The State of Wikimedia Research: 2015-2016

Benjamin Mako Hill
Tilman Bayer
Wikimania 2016, Esino Lario
June 24, 2016

I've been doing this for many years. I started in 2008 and have done this almost every single year since.

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

"This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – 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 my Wikimania 2008 Submission**

"This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – 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 my Wikimania 2008 Submission

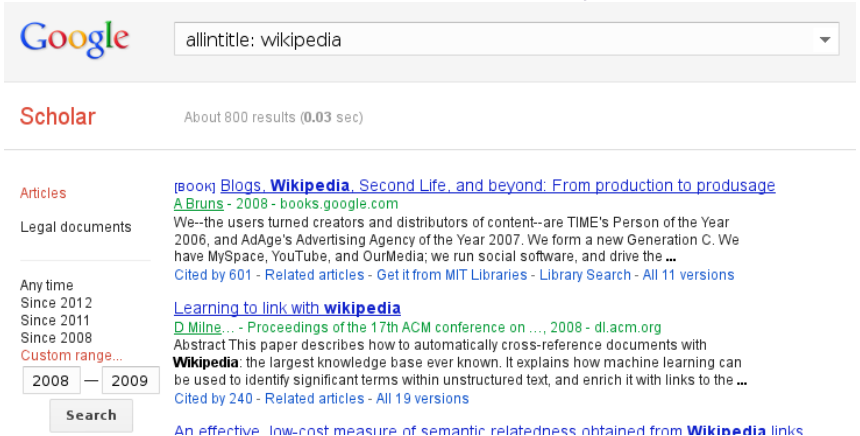
Back in Wikimania 2008, I set out to run a session at Wikimania that would provide a comprehensive literature review of articles in Wikipedia published in the last year.

> *"This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – 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 my 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 – a literature review in the scholarly parlance – 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 my Wikimania 2008 Submission

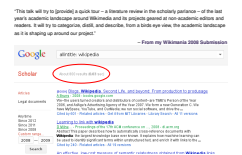


I tried to import the whole list into Zotero and managed to get banned for abusing the Google Scholar because they thought that no human being could realistically consume the amount of material published on Wikipedia that year.

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

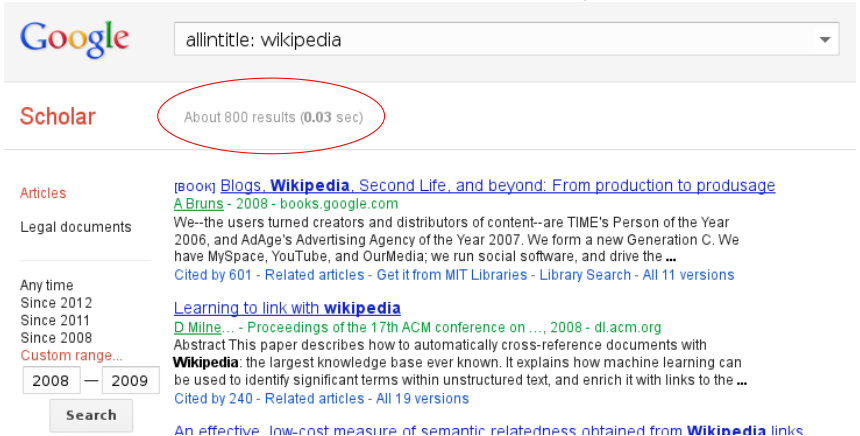
And believe it or not, this year is even bigger.

And my talk is even shorter.



"This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – 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 my Wikimania 2008 Submission

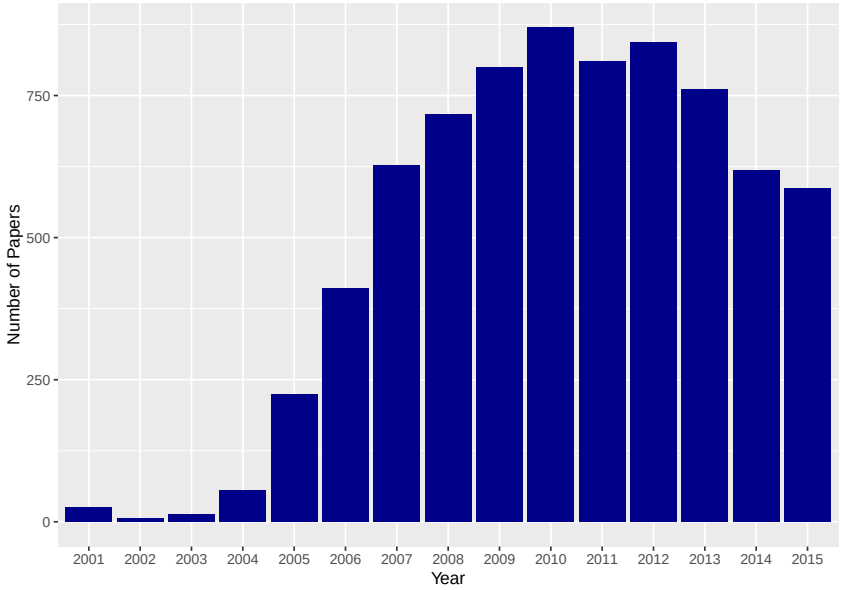


I tried to import the whole list into Zotero and managed to get banned for abusing the Google Scholar because they thought that no human being could realistically consume the amount of material published on Wikipedia that year.

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

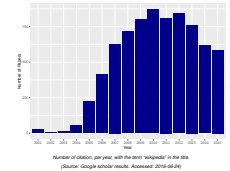
And believe it or not, this year is even bigger.

And my talk is even shorter.



Number of citation, per year, with the term "wikipedia" in the title.

(Source: Google scholar results. Accessed: 2016-06-24)

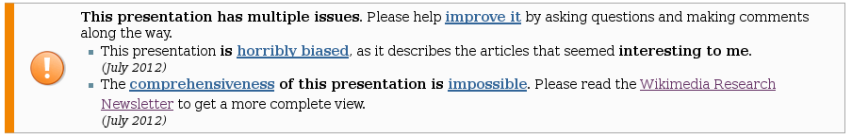


Academics have written **a lot** of papers about Wikipedia. There are more than 500 papers published about Wikipedia each year and although we've reached and moved past a peak it seems, it's not slowing by much.

- **6,037** Wikipedia-related publications in the Scopus database as of June 2016
- **202** recent publications covered in the 12 issues of the **Wikimedia Research Newsletter** from June 2015 to May 2016

- **6,037** Wikipedia-related publications in the Scopus database as of June 2016
- **202** recent publications covered in the 12 issues of the **Wikimedia Research Newsletter** from June 2015 to May 2016

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).

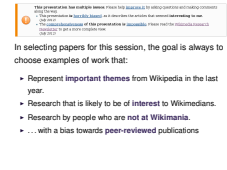


This presentation has multiple issues. Please help improve it by asking questions and making comments along the way.

- 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 my disclaimer slide...

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

Gender and Wikipedia

Tilman

Gender and Wikipedia

Hinnosaar, M. (2015). Gender Inequality in New Media: Evidence from Wikipedia (Carlo Alberto Notebooks No. 411). Working Paper. Collegio Carlo Alberto. Retrieved from http://econpapers.repec.org/paper/ccawpaper/411.htm

Gender and Wikipedia

Hinnosaar, M. (2015). Gender Inequality in New Media: Evidence from Wikipedia (Carlo Alberto Notebooks No. 411). Working Paper. Collegio Carlo Alberto. Retrieved from http://econpapers.repec.org/paper/ccawpaper/411.htm

Study of Wikipedia's gender gap has progressed from examing the demographics of contributors to the gap's possible effects on Wikipedia's content (see also last year's State of Wikimedia Scholarship presentation), and its causes.

Gender and Wikipedia

Data sources:

- Survey and experiment with 1000 Amazon Mechanical Turk users
- Dataset of biographical articles (with gender from Wikidata)
- Self-stated gender (only provided by small minority of editors)
- Pageview data

Gender and Wikipedia

Data sources:

- Survey and experiment with 1000 Amazon Mechanical Turk users
- Dataset of biographical articles (with gender from Wikidata)
- Self-stated gender (only provided by small minority of editors)
- Pageview data

Turkers from the US only, paid $1.50 for a 20 minutes task

Gender and Wikipedia

.... the number of readers per editor is higher for articles about women ... On a typical (median) day in September 2014, no one read 26 percent of the biographies of men versus only 16 percent of the biographies of women.

...almost half of the gender gap in Wikipedia writing is explained by gender differences in two characteristics: frequency of Wikipedia use and belief about one's competence ... The gender difference in the belief about competence could be due to women being less competent or due to women underestimating their competence.

women are about twice as likely as men to contribute to Wikipedia articles about women

Gender and Wikipedia

... the number of readers per editor is higher for articles about women ... On a typical (median) day in September 2014, no one read 26 percent of the biographies of men versus only 16 percent of the biographies of women.

...almost half of the gender gap in Wikipedia writing is explained by gender differences in two characteristics: frequency of Wikipedia use and belief about one's competence ... The gender difference in the belief about competence could be due to women being less competent or due to women underestimating their competence.

women are about twice as likely as men to contribute to Wikipedia articles about women

Gender and Wikipedia

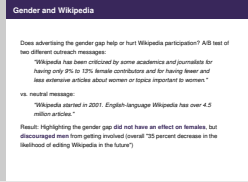
Does advertising the gender gap help or hurt Wikipedia participation? A/B test of two different outreach messages:

> *"Wikipedia has been criticized by some academics and journalists for having only 9% to 13% female contributors and for having fewer and less extensive articles about women or topics important to women."*

vs. neutral message:

> *"Wikipedia started in 2001. English-language Wikipedia has over 4.5 million articles."*

Result: Highlighting the gender gap **did not have an effect on females**, but **discouraged men** from getting involved (overall "35 percent decrease in the likelihood of editing Wikipedia in the future")



"The result provides an example where encouraging gender equality can partially backfire. Wikipedia has set a goal to increase the share of female editors. One way to achieve this is by discouraging male editors. However, this might not be desirable"

Student Use of Wikipedia

Tilman

Student Use of Wikipedia

Selwyn, N., & Gorard, S. (2016). Students' use of Wikipedia as an academic resource — Patterns of use and perceptions of usefulness. The Internet and Higher Education, 28, 28–34.
http://doi.org/10.1016/j.iheduc.2015.08.004

Student Use of Wikipedia

Selwyn, N., & Gorard, S. (2016). Students' use of Wikipedia as an academic resource — Patterns of use and perceptions of usefulness. The Internet and Higher Education, 28, 28–34.
http://doi.org/10.1016/j.iheduc.2015.08.004

Student Use of Wikipedia

- Survey at two Australian universities: 1658 (self-selecting) respondents followed by group interviews
- Asked about whether students used Wikipedia for academic work, and how useful they rated it compared to other tools (e.g. library website, Facebook...)

Results:

- Wikipedia kind of in the middle of the field regarding frequency of use and perceived usefulness
- Gender difference: "Looking for information on Wikipedia was perceived to be more useful by males (76.7%) as opposed to females (58.7%)."

Student Use of Wikipedia

- Survey at two Australian universities: 1658 (self-selecting) respondents followed by group interviews
- Asked about whether students used Wikipedia for academic work, and how useful they rated it compared to other tools (e.g. library website, Facebook...)

Results:

- Wikipedia kind of in the middle of the field regarding frequency of use and perceived usefulness
- Gender difference: "Looking for information on Wikipedia was perceived to be more useful by males (76.7%) as opposed to females (58.7%)."

Student Use of Wikipedia

Gender difference probably partly due to differences by subject: "78.2% of respondents studying Engineering, Computer Science & Maths subjects reporting Wikipedia as useful, as compared to 34.4% of students studying Education subjects."

	Make use of Wikipedia as part of their academic studies	If using Wikipedia, then find it to be 'useful' or 'very useful' for academic studies
Engineering, Computer Science & Maths	96.8	78.2
Law	91.9	53.2
Creative arts and design	91.5	68.2
Sciences (physical and biological)	90.6	72.7
Humanities, languages and library studies	86.8	61.7
Medicine (and allied subjects)	84.8	67.7
Business	83.9	64.6
Social sciences, economics and politics	80.6	67.1
Education	72.3	34.4

Student Use of Wikipedia

Gender difference probably partly due to differences by subject: "78.2% of respondents studying Engineering, Computer Science & Maths subjects reporting Wikipedia as useful, as compared to 34.4% of students studying Education subjects."

Physical Models of Wikipedia

This was the year of physcists studying wikipedia.

Physical Models of Wikipedia

Iñiguez, G., Török, J., Yasseri, T., Kaski, K., & Kertész, J. (2014). Modeling social dynamics in a collaborative environment. EPJ Data Science, 3(1), 1–20. http://doi.org/10.1140/epjds/s13688-014-0007-z

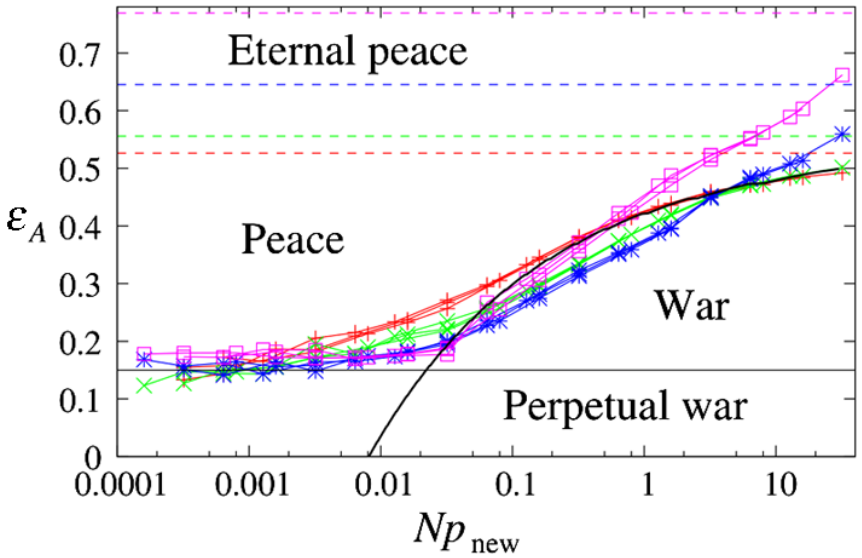
Physical Models of Wikipedia

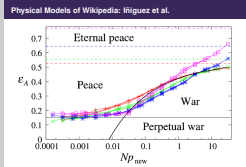
Iñiguez, G., Török, J., Yasseri, T., Kaski, K., & Kertész, J. (2014). Modeling social dynamics in a collaborative environment. EPJ Data Science, 3(1), 1–20. http://doi.org/10.1140/epjds/s13688-014-0007-z

The approach in this whole kind of work is to start out and create a simple mathematical model of a phenomena. The goal is to show that using 4-5 variables, you can make good predictions of certain outcomes.

In this case, the goal is predict the amount of conflict in Wikipedia articles.

Physical Models of Wikipedia: Iñiguez et al.

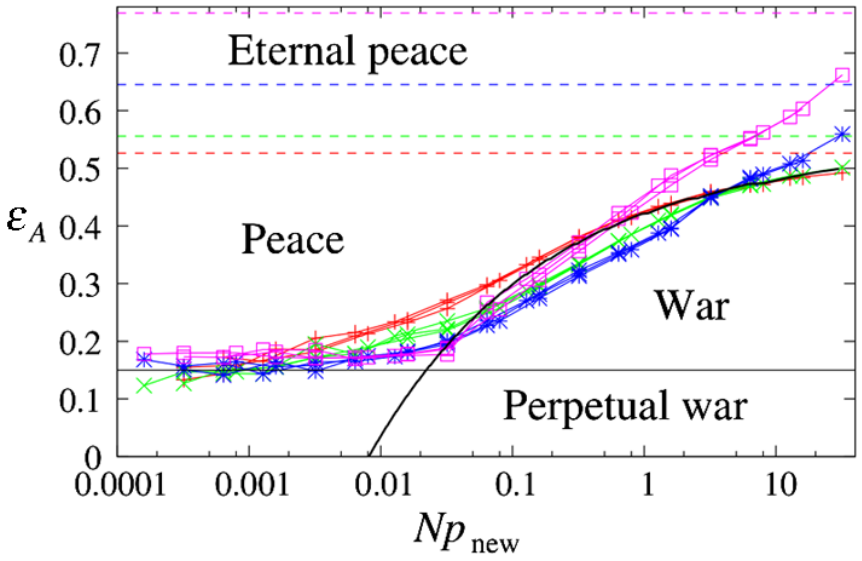


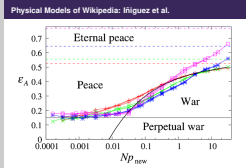


This is their model (no data is used here. This is purely what your model predicts. The parameters are:

- Everything on the top of the graph is conflict. Everything on the bottom of the graph is not.
- The threshold for conflict (or how close to people need to be to start compromising) (ϵ).
- **Not shown.** The rate or speed to which people will converge toward each others (u).
- The chance that a new editor will join (and replace an old one) (N).

Physical Models of Wikipedia: Iñiguez et al.





The takeaway is this:

If people have a have a low threshold for compromise (basically, they're willing to put up with a lot) and let things stay in an article even if they disagree with them, things will be stable.

If people are unwilling to compromise, it will be constant conflict.

Other than that, it's a function of how many newcomers are showing up with new opions. Articles in constant church will be in rough shape unless folks are all saints and are willing to put up with anything.

Physical Models of Wikipedia: Iñiguez et al.

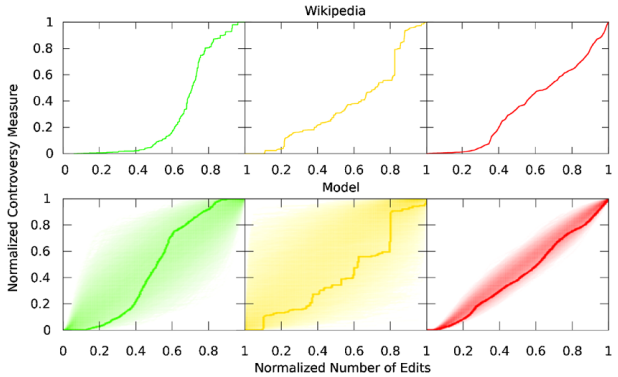
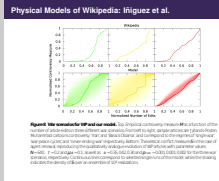


Figure 8 War scenarios for WP and our model. Top: Empirical controversy measure M as a function of the number of article edits in three different war scenarios. From left to right, sample articles are 'Jyllands-Posten Muhammad cartoons controversy', 'Iran', and 'Barack Obama'; and correspond to the regimes of 'single war', 'war-peace cycles', and 'never-ending war' respectively. Bottom: Theoretical conflict measure S in the case of agent renewal, reproducing the qualitatively analogue evolution of WP articles with parameter values $N = 640$, $\tau = 0.2$ and $\mu_A = 0.1$, as well as $A = 0.35, 0.42, 0.30$ and $p_{new} = 0.001, 0.001, 0.002$ for the three war scenarios, respectively. Continuous lines correspond to selected single runs of the model, while the shading indicates the density of S over an ensemble of 10^4 realizations.



The question now is, is this model any good? Does it describe the way that Wikipedia works?

- The y-axis is the amount of controversy.
- The x-axis is the number of edits in a particular period.
- The colors are three articles: 'Jyllands-Posten Muhammad cartoons controversy', 'Iran', and 'Barack Obama'; and correspond to the regimes of 'single war', 'war-peace cycles', and 'never-ending war' respectively
- The top is the actual amount of controversy over time.
- The bottom is the amount predicted by the model!

The thing to takeaway is that the model works! It provides a good prediction of when controversy will happen across various kinds of articles as a function of the amount of editing.

Wikipedia and Media Ecosystems

Mako

Wikipedia and Media Ecosystems

Zangerle, E., Schmidhammer, G., & Specht, G. (2015). #Wikipedia on Twitter: Analyzing Tweets About Wikipedia. In Proceedings of the 11th International Symposium on Open Collaboration (p. 14:1–14:8). New York, NY, USA: ACM.
http://doi.org/10.1145/2788993.2789845

Also see:

- Wikipedia on Reddit [Carson et al. (2015) Determining the influence of reddit posts on wikipedia pageviews.]
- News coverage and Wikipedia [Geiß et al. (2015) The interplay between media-for-monitoring and media-for-searching: How news media trigger searches and edits in Wikipedia]

Wikipedia and Media Ecosystems
Zangerle, E., Schmidhammer, G., & Specht, G. (2015). #Wikipedia on Twitter: Analyzing Tweets About Wikipedia. In Proceedings of the 11th International Symposium on Open Collaboration (p. 14:1–14:8). New York, NY, USA: ACM.
http://doi.org/10.1145/2788993.2789845
Also see:
- Wikipedia on Reddit [Carson et al. (2015) Determining the influence of reddit posts on wikipedia pageviews.]
- News coverage and Wikipedia [Geiß et al. (2015) The interplay between media-for-monitoring and media-for-searching: How news media trigger searches and edits in Wikipedia]

These articles are all about looking at how activity (e.g., in social media and the news) affects Wikipedia. And about tracking about how Wikipedia is tweeted about and categorizing the types of ways that Wikipedia is talked about from outside.

Wikipedia and Media Ecosystems

Teplitskiy, M., Lu, G., & Duede, E. (2016 Forthcoming). Amplifying the Impact of Open Access: Wikipedia and the Diffusion of Science. Journal of the Association for Information Science and Technology. http://doi.org/10.1002/asi.23687

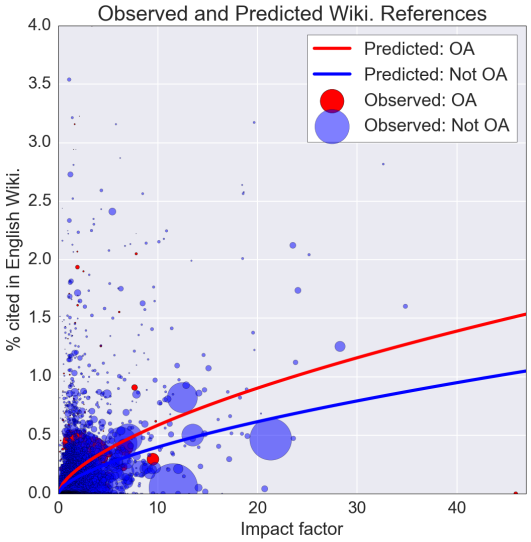
Wikipedia and Media Ecosystems
Teplitskiy, M., Lu, G., & Duede, E. (2016 Forthcoming). Amplifying the Impact of Open Access: Wikipedia and the Diffusion of Science. Journal of the Association for Information Science and Technology.
http://doi.org/10.1002/asi.23687

This article does kind of the opposite. It tries to understand of the effectiveness of other forms of publishing by looking at Wikipedia as an outcome or a measure of success. In this case, the research question is: Does open access work? Does it increase exposure to scientific work?

We'll measure this by looking at whether or not articles are cited in Wikipedia.

This work takes a dataset of every article published in the top 250 journals in 26 fields (5000 journals) and looks at every article and measures what portions of articles that have been published in the journal were cited in Wikipedia.

Wikipedia and Media Ecosystems: Teplitskiy et al.



The answer is not many! The top journal was 4% and the mean was 0.5%.
Of course, impact factor of journals has a modest relation/correlation to the proportion of articles. High prestige (i.e., more cited) journals tend to be cited in Wikipedia more. But the relationship is not super strong. The slope of the line shows that relationship.

The real takeaway is about the OA versus non-OA comparison as you can see in the line.

Beyond Wikipedia: Studies of Sister Projects

Tilman

Beyond Wikipedia: Studies of Sister Projects

- Vast majority of research about Wikimedia projects is about Wikipedia
  - (ca. 95% of Google Scholars search results since 2015)
- **Wikidata** and **Wiktionary** are the sister projects with the most recent research attention.

Beyond Wikipedia: Studies of Sister Projects
- Vast majority of research about Wikimedia projects is about Wikipedia
  - (ca. 95% of Google Scholars search results since 2015)
- **Wikidata** and **Wiktionary** are the sister projects with the most recent research attention.

- Most papers focus on the data they offer for reuse:
  - Wikidata: Structured, machine-readable information
  - Wiktionary: Dictionary information (somewhat structured but less machine readabl)
- But there has been research on community processes too.

Beyond Wikipedia: Studies of Sister Projects

Müller-Birn, C., Karran, B., Lehmann, J., & Luczak-Rösch, M. (2015). Peer-production System or Collaborative Ontology Engineering Effort: What is Wikidata? In Proceedings of the 11th International Symposium on Open Collaboration (p. 20:1–20:10). New York, NY, USA: ACM. http://doi.org/10.1145/2788993.2789836

- Analysis of nearly 165 million edits from Wikidata's first two years (October 2012-October 2014)
  - ca. 85% by 160 bots, 15% by registered users, < 1% by anonymous users

Beyond Wikipedia: Studies of Sister Projects

Müller-Birn, C., Karran, B., Lehmann, J., & Luczak-Rösch, M. (2015). Peer-production System or Collaborative Ontology Engineering Effort: What is Wikidata? In Proceedings of the 11th International Symposium on Open Collaboration (p. 20:1–20:10). New York, NY, USA: ACM. http://doi.org/10.1145/2788993.2789836

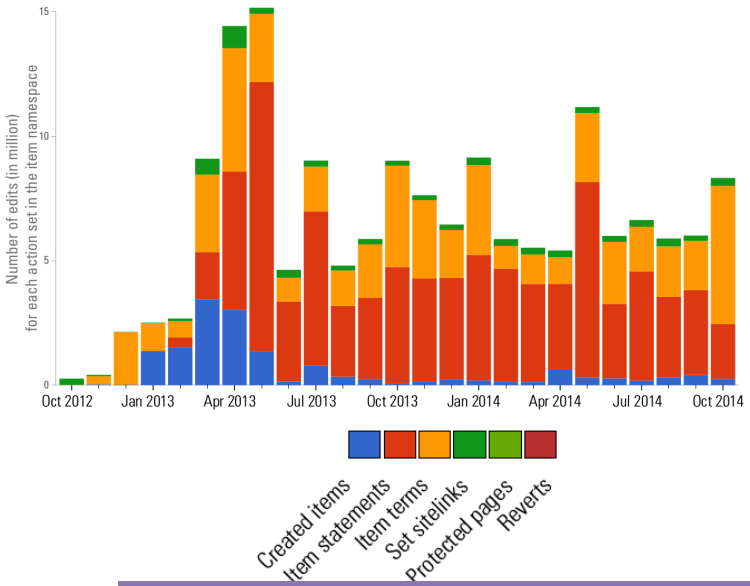
- Analysis of nearly 165 million edits from Wikidata's first two years (October 2012-October 2014)
  - ca. 85% by 160 bots, 15% by registered users, < 1% by anonymous users

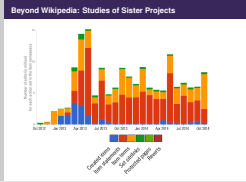
Beyond Wikipedia: Studies of Sister Projects

Classified Wikidata edits into action types, such as:

- Create items (e.g. Q30 for the United States)
- Edit statements about items (e.g.: "Mount McKinley is the highest point of the United States")
- Edit terms (e.g. descriptions: "country in North America" and labels: "Stati Uniti d'America")

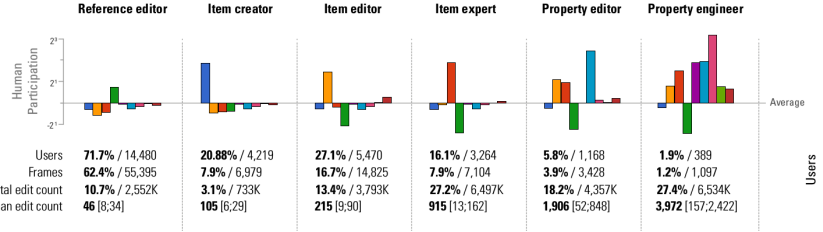
Beyond Wikipedia: Studies of Sister Projects





- Wikipedia articles were imported in bulk as items in early 2013, and later added/modified as "sitelinks"
- After that, statement and term edits dominate.

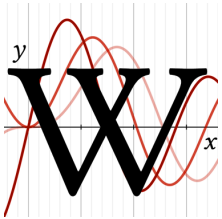
Clustered human and bot editors into "patterns of participation".
For humans: Reference Editor (mainly adds sitelinks to Wikipedia etc. - not reference in the sense of citations), Item Creator, Item Editor (mainly edits terms), Item Expert (sets statements), Property Editor (sets terms on properties), Property Engineer (creates new properties using special user right, discusses them)



Conclusion: "Wikidata finds itself between two approaches – 'classic' peer-production and collaborative ontology engineering. .. It seems that the simplified user interface of Wikidata is very valuable for data contributions, but less valuable for [systematic] data modelling." However, the researchers speculated that it could veer more into more systematic ontology engineering in the future.

More Resources

- **Wikimedia Research Newsletter**
  [[:meta:Research:Newsletter]] / @WikiResearch
- **WikiSym/OpenSym** (This August in Berlin!)
- **WikiPapers Repository** [http://wikipapers.referata.com]
- **Much More**



More Resources
- Wikimedia Research Newsletter [[:meta:Research:Newsletter]] / @WikiResearch
- WikiSym/OpenSym (This August in Berlin!)
- WikiPapers Repository [http://wikipapers.referata.com]
- Much More

Those are my six exemplary studies from the past year.
There has been just tons and tons of work in this area. Trying to talk about this in 20 minutes strikes me as increasingly crazy every year I try to do it.
The most important source, now going for a couple years, is the Wikimedia Research Newsletter which is published monthly in the (English) Signpost and syndicated on the Wikimedia Research.
But there are other resources as well. And I encourage you to get involved.