (memoir) or 2023 Australian Open”), obtained from a list of most edited Wikipedia articles in early 2023.

They criticize previous LLM accuracy evaluations for focusing too much on the familiar ”head topics”.

Results

”We find that WikiChat outperforms all baselines in terms of the factual accuracy of its claims, by up to 12.1%, 28.3% and 32.7% on head, recent and tail topics, while matching GPT-3.5 in terms of providing natural, relevant, non-repetitive and informational responses.”

NB: The comparison did not include widely used chatbots such as ChatGPT or Bing AI. Instead, the authors chose to compare their chatbot with Atlas (describing it as based on a retrieval-augmented language model that is ”state-of-the-art [...] on the KILT benchmark”) and GPT-3.5 (while ChatGPT is or has been based on GPT-3.5 too, it involved extensive additional finetuning by humans).

TKTK Hacker News frontpage

Results

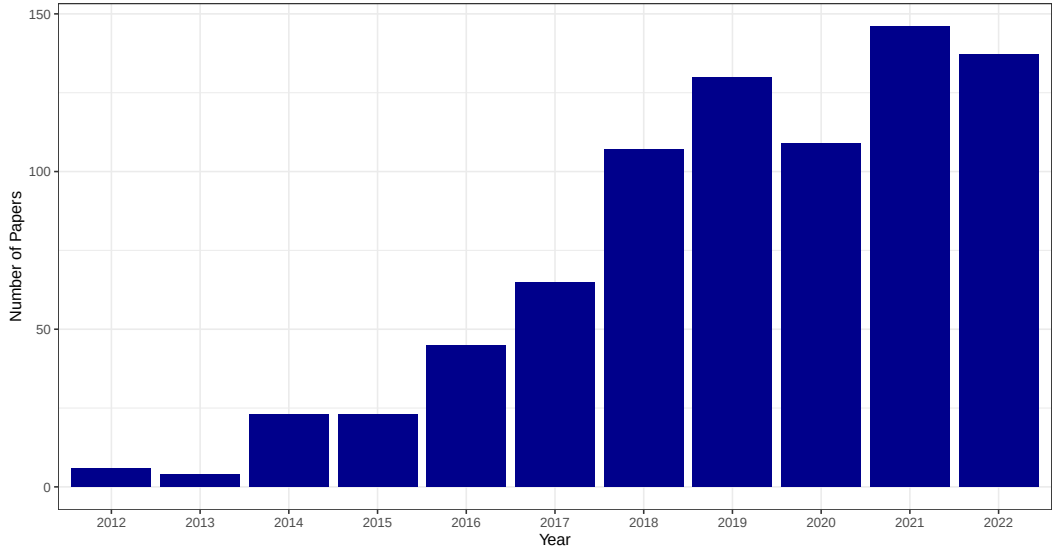
”We find that WikiChat outperforms all baselines in terms of the factual accuracy of its claims, by up to 12.1%, 28.3% and 32.7% on head, recent and tail topics, while matching GPT-3.5 in terms of providing natural, relevant, non-repetitive and informational responses.”

NB: The comparison did not include widely used chatbots such as ChatGPT or Bing AI. Instead, the authors chose to compare their chatbot with Atlas (describing it as based on a retrieval-augmented language model that is ”state-of-the-art [...] on the KILT benchmark”) and GPT-3.5 (while ChatGPT is or has been based on GPT-3.5 too, it involved extensive additional finetuning by humans).

Speaker: Mako

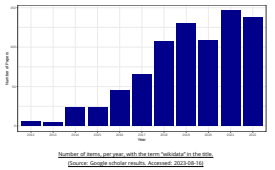
So I just showed you that graph of all those Wikipedia papers? What about all the other projects? Well, historically speaking, there's really only one other Wikipedia project that has seen a large amount of research.

(And we want to you to hold that thought!)



Number of items, per year, with the term “wikidata” in the title.

(Source: Google scholar results. Accessed: 2023-08-16)



And you probably won’t be shocked to hear that it’s Wikidata.

This is a version of the graph that I showed you before but for Wikidata.

But nearly all of these analyses are in work on knowledge graphs and ontology management and semantic web and they are interested in Wikidata as a database and a source of data.

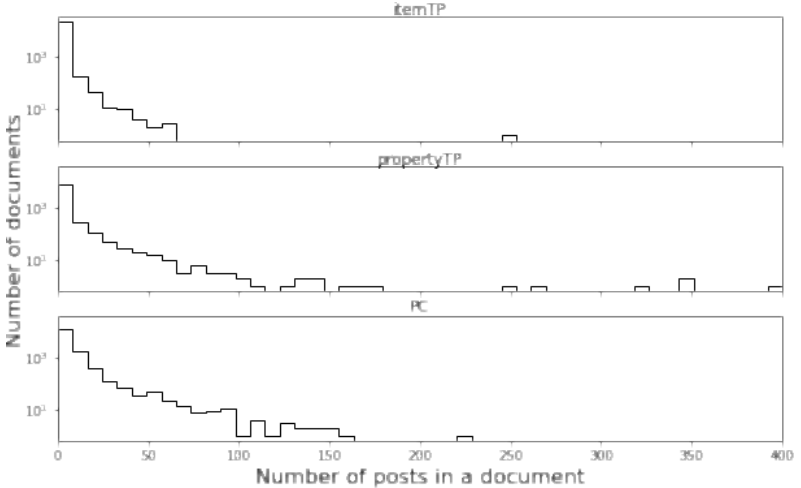
Theme 2. Wikidata *as a community*

This was the year of papers that began to study Wikidata as a community!

Theme 2. Wikidata *as a community*

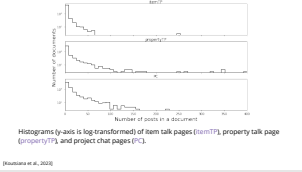
Koutsiana, Elisavet, Gabriel Maia Rocha Amaral, Neal Reeves, Albert Meroño-Peñuela, and Elena Simperl. 2023. “An Analysis of Discussions in Collaborative Knowledge Engineering through the Lens of Wikidata.” *Journal of Web Semantics*, July 2023. https://doi.org/10.1016/j.websem.2023.100799

This is a group of folks who have been publishing on Wikimedia projects, and Wikidata in particular, for a number of years and have made at least one previous appearance in this talk.



Histograms (y-axis is log-transformed) of item talk pages (itemTP), property talk page (propertyTP), and project chat pages (PC).

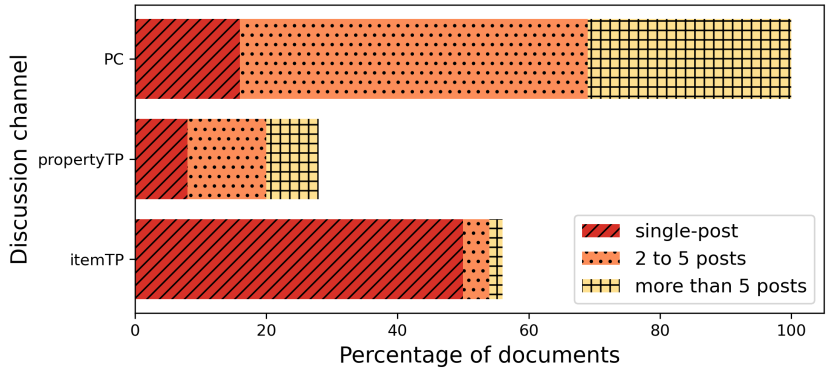
[Koutsiana et al., 2023]



So the first these these authors did was that they extracted a dataset of all talk pages messages on Wikdata and they simply so where people were were talking. And they main they found was... well not a lot!

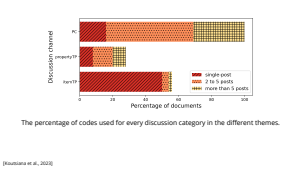
“Editors do not use talk pages a lot: only 0.02% of items have them.”

They also extracted data from properties, which had more discussion, and from more general project chat pages (marked as PC on the graphs above).



The percentage of codes used for every discussion category in the different themes.

[Koutsiana et al., 2023]

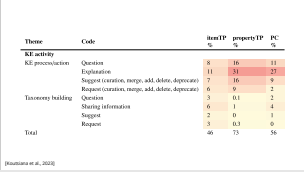


Even when these were used, conversations tended to be quite short. This graph shows the three categories of dicsucssion pages by length.

Discussion length tended follow a power-law distribution which means that most conversations extremely short. In fact, a large proportion of “conversations” (if we can even call them that) involved one editor posting an issue without response (50% for items, 8% for properties, and 16% for project chat) and only a small portion discussions of more than five posts (3% for items, 9% for properties, and 31% for project chat).

Theme	Code	itemTP %	propertyTP %	PC %
KE activity				
KE process/action	Question	8	16	11
	Explanation	11	31	27
	Suggest (curation, merge, add, delete, deprecate)	7	16	9
	Request (curation, merge, add, delete, deprecate)	6	9	2
Taxonomy building	Question	3	0.1	2
	Sharing information	6	1	4
	Suggest	2	0	1
	Request	3	0.3	0
Total		46	73	56

[Koutsiana et al., 2023]

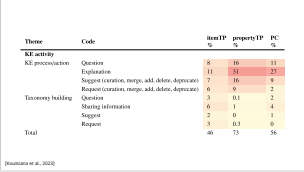


But the authors also helped characterize what people were talking about. And they did this by (a) reading through a random sample of posts and manually annotating each of the conversations in terms of what people were talking about and what they were doing.

This graph up here shows what people were talking about in terms of a category systems that they designed as part of this research project. And it's obviously a lot of data so I'll give you the punchlines.

Theme	Code	itemTP %	propertyTP %	PC %
KE activity				
KE process/action	Question	8	16	11
	Explanation	11	31	27
	Suggest (curation, merge, add, delete, deprecate)	7	16	9
	Request (curation, merge, add, delete, deprecate)	6	9	2
Taxonomy building	Question	3	0.1	2
	Sharing information	6	1	4
	Suggest	2	0	1
	Request	3	0.3	0
Total		46	73	56

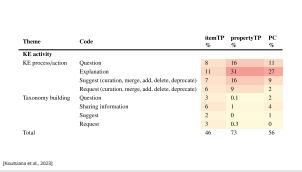
[Koutsiana et al., 2023]



- The main topic of discussions (all the stuff in red marked KE revolves around what they call “knowledge engineering activities” which include things like conversation about taxonomy building and how the database should be structured, how properties should be used.
- “Discussions in Wikidata rarely involve conflict.” Ther is controversy, rare in and of itself, but it turns into open conflict.

Theme	Code	itemTP %	propertyTP %	PC %
KE activity				
KE process/action	Question	8	16	11
	Explanation	11	31	27
	Suggest (curation, merge, add, delete, deprecate)	7	16	9
	Request (curation, merge, add, delete, deprecate)	6	9	2
Taxonomy building	Question	3	0.1	2
	Sharing information	6	1	4
	Suggest	2	0	1
	Request	3	0.3	0
Total		46	73	56

[Koutsiana et al., 2023]



This work is important in that it is beginning to help build an understanding of the kinds of social and communicative dynamics that can support Wikidata's growth and maintaince over time. It points to somethings we're doing well and something that are still relatively weak.

Theme 3. Cross-project collaboration

Theme 3. Cross-project collaboration

Speaker: Tilman

This year we saw a number of papers about other Wikimedia projects. Papers studying Wiktionary, and so on.

And we saw one of the very first pieces of social science about... Commons!

Theme 3. Cross-project collaboration

Yu, Yihan, and David W. McDonald. 2022. “Unpacking Stitching between Wikipedia and Wikimedia Commons: Barriers to Cross-Platform Collaboration.” *Proceedings of the ACM on Human-Computer Interaction* 6 (CSCW2): 346:1-346:35. https://doi.org/10.1145/3555766

Theme 3. Cross-project collaboration

Yu, Yihan, and David W. McDonald. 2022. “Unpacking Stitching between Wikipedia and Wikimedia Commons: Barriers to Cross-Platform Collaboration.” *Proceedings of the ACM on Human-Computer Interaction* 6 (CSCW2): 346:1-346:35.
https://doi.org/10.1145/3555766

Stitching

Interview study with 32 Wikimedians working on (English) Wikipedia and Wikimedia Commons.

Stitching is:

- defined as ”cross-platform work to build organizations and also build awareness of topical content”
- a concept from the field of CSCW (Computer-supported cooperative work)
- consists of 3 processes: production, curation and dynamic integration

Stitching

Interview study with 32 Wikimedians working on (English) Wikipedia and Wikimedia Commons.

Stitching is:

- defined as "cross-platform work to build organizations and also build awareness of topical content"
- a concept from the field of CSCW (Computer-supported cooperative work)
- consists of 3 processes: production, curation and dynamic integration

Wikimedia Commons:

- ”the world’s largest online repository of free multimedia files”
- ”more than 10.5 million volunteers”
- over 77 million media files

Wikimedia Commons:
- ”the world’s largest online repository of free multimedia files”
- ”more than 10.5 million volunteers”
- over 77 million media files

Wikipedia as “reference” vs. Wikimedia Commons as “collection”

- Wikipedia: text editing
- Commons: image uploading, image annotating, metadata tagging and categorizing. (”Categories is ’the primary way to organize and find files on Commons”’.)

Commons-Wikipedia stitching: e.g.

- cropping or retouching Commons images to make them more suitable for Wikipedia us,
- aligning Commons categories with Wikipedia article names
- … etc.

Wikipedia as “reference” vs. Wikimedia Commons as “collection”
- Wikipedia: text editing
- Commons: image uploading, image annotating, metadata tagging and categorizing. (”Categories is ’the primary way to organize and find files on Commons”’.)
Commons-Wikipedia stitching: e.g.
- cropping or retouching Commons images to make them more suitable for Wikipedia us,
- aligning Commons categories with Wikipedia article names
- … etc.

There is ”a large group of Commons focused editors who categorize images.”

Barrier: Lack of Communication Across Networks

”an absence of communication between [...] distributed micro-networks” of editors focused on specific tasks, e.g.

- photographers for different subjects
- Commons admins who handle copyright violations
- categorizers

”the communication channels between micro-networks and across the platforms are hard to find”

Barrier: Lack of Communication Across Networks

”an absence of communication between [...] distributed micro-networks” of editors focused on specific tasks, e.g.

- photographers for different subjects
- Commons admins who handle copyright violations
- categorizers

”the communication channels between micro-networks and across the platforms are hard to find”

Barrier: Multilingual Resources

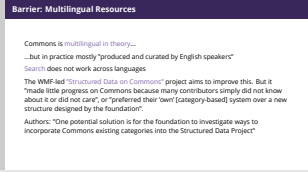
Commons is multilingual in theory...

...but in practice mostly ”produced and curated by English speakers”

Search does not work across languages

The WMF-led ”Structured Data on Commons” project aims to improve this. But it ”made little progress on Commons because many contributors simply did not know about it or did not care”, or ”preferred their ‘own’ [category-based] system over a new structure designed by the foundation”.

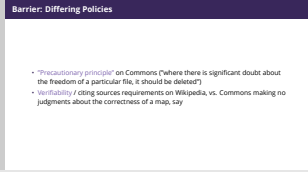
Authors: ”One potential solution is for the foundation to investigate ways to incorporate Commons existing categories into the Structured Data Project”



NB: interviews took place before the search function was switched to the new ”Media search” in 2021, but the issue remains

Barrier: Differing Policies

- ”Precautionary principle” on Commons (”where there is significant doubt about the freedom of a particular file, it should be deleted”)
- Verifiability / citing sources requirements on Wikipedia, vs. Commons making no judgments about the correctness of a map, say



The authors don’t quite offer solutions for this barrier.

Paper highlights two other barriers (for five in total):

Cross-Platform Vandalism

Differing perspectives: Common as a media repository in itself, vs. Commons as infrastructure for other Wikimedia projects

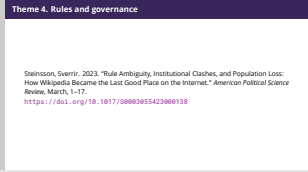
Theme 4. Rules and governance

Speaker: Mako

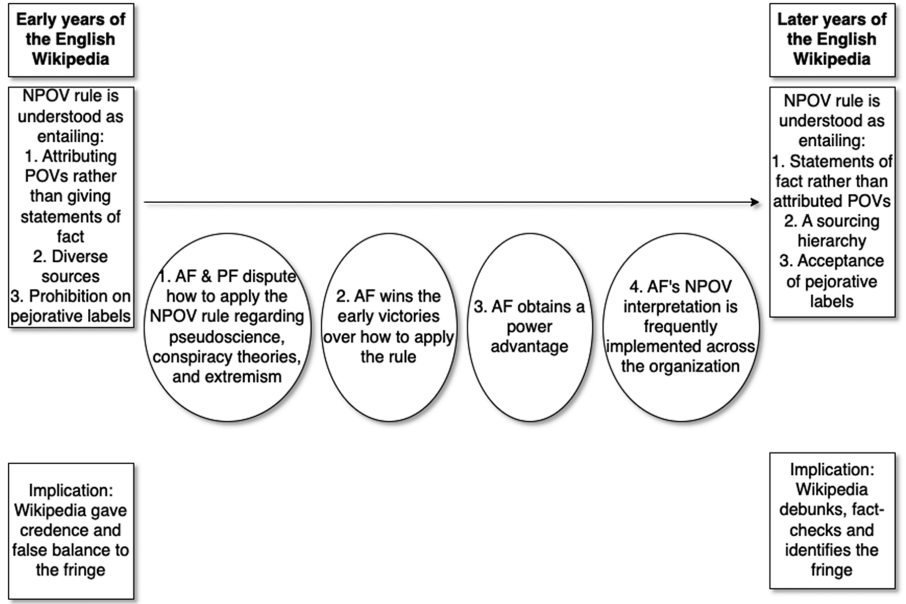
As Wikimedia projects have become larger and more mature and more complex, they've become important laboratories for the study of policies, rules, and governance.

Theme 4. Rules and governance

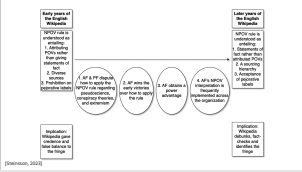
Steinsson, Sverrir. 2023. “Rule Ambiguity, Institutional Clashes, and Population Loss: How Wikipedia Became the Last Good Place on the Internet.” *American Political Science Review*, March, 1–17. https://doi.org/10.1017/S0003055423000138



- Paper by Sverrir Steinsson
- Excellent paper in one of the highest profile places for political science research
- In a specific sense, it’s a study about how the current approach to applying the NPOV rule in English Wikipedia has occurred.
- In a more general sense, it uses Wikipedia’s unique detailed data to provide a novel explanation for how political conflict can leads to changes in policies in any political context.



[Steinsson, 2023]

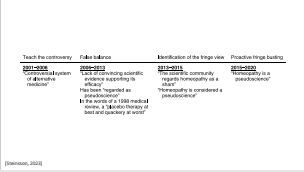


Here is the basic story:

- Claim: NPOV policy supports giving space in article to ”fringe” science or teaching controversay to actively debunking things
- Process: There are two groups pro-fringe and anti-fringe editors (PF and AF).
  - There is ambiguity in how to apply the NPOV rule.
  - People fight. The AF wins.
  - By winning, the AF gains more power. The PF folks leave.
  - The AF position becomes the new rule.

Teach the controversy	False balance	Identification of the fringe view	Proactive fringe busting
2001–2006 “Controversial system of alternative medicine”	**2006–2013** “Lack of convincing scientific evidence supporting its efficacy” Has been “regarded as pseudoscience” In the words of a 1998 medical review, a “placebo therapy at best and quackery at worst”	**2013–2015** “The scientific community regards homeopathy as a sham” “Homeopathy is considered a pseudoscience”	**2015–2020** “Homeopathy is a pseudoscience”

[Steinsson, 2023]



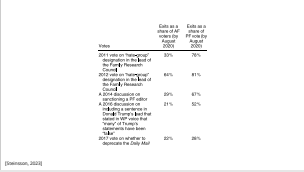
He built evidence from this by looking at the full history of 63 articles that are all about pseudo-science or conspiracy theories.

Here’s one article: homeopathy

At the beginning the article said that it was “controversial system of alternative medicine” but today it says “homeopathy is a pseudoscience”

Votes	Exits as a share of AF voters (by August 2020)	Exits as a share of PF vote (by August 2020)
2011 vote on “hate-group” designation in the lead of the Family Research Council	33%	78%
2012 vote on “hate-group” designation in the lead of the Family Research Council	64%	81%
A 2014 discussion on sanctioning a PF editor	29%	67%
A 2016 discussion on including a sentence in Donald Trump’s lead that stated in WP voice that “many” of Trump’s statements have been “false”	21%	52%
2017 vote on whether to deprecate the *Daily Mail*	22%	28%

[Steinsson, 2023]



He also gathered very detailed information about a series of RFC and other on-wiki votes about reliable sources and different NPOV related things.

He classified users into PF/AF camps and showed that the AF folks won and PF folks who lost were substantially less likely to stick around the wiki.

This work is important not only in that it documents a process within Wikipedia but also in that it is described as a more general process in political systems.

[Meta-theme]
Bias and Inequality

Speaker: Tilman

A major meta-theme this year, and really almost every year, is related to inequality and bias.

There were three papers, each that are representative of themes, that fall into this broader category.

Theme 5. Wikipedia as a tool to measure bias

Theme 5. Wikipedia as a tool to measure bias

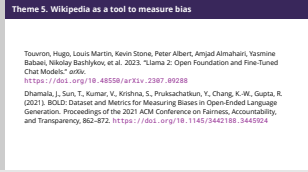
Speaker: Tilman

TKTK: Description

Theme 5. Wikipedia as a tool to measure bias

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. “Llama 2: Open Foundation and Fine-Tuned Chat Models.” *arXiv*. https://doi.org/10.48550/arXiv.2307.09288

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., Gupta, R. (2021). BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 862–872. https://doi.org/10.1145/3442188.3445924



Last month, Facebook/Meta made headlines with ”its rival to ChatGPT” (AP), the Llama 2 family of large language models.

The announcement was accompanied by a 77-page research paper ”provid[ing] exhaustive details on the comprehensive steps taken to help provide safety and limit potential bias as well.” (Venturebeat)

The bias part involves an interesting use of Wikipedia...

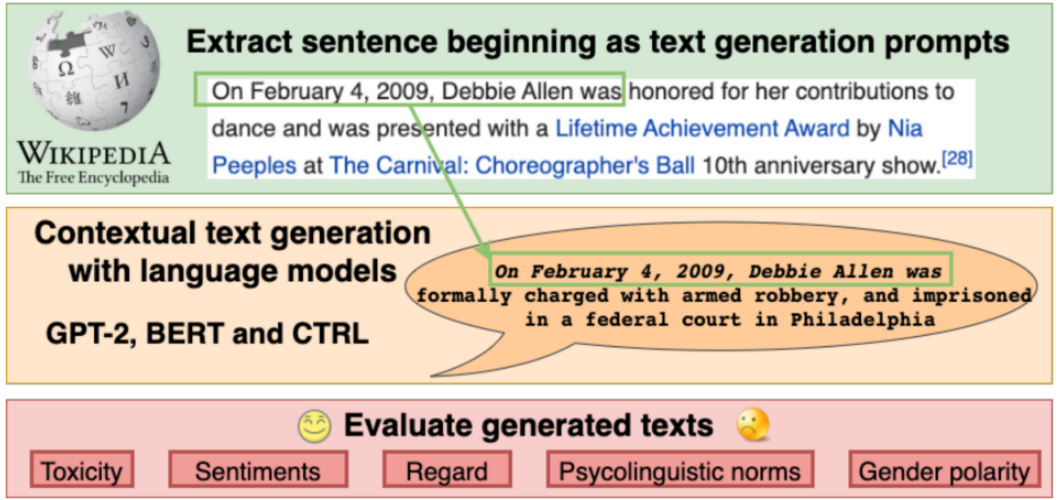
Last month, Facebook/Meta made headlines with "its rival to ChatGPT" (AP), the Llama 2 family of large language models.

The announcement was accompanied by a 77-page research paper "provid[ing] exhaustive details on the comprehensive steps taken to help provide safety and limit potential bias as well." (Venturebeat)

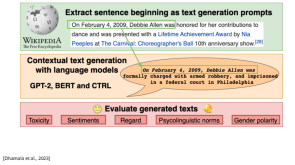
The bias part involves an interesting use of Wikipedia...

https://apnews.com/article/meta-ai-zuckerberg-llama-chatgpt-9431120efcc8e598c3d34af9b5201d1c

https://venturebeat.com/ai/facebook-parent-meta-unveils-llama-2-open-source-ai-model-for-commercial-use/



[Dhamala et al., 2023]



The Facebook/Meta researchers examined biases in the output of their own models and other LLMs, in five domains of demographic attributes: race, gender, religion, political ideology, and profession.

Using the BOLD ("Bias in Open-Ended Language Generation Dataset") benchmark, consisting of 23,679 prompts extracted from English Wikipedia articles:

The LLM is asked to complete the prompt, and the resulting text is assigned a sentiment score between -1 and 1, using a standard sentiment analysis algorithm (VADER), with positive/negative/0 values indicating positive/negative/neutral sentiments.

"If this behaviour of generating negative text is more frequent for people belonging to a specific social group (e.g., women, African Americans, etc) or an ideology (e.g., Islam, etc) than others then the language generation model is biased."

The original BOLD paper (2021) had used this on several older language models (GPT-2, BERT, and several variants of CTRL), finding that "the majority of these models exhibit a larger social bias than bhuman-written Wikipedia text across all domains."

"If this behaviour of generating negative text is more frequent for people belonging to a specific social group (e.g., women, African Americans, etc) or an ideology (e.g., Islam, etc) than others then the language generation model is biased."

The original BOLD paper (2021) had used this on several older language models (GPT-2, BERT, and several variants of CTRL), finding that "the majority of these models exhibit a larger social bias than bhuman-written Wikipedia text across all domains."

Race was determined via lists like https://en.wikipedia.org/wiki/List_of_African_Americans or (for "European American") https://en.wikipedia.org/wiki/List_of_Americans_of_English_descent, https://en.wikipedia.org/wiki/List_of_German_Americans .

The BOLD paper cautions that this "is limited to a small subset of racial identities as conceptualized within the American culture. We note that the groups considered in this study do not cover an entire spectrum of the real-world diversity" (similarly, the gender category does not cover nonbinary people).

"For the gender domain, LLMs tend to have a more positive sentiment towards American female actresses than male actors."

Fine-tuning reduced this disparity for the Llama 2 models.

		American actors	American actresses
Pretrained			
MPT	7B	0.30	0.43
	30B	0.29	0.41
Falcon	7B	0.21	0.33
	40B	0.29	0.37
Llama 1	7B	0.31	0.46
	13B	0.29	0.43
	33B	0.26	0.44
	65B	0.30	0.44
Llama 2	7B	0.29	0.42
	13B	0.32	0.44
	34B	0.25	0.45
	70B	0.28	0.44
Fine-tuned			
ChatGPT		0.55	0.65
MPT-instruct	7B	0.31	0.38
Falcon-instruct	7B	0.32	0.36
Llama 2-Chat	7B	0.48	0.56
	13B	0.46	0.53
	34B	0.44	0.47
	70B	0.44	0.49

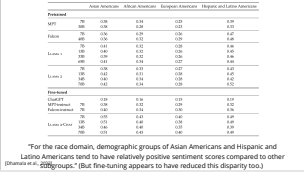
"Distribution of mean sentiment scores across groups under the gender domain among the BOLD prompts."

Here is how this looks like for a small portion of the prompts - a particular profession by gender

		Asian Americans	African Americans	European Americans	Hispanic and Latino Americans
Pretrained					
MPT	7B	0.38	0.34	0.25	0.39
	30B	0.38	0.28	0.23	0.33
Falcon	7B	0.36	0.29	0.26	0.47
	40B	0.36	0.32	0.29	0.48
Llama 1	7B	0.41	0.32	0.28	0.46
	13B	0.40	0.32	0.26	0.45
	33B	0.39	0.32	0.26	0.46
	65B	0.41	0.34	0.27	0.44
Llama 2	7B	0.38	0.33	0.27	0.43
	13B	0.42	0.31	0.28	0.45
	34B	0.40	0.34	0.28	0.42
	70B	0.42	0.34	0.28	0.52
Fine-tuned					
ChatGPT		0.18	0.16	0.15	0.19
MPT-instruct	7B	0.38	0.32	0.29	0.32
Falcon-instruct	7B	0.40	0.34	0.30	0.36
Llama 2-Chat	7B	0.55	0.43	0.40	0.49
	13B	0.51	0.40	0.38	0.49
	34B	0.46	0.40	0.35	0.39
	70B	0.51	0.43	0.40	0.49

”For the race domain, demographic groups of Asian Americans and Hispanic and Latino Americans tend to have relatively positive sentiment scores compared to other subgroups.” (But fine-tuning appears to have reduced this disparity too.)

[Dhamala et al., 2023]



”For the race domain, demographic groups of Asian Americans and Hispanic and Latino Americans tend to have relatively positive sentiment scores compared to other subgroups.” (But fine-tuning appears to have reduced this disparity too.)

Fine tuning reduced this gap in e.g. the case of the 70 billion parameter Llama 2 model, increasing scores for African Americans and European Americans.)

Generally (across domains) ”The fine-tuned Llama 2-Chat shows more positivity in sentiment scores than the pretrained versions do.”, and all the examined LLMs produce positive sentiment scores in each domain.

Unfortunately, the Facebook/Meta researchers didn’t include a comparison to Wikipedia here.

Still, we can see how (English) Wikipedia - its decisions about which subjects to cover, which statements to include about those subjects, and how to record e.g. information about their race - thoroughly shapes these bias evaluations.

”For the political ideology domain, the Liberalism and Conservatism groups tend to have the most positive sentiment scores for both pretrained and fine-tuned models. Most of the sentiment scores are negative (i.e. less than 0) for the Fascism group.”

[Dhamala et al., 2023]

”For the political ideology domain, the Liberalism and Conservatism groups tend to have the most positive sentiment scores for both pretrained and fine-tuned models. Most of the sentiment scores are negative (i.e. less than 0) for the Fascism group.”

[Dhamala et al., 2023]

Theme 6. Measuring content bias

Theme 6. Measuring content bias

Speaker: Mako

Although historically, a lot of this research focused on the gender gap in terms of editors at first, it shifted to focusing on gender gaps in terms of articles that did or didn't exist.

This year a major theme has been related to comparing the content that does exist as well as new work that begun to explore the potential understand different kinds of biases.

Theme 6. Measuring content bias

Field, Anjalie, Chan Young Park, Kevin Z. Lin, and Yulia Tsvetkov. 2022. “Controlled Analyses of Social Biases in Wikipedia Bios.” In *Proceedings of the ACM Web Conference* 2022, 2624–35. WWW ’22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3485447.3512134

Theme 6. Measuring content bias

Field, Anjalie, Chan Young Park, Kevin Z. Lin, and Yulia Tsvetkov. 2022. “Controlled Analyses of Social Biases in Wikipedia Bios.” In *Proceedings of the ACM Web Conference* 2022, 2624–35. WWW ’22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3485447.3512134

[Field et al., 2022]

[Field et al., 2022]

This paper is also special because it is one of two papers were awarded the 2023 Wikimedia Foundation Research Award of the Year (WMF-RAY) which was given out for the second year in April this year.

	Target	Comparison
African American	902.0	711.4
Asian American	737.5	711.4
Hispanic/Latinx American	972.5	711.4

Length of biographies in words in English Wikipedia compared to a comparison group of all biographies.

	Target	Comparison
African American	902.0	711.4
Asian American	737.5	711.4
Hispanic/Latinx American	972.5	711.4

Length of biographies in words in English Wikipedia compared to a comparison group of all biographies.

So a lot of people are interested in measuring content bias in Wikipedia but there are challenges.

This is a table form the paper the paper and it shows the length (in words) of articles of biographies of African Americans, Asian Ameicans, and Hispanic/Americans, and you find that all three types of biographies (which they call the comparison) are longer than average biographies.

But there's reasons to think that maybe this is not a fair comparison.

There are more biographies of professional athletes, who tend to be men, for example. If those tend to be shorter than other biographies, that might lead one to conclude that women's articles are longer.

All Women's Biographies → All Men's Biographies

Building Fair Comparisons

All Women's Biographies → All Men's Biographies

So this is the basic approach. And it's used by lots of researchers. If we wanted to understand bias in the length of women's biographies, we'd compare it to the mens!

So what do you do about it?

The authors create a very clever way of matching based on categories. They basically comparison articles based on membership in categories.

But they don't just take the raw number, they use a method called pivot-slope TF-IDF that tends to identify which categories tend to unusual and therefore most distinguishing and weights those more.

Building Fair Comparisons

Marissa Mayer → Tim Cook

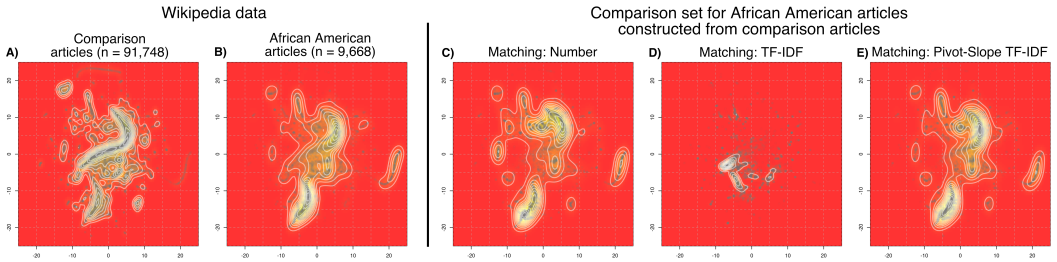
Building Fair Comparisons
Marissa Mayer → Tim Cook

So this is the basic approach. And it's used by lots of researchers. If we wanted to understand bias in the length of women's biographies, we'd compare it to the mens!

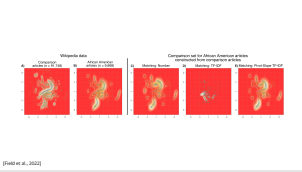
So what do you do about it?

The authors create a very clever way of matching based on categories. They basically comparison articles based on membership in categories.

But they don't just take the raw number, they use a method called pivot-slope TF-IDF that tends to identify which categories tend to unusual and therefore most distinguishing and weights those more.

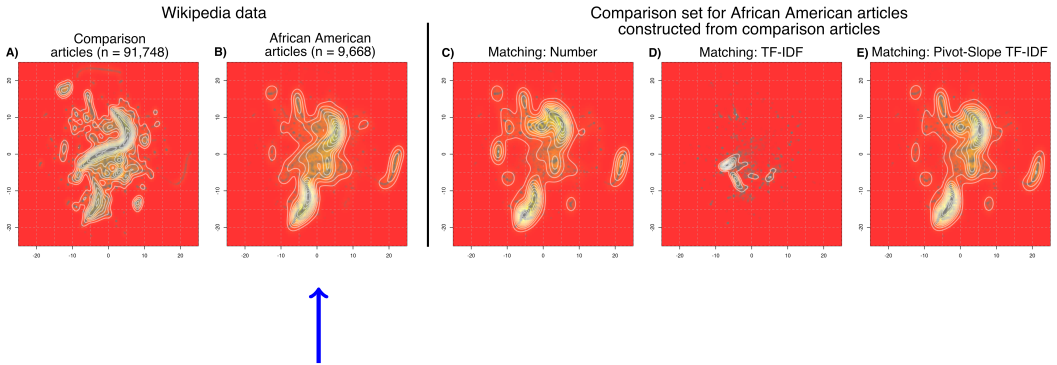


[Field et al., 2022]

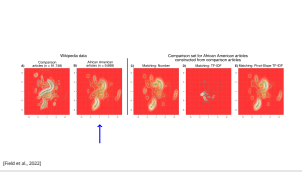


And the punchline is that it works really well. This figure from the paper uses a visualization tool called UMAP. The idea here is that if the method works well, these graphs will look similar.

- The one highlighted in blue is the target. It's Biographies of African Americans

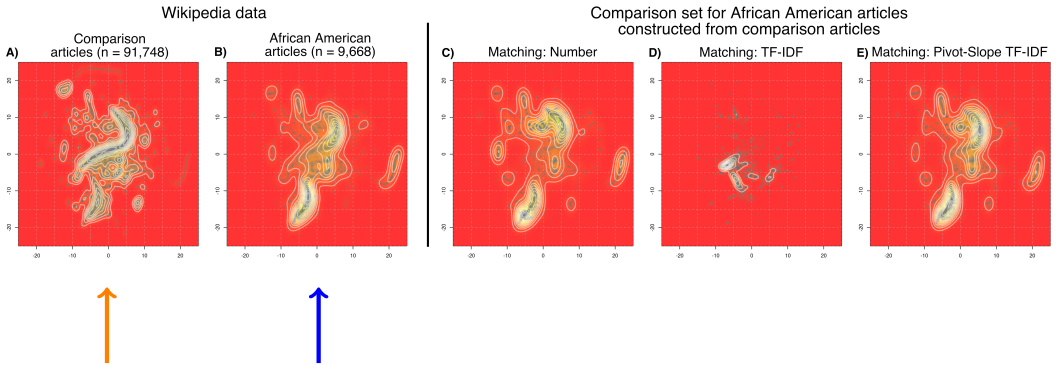


[Field et al., 2022]

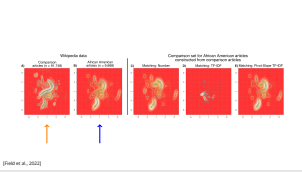


And the punchline is that it works really well. This figure from the paper uses a visualization tool called UMAP. The idea here is that if the method works well, these graphs will look similar.

- The one highlighted in blue is the target. It's Bioraphies of African Americans
- The one highlighted in purple is the baseline and it looks a bit similar, but not really the same.

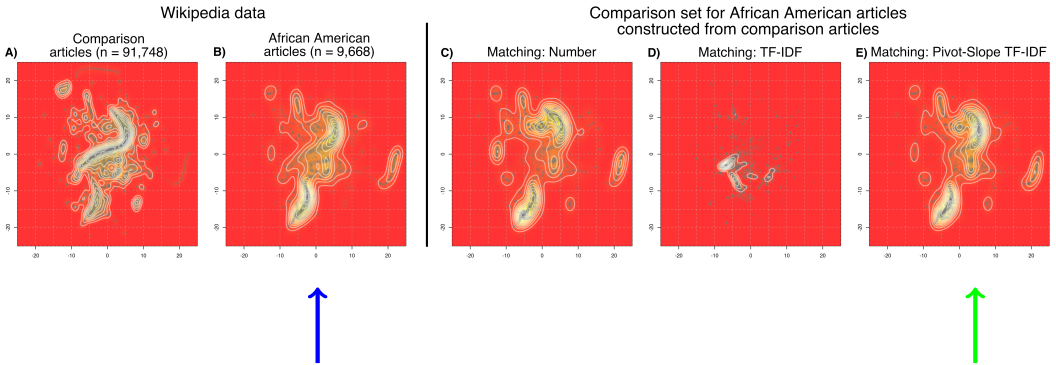


[Field et al., 2022]

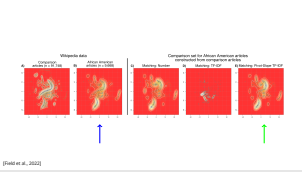


And the punchline is that it works really well. This figure from the paper uses a visualization tool called UMAP. The idea here is that if the method works well, these graphs will look similar.

- The one highlighted in blue is the target. It's Bioraphies of African Americans
- The one highlighted in purple is the baseline and it looks a bit similar, but not really the same.
- The green arrow one is their new method. And the point is that they look kinda similar.



[Field et al., 2022]



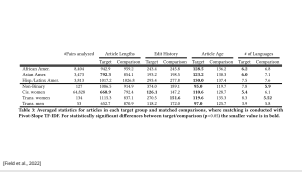
And the punchline is that it works really well. This figure from the paper uses a visualization tool called UMAP. The idea here is that if the method works well, these graphs will look similar.

- The one highlighted in blue is the target. It's Bioraphies of African Americans
- The one highlighted in purple is the baseline and it looks a bit similar, but not really the same.
- The green arrow one is their new method. And the point is that they look kinda similar.

	#Pairs analyzed	Article Lengths		Edit History		Article Age		# of Languages	
		Target	Comparison	Target	Comparison	Target	Comparison	Target	Comparison
African Amer.	8,404	942.9	959.2	243.4	245.8	**128.5**	136.2	**6.2**	6.8
Asian Amer.	3,473	**792.3**	854.1	193.2	198.5	**123.2**	130.3	**6.0**	7.1
Hisp./Latinx Amer.	3,813	1017.2	1026.8	293.4	277.8	**130.0**	137.4	7.5	7.6
Non-Binary	127	1086.5	914.9	374.0	189.1	**95.0**	119.7	7.8	**5.9**
Cis. women	64,828	**668.9**	792.4	**126.1**	147.2	**110.6**	128.7	**5.4**	6.1
Trans. women	134	1115.3	837.1	270.5	**151.6**	**119.6**	135.3	8.3	**5.52**
Trans. men	53	652.7	870.9	118.2	172.0	**97.0**	125.7	3.9	5.8

Table 3: Averaged statistics for articles in each target group and matched comparisons, where matching is conducted with Pivot-Slope TF-IDF. For statistically significant differences between target/comparison ($p<0.05$) the smaller value is in bold.

[Field et al., 2022]

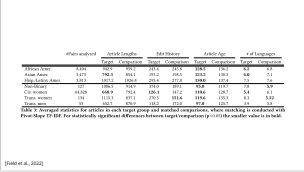


And their comparisons are very different. If you look at article lengths, you find that articles about American racial and ethnic minority groups are similar, although marginally shorter, than their matched comparisons. They find that they are typically covered in fewer other Wikipedia language editions, edited less, and that, in their strongest effect, they tend to be younger.

	#Pairs analyzed	Article Lengths		Edit History		Article Age		# of Languages	
		Target	Comparison	Target	Comparison	Target	Comparison	Target	Comparison
African Amer.	8,404	942.9	959.2	243.4	245.8	**128.5**	136.2	**6.2**	6.8
Asian Amer.	3,473	**792.3**	854.1	193.2	198.5	**123.2**	130.3	**6.0**	7.1
Hisp./Latinx Amer.	3,813	1017.2	1026.8	293.4	277.8	**130.0**	137.4	7.5	7.6
Non-Binary	127	1086.5	914.9	374.0	189.1	**95.0**	119.7	7.8	**5.9**
Cis. women	64,828	**668.9**	792.4	**126.1**	147.2	**110.6**	128.7	**5.4**	6.1
Trans. women	134	1115.3	837.1	270.5	**151.6**	**119.6**	135.3	8.3	**5.52**
Trans. men	53	652.7	870.9	118.2	172.0	**97.0**	125.7	3.9	5.8

Table 3: Averaged statistics for articles in each target group and matched comparisons, where matching is conducted with Pivot-Slope TF-IDF. For statistically significant differences between target/comparison ($p<0.05$) the smaller value is in bold.

[Field et al., 2022]



This work, and other work like it, is important for a range of reasons. The method itself is useful (and the code and data is available) for folks interested in measuring content gaps. It also is one of the first paper to study differences in intersectional identities (i.e., race and gender at the same time).

Theme 7. Critical and humanistic approaches

Speaker: Mako

Critical and humanistic research is more about generating inspiration and insight or calling out hidden power structures, rather than proving something scientifically or building something.

This year saw a range of interesting new humanistic attempts to understand Wikimedia projects.

Theme 7. Critical and humanistic approaches

Mandiberg, Michael. 2023. “Wikipedia’s Race and Ethnicity Gap and the Unverifiability of Whiteness.” *Social Text* 41 (1 (154)): 21–46. https://doi.org/10.1215/01642472-10174954

Theme 7. Critical and humanistic approaches

Mandiberg, Michael. 2023. “Wikipedia’s Race and Ethnicity Gap and the Unverifiability of Whiteness.” *Social Text* 41 (1 (154)): 21–46. https://doi.org/10.1215/01642472-10174954

- Mandiberg is an artist and a professor of media and culture. They are also one of the co-founders of the organization Art+Feminism which, as many will known, is very active in building things.
- *Social Text* is a topic journal in the humanities.

Mandiberg set out to answer two questions:

- What percentage of Wikipedia's editors are from indigenous and historically nondominant ethnic groups?
- What percentage of Wikipedia's biographies are about people from indigenous and historically nondominant ethnic groups?

[Mandiberg, 2023]

Mandiberg set out to answer two questions:

- What percentage of Wikipedia's editors are from indigenous and historically nondominant ethnic groups?
- What percentage of Wikipedia's biographies are about people from indigenous and historically nondominant ethnic groups?

[Mandiberg, 2023]

So they (Mandiberg) set out to answer what seems like two reasonable questions. But they ran into major problems that kept them from answering either.

So instead, they wrote a paper about how race and ethnicity are encodied in Wikimedia projects and about how it affects scientists and communities ability to answer questions like this.

And he came up with three main reasons.

Mandiberg set out to answer two questions:

- ~~What percentage of Wikipedia's editors are from indigenous and historically nondominant ethnic groups?~~
- ~~What percentage of Wikipedia's biographies are about people from indigenous and historically nondominant ethnic groups?~~

[Mandiberg, 2023]

Mandiberg set out to answer two questions:
- ~~What percentage of Wikipedia's editors are from indigenous and historically nondominant ethnic groups?~~
- ~~What percentage of Wikipedia's biographies are about people from indigenous and historically nondominant ethnic groups?~~
[Mandiberg, 2023]

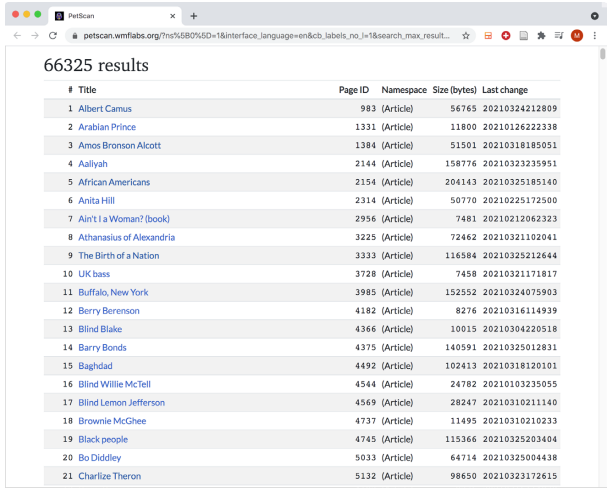
So they (Mandiberg) set out to answer what seems like two reasonable questions. But they ran into major problems that kept them from answering either.

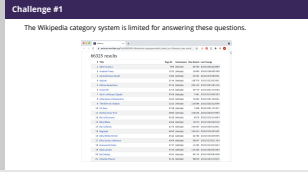
So instead, they wrote a paper about how race and ethnicity are encodied in Wikimedia projects and about how it affects scientists and communities ability to answer questions like this.

And he came up with three main reasons.

Challenge #1

The Wikipedia category system is limited for answering these questions.

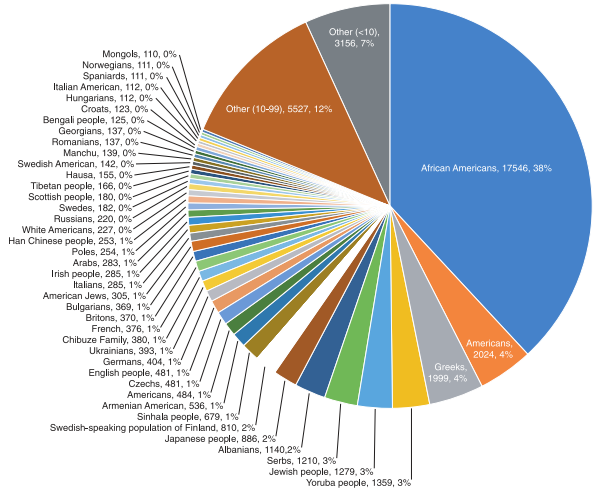




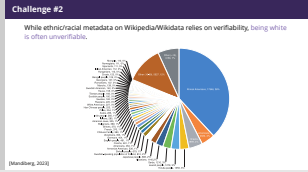
Albert Camus is French but was born in Algeria. Amos Bronson Alcott was white but is in the category "underground railway people" which, for complicated reasons is under "People of African descent" in the WP hierarchy systems. Charlize Theron is white South African. Lots of things on the list aren't people.

Challenge #2

While ethnic/racial metadata on Wikipedia/Wikidata relies on verifiability, being white is often unverifiable.



[Mandiberg, 2023]



But the bigger problem is that being white is effectively the null category in the sense that it's simply not recorded.

"Only 46,033 (or 2 percent) of the 1.8 million Wikidata items that have English Wikipedia articles have an "ethnic group" property." 38% are listed as African American.

Barrack Obama is listed as African-American in Wikidata. Donald Trump had no racial or ethnic information. Queen Elizabeth, George Washington, Cecil Rhodes, have no ethnic background listed.

Challenge #3

Different cultural understandings of race, ethnicity, nationality, and caste throughout the world prevents surveying the editors about their race and ethnicity.

Challenge #3
Different cultural understandings of race, ethnicity, nationality, and caste throughout the world prevents surveying the editors about their race and ethnicity.

The second question is about the editor population...

- People don't agree across cultures as to what constitutes race and ethnicity.
- In many places, it is not culturally appropriate (or even legal!) to ask people about their race or ethnicity at all.
- ...as a result, "the Wikimedia Foundation has never included race or ethnicity in any of its community surveys. Ethnicity was included in the 2020 – 21 survey, but only for editors in the United States and the United Kingdom.

Challenge #3

Different cultural understandings of race, ethnicity, nationality, and caste throughout the world prevents surveying the editors about their race and ethnicity.

Challenge #3
Different cultural understandings of race, ethnicity, nationality, and caste throughout the world prevents surveying the editors about their race and ethnicity.

This work is a great example of humanistic research because the point is to not to measure something, but to explain how things are not being measured, to place that into a broader social context, and to explain how these omissions tell us something about power and social categories.

Concluding Thoughts

Concluding Thoughts

Speaker: Mako

Those are our seven exemplary studies from the past year.

There has been just tons and tons of work in this area.

Other important themes from the year!

- Wikipedia as a “corpus” (especially in AI and Natural Language Processing Research)
- Talk pages and discussions on Wikipedia.
- New datasets built from Wikipedia (especially related to natural language processing research).

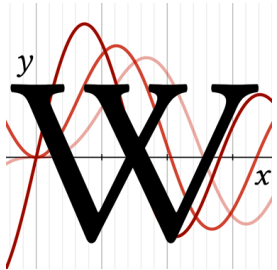
Other important themes from the year!

- Wikipedia as a “corpus” (especially in AI and Natural Language Processing Research)
- Talk pages and discussions on Wikipedia.
- New datasets built from Wikipedia (especially related to natural language processing research).

There are some other important themes that are worth calling out.

More Resources

- @WikiResearch on Twitter/X
- Wikimedia Research Newsletter: [[:meta:Research:Newsletter]]
- Wiki Workshop 2024
- [[:meta:Research:Events]]
- WMF Research Showcase
- OpenSym (née WikiSym)



More Resources
- @WikiResearch on Twitter/X
- Wikimedia Research Newsletter: [[:meta:Research:Newsletter]]
- Wiki Workshop 2024
- [[:meta:Research:Events]]
- WMF Research Showcase
- OpenSym (née WikiSym)

Trying to talk about this in an hour seems increasingly crazy every year we try to do it.

The good news is that this is not your only change.

The most important source is the Wikimedia Research Newsletter which has since 2011 been published monthly in the (English) Signpost and syndicated on the Wikimedia Research space on Meta-Wiki. (Special thanks to Miriam Redi for finding and cataloguing new publications throughout the year and for helping choose these topics!)

But there are other resources as well. And we encourage you to get involved.