The State of Wikimedia Research: 2017–2018

Tilman Bayer
Benjamin Mako Hill
Reem Al-Kashif
Mohammed Sadat Abdulai

Wikimania 2018, Cape Town
July 21, 2018

2018-07-21

Presentation Title

The State of Wikimedia Research: 2017–2018

Tilman Bayer
Benjamin Mako Hill
Reem Al-Kashif
Mohammed Sadat Abdulai
Wikimania 2018, Cape Town
July 21, 2018

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

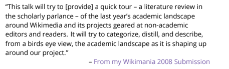
This began as an excuse for me to make sure I was up to date on Wikimedia Research.

"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

2018-07-21

Presentation Title



"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

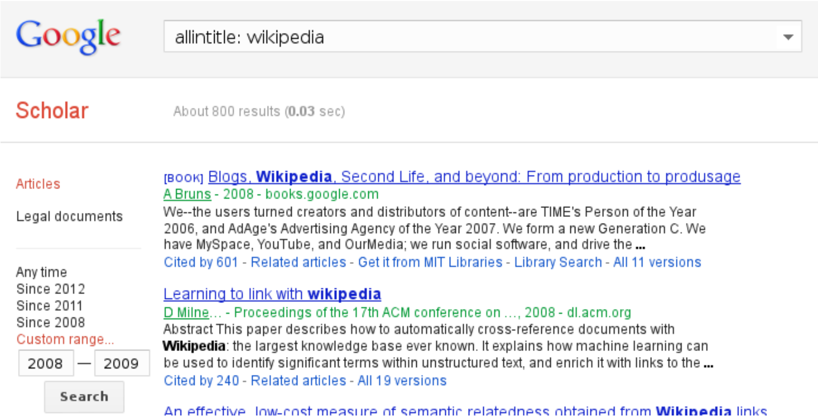
Back at Wikimania 2008, I set out to run a session that would provide a comprehensive literature review of articles in Wikipedia published in the last year.

> *"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



About 800 results (0.03 sec)

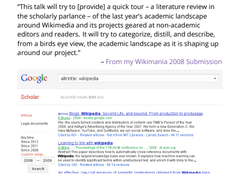
[BOOK] Blogs, **Wikipedia**, Second Life, and beyond: From production to produsage
A Bruns - 2008 - books.google.com
We--the users turned creators and distributors of content--are TIME's Person of the Year 2006, and AdAge's Advertising Agency of the Year 2007. We form a new Generation C. We have MySpace, YouTube, and OurMedia; we run social software, and drive the ...
Cited by 601 - Related articles - Get it from MIT Libraries - Library Search - All 11 versions

Learning to link with **wikipedia**
D Milne... - Proceedings of the 17th ACM conference on ..., 2008 - dl.acm.org
Abstract This paper describes how to automatically cross-reference documents with **Wikipedia**: the largest knowledge base ever known. It explains how machine learning can be used to identify significant terms within unstructured text, and enrich it with links to the ...
Cited by 240 - Related articles - All 19 versions

An effective, low-cost measure of semantic relatedness obtained from **Wikipedia** links

2018-07-21

Presentation Title



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.

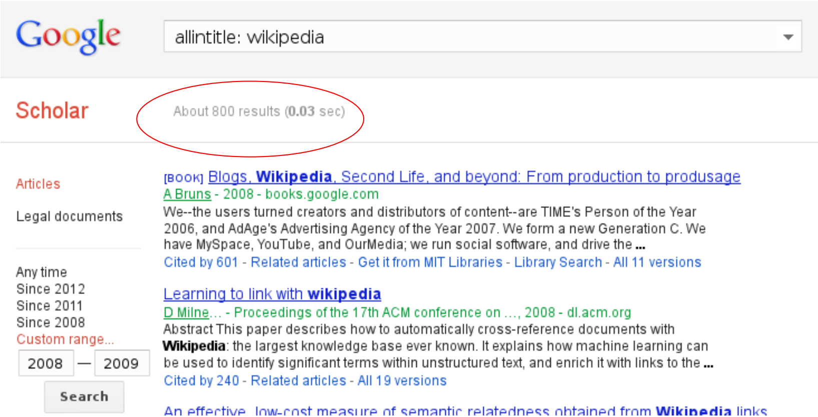
So anyway, I had a 45 minute talk so it worked out to 3.45 seconds to per paper...

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

And this 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



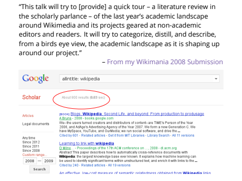
[BOOK] Blogs, Wikipedia, Second Life, and beyond: From production to produsage
A Bruns - 2008 - books.google.com
We--the users turned creators and distributors of content--are TIME's Person of the Year 2006, and AdAge's Advertising Agency of the Year 2007. We form a new Generation C. We have MySpace, YouTube, and OurMedia; we run social software, and drive the ...
Cited by 601 - Related articles - Get it from MIT Libraries - Library Search - All 11 versions

Learning to link with wikipedia
D Milne... - Proceedings of the 17th ACM conference on ..., 2008 - dl.acm.org
Abstract This paper describes how to automatically cross-reference documents with **Wikipedia**: the largest knowledge base ever known. It explains how machine learning can be used to identify significant terms within unstructured text, and enrich it with links to the ...
Cited by 240 - Related articles - All 19 versions

An effective, low-cost measure of semantic relatedness obtained from Wikipedia links

Presentation Title

2018-07-21

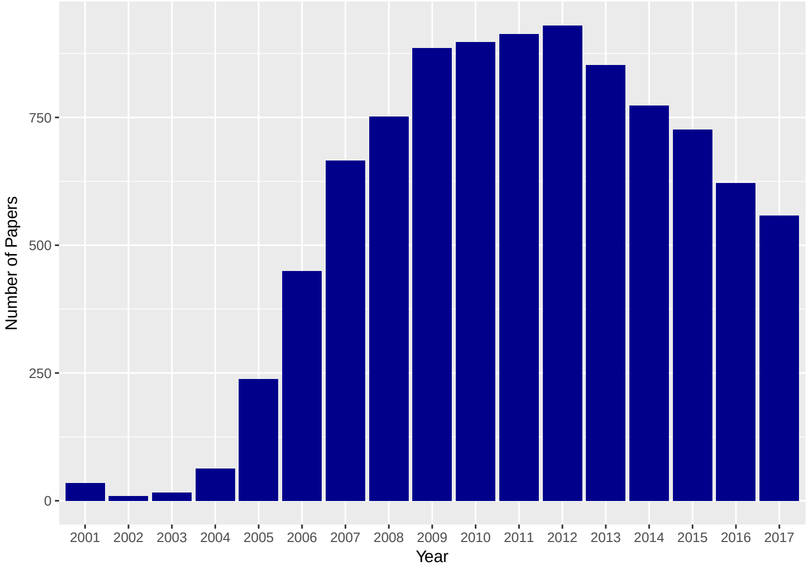

"This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – of the last year's academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape 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 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 this talk is even shorter.

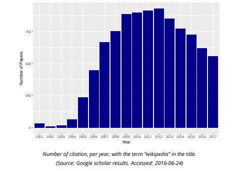


Number of citation, per year, with the term “wikipedia” in the title.

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

2018-07-21

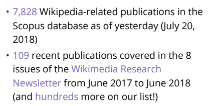
Presentation Title


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

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

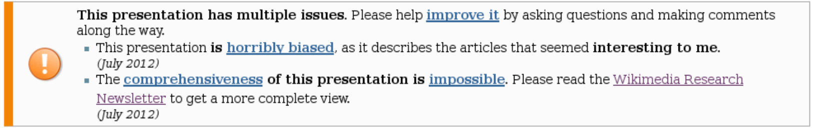
Academics have written a lot of papers about Wikipedia. There are more than 500 papers published about Wikipedia each year and although we’ve reached and moved past a peak it seems, it’s not slowing by much.

- 7,828 Wikipedia-related publications in the Scopus database as of yesterday (July 20, 2018)
- 109 recent publications covered in the 8 issues of the Wikimedia Research Newsletter from June 2017 to June 2018 (and hundreds more on our list!)



- 7,828 Wikipedia-related publications in the Scopus database as of yesterday (July 20, 2018)
- 109 recent publications covered in the 8 issues of the Wikimedia Research Newsletter from June 2017 to June 2018 (and hundreds more on our list!)

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.

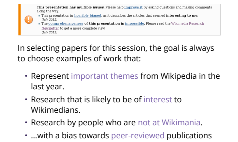
- This presentation **is** horribly biased, as it describes the articles that seemed **interesting to me.** *(July 2012)*
- The comprehensiveness **of this presentation is** impossible. Please read the Wikimedia Research Newsletter to get a more complete view. *(July 2012)*

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

2018-07-21

Presentation Title



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...

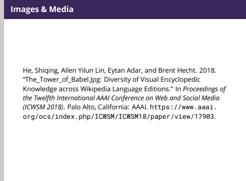
Within these goals, the selections are incomplete, and wrong.

Images & Media

Images & Media

Images & Media

He, Shiqing, Allen Yilun Lin, Eytan Adar, and Brent Hecht. 2018. “The_Tower_of_Babel.Jpg: Diversity of Visual Encyclopedic Knowledge across Wikipedia Language Editions.” In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*. Palo Alto, California: AAAI. `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17903`.



Images & Media

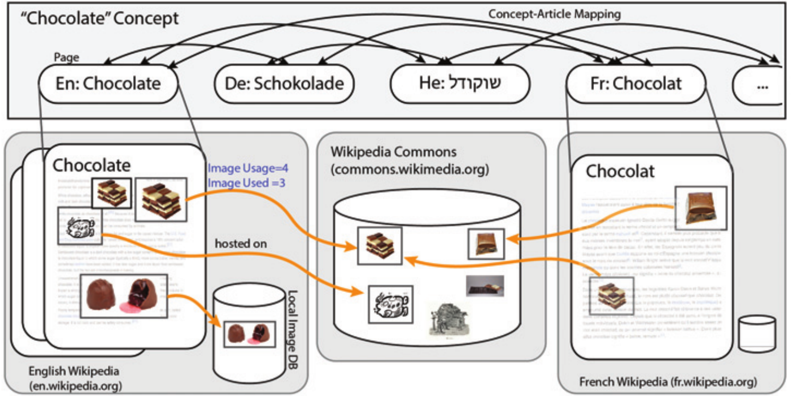
He, Shiqing, Allen Yilun Lin, Eytan Adar, and Brent Hecht. 2018. “The_Tower_of_Babel.Jpg: Diversity of Visual Encyclopedic Knowledge across Wikipedia Language Editions.” In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*. Palo Alto, California: AAAI. `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17903`.

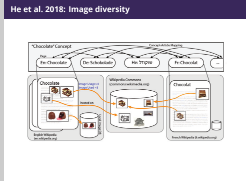
Mako

This paper is by a team at the University of Michigan and Northwestern University and it looks as image use.

Image use is something that has historically been studied very little. This year, it sort of exploded in popularity and there were a series of papers on the topic.

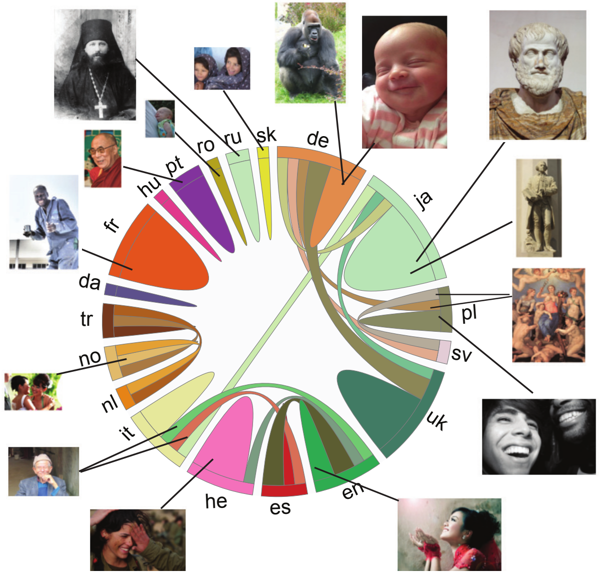
He et al. 2018: Image diversity

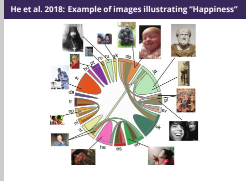



He et al. 2018: Image diversity

This paper really focused on understanding "image diversity" and it looks at it in the biggest 25 language editions of Wikipedia. This is what they mean by image diversity is that they found articles on the same topic (from inter-wiki links stored in WikiData and in the individual wikis) and then they looked at overlap in terms of images in commons.

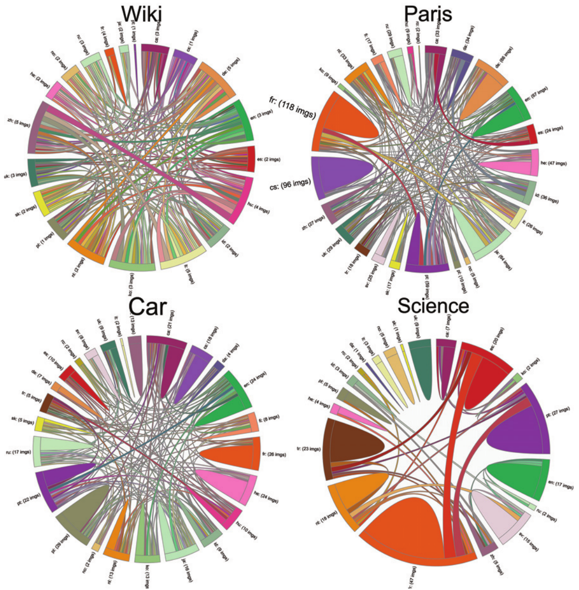
He et al. 2018: Example of images illustrating “Happiness”

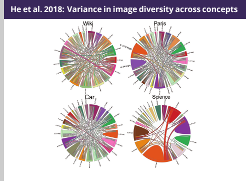



He et al. 2018: Example of images illustrating “Happiness”

Here is an example from the article on happiness. German shows a gorilla. Some images show up in a few. But—in general—there’s a ton diversity.

He et al. 2018: Variance in image diversity across concepts

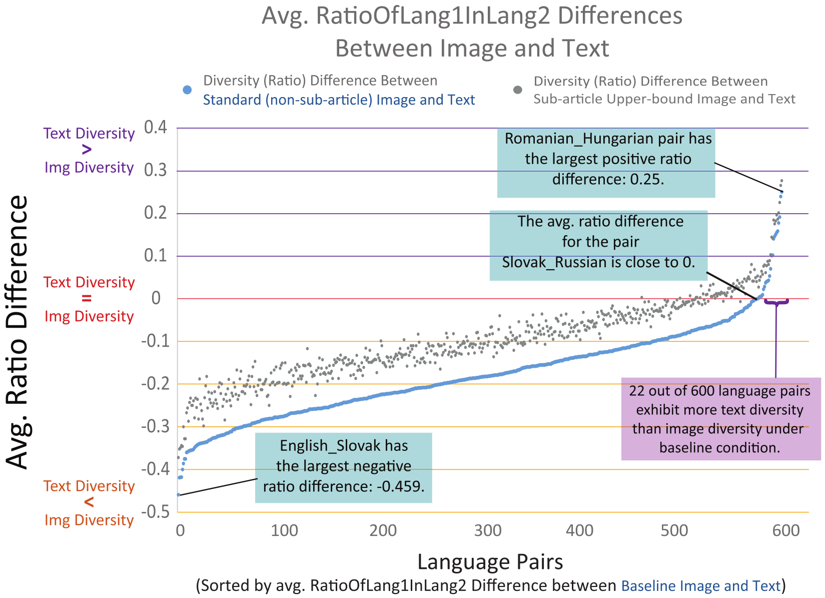




They found that 67% of images appear in only one of the 25 editions.

Some concepts—like wiki—have a tone of overlap. Other concepts–like science—have a huge amount of diversity.

He et al. 2018: Diversity in text and images



Avg. RatioOfLang1InLang2 Differences Between Image and Text

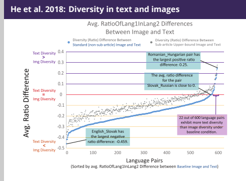
Romanian_Hungarian pair has the largest positive ratio difference: 0.25.

The avg. ratio difference for the pair Slovak_Russian is close to 0.

22 out of 600 language pairs exhibit more text diversity than image diversity under baseline condition.

English_Slovak has the largest negative ratio difference: -0.459.

● Diversity (Ratio) Difference Between Standard (non-sub-article) Image and Text
● Diversity (Ratio) Difference Between Sub-article Upper-bound Image and Text
Avg. Ratio Difference
Text Diversity > Img Diversity
Text Diversity = Img Diversity
Text Diversity < Img Diversity
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
0
100
200
300
400
500
600
Language Pairs
(Sorted by avg. RatioOfLang1InLang2 Difference between Baseline Image and Text)



Previous work has shown that text is very diverse in Wikipedia in the sense that different languages use different concepts to talk about a particular topic.

There was reason to believe that there might be less in image since they are hosted in commons and don't need to be localized.

Every dot on this graph is a language pair. Things below the red line have more image diversity than text diversity.

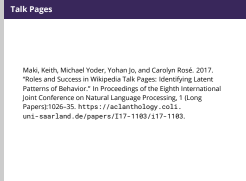
As you can see, there is generally much more image diversity than text diversity.

Talk Pages

Talk Pages

Talk Pages

Maki, Keith, Michael Yoder, Yohan Jo, and Carolyn Rosé. 2017. “Roles and Success in Wikipedia Talk Pages: Identifying Latent Patterns of Behavior.” In Proceedings of the Eighth International Joint Conference on Natural Language Processing, 1 (Long Papers):1026–35. `https://aclanthology.coli.uni-saarland.de/papers/I17-1103/i17-1103`.



Talk Pages

Maki, Keith, Michael Yoder, Yohan Jo, and Carolyn Rosé. 2017. “Roles and Success in Wikipedia Talk Pages: Identifying Latent Patterns of Behavior.” In Proceedings of the Eighth International Joint Conference on Natural Language Processing, 1 (Long Papers):1026–35. `https://aclanthology.coli.uni-saarland.de/papers/I17-1103/i17-1103`.

Reem

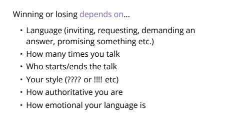
Drawn by J.H. Mortimer.
Killing an Enemy.
Etched by R. Blyth.
London. Published as the Act directs 15th Nov. 1779, by R. Blyth N. 27, Great Castle Street, Cavendish Square.

Whose suggestions/opinions make it to the article and do not get reverted?
53k+ instances of interaction on talk pages paired with edit actions were analyzed.

Whose suggestions/opinions make it to the article and do not get reverted?
53k+ instances of interaction on talk pages paired with edit actions were analyzed.

Winning or losing depends on...

- Language (inviting, requesting, demanding an answer, promising something etc.)
- How many times you talk
- Who starts/ends the talk
- Your style (???? or !!!! etc)
- How authoritative you are
- How emotional your language is

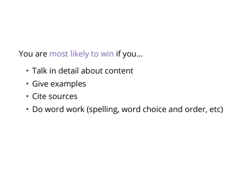


Winning or losing depends on...

- Language (inviting, requesting, demanding an answer, promising something etc.)
- How many times you talk
- Who starts/ends the talk
- Your style (???? or !!!! etc)
- How authoritative you are
- How emotional your language is

You are most likely to win if you...

- Talk in detail about content
- Give examples
- Cite sources
- Do word work (spelling, word choice and order, etc)

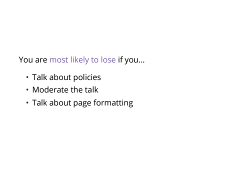


You are most likely to win if you...

- Talk in detail about content
- Give examples
- Cite sources
- Do word work (spelling, word choice and order, etc)

You are most likely to lose if you...

- Talk about policies
- Moderate the talk
- Talk about page formatting


You are most likely to lose if you...

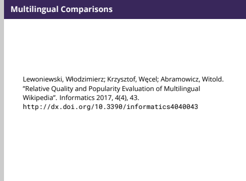
- Talk about policies
- Moderate the talk
- Talk about page formatting

Multilingual Comparisons

Multilingual Comparisons

Multilingual Comparisons

Lewoniewski, Włodzimierz; Krzysztof, Węcel; Abramowicz, Witold. ”Relative Quality and Popularity Evaluation of Multilingual Wikipedia”. Informatics 2017, 4(4), 43. `http://dx.doi.org/10.3390/informatics4040043`


Multilingual Comparisons

Lewoniewski, Włodzimierz; Krzysztof, Węcel; Abramowicz, Witold. ”Relative Quality and Popularity Evaluation of Multilingual Wikipedia”. Informatics 2017, 4(4), 43. http://dx.doi.org/10.3390/informatics4040043

Tilman

Knowledge gaps are the theme of this Wikimania, and in (one form or the other) they have been a big theme in research this year too.

Some of this research is already being presented elsewhere here, so it’s out of scope for this talk. E.g. yesterday’s keynote by Martin Dittus about geographical imbalances, the ”Wikipedia Cultural Diversity Observatory” (which goes beyond geolocation to incorporate other data for a fuller picture of diversity), and the Wikimedia Foundation’s own research and technology development to bridge such knowledge gaps.

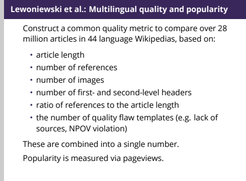
Lewoniewski et al.: Multilingual quality and popularity

Construct a common quality metric to compare over 28 million articles in 44 language Wikipedias, based on:

- article length
- number of references
- number of images
- number of first- and second-level headers
- ratio of references to the article length
- the number of quality flaw templates (e.g. lack of sources, NPOV violation)

These are combined into a single number.

Popularity is measured via pageviews.



Lewoniewski et al.: Multilingual quality and popularity

Construct a common quality metric to compare over 28 million articles in 44 language Wikipedias, based on:

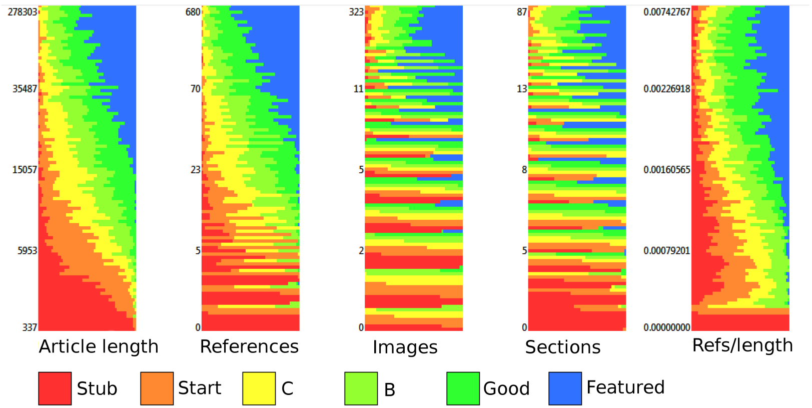
- article length
- number of references
- number of images
- number of first- and second-level headers
- ratio of references to the article length
- the number of quality flaw templates (e.g. lack of sources, NPOV violation)

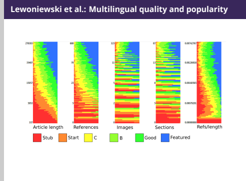
These are combined into a single number.

Popularity is measured via pageviews.

As the authors point out, more sophisticated quality metrics exist, including the Wikimedia Foundation's ORES service, which is machine learning based. They didn't use it because it was only available for three languages.

Lewoniewski et al.: Multilingual quality and popularity



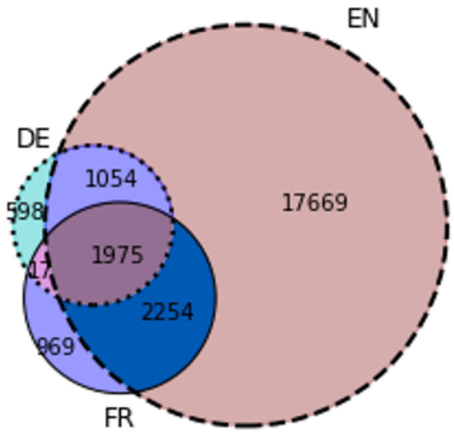


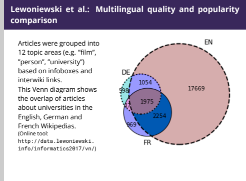
These five metrics are positively correlated with the quality grades that editors assign manually on the English Wikipedia.

E.g. on the left you can see that there are almost no featured articles (blue) with less than 15000 bytes length. But more than half of the articles over 250k have featured article status.

Lewoniewski et al.: Multilingual quality and popularity comparison

Articles were grouped into 12 topic areas (e.g. ”film”, ”person”, ”university”) based on infoboxes and interwiki links.
This Venn diagram shows the overlap of articles about universities in the English, German and French Wikipedias.
(Online tool: `http://data.lewoniewski.info/informatics2017/vn/`)




Lewoniewski et al.: Multilingual quality and popularity comparison

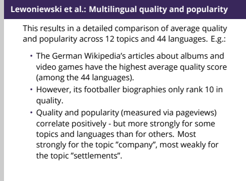
Articles were grouped into 12 topic areas (e.g. ”film”, ”person”, ”university”) based on infoboxes and interwiki links.
This Venn diagram shows the overlap of articles about universities in the English, German and French Wikipedias.
(Online tool: `http://data.lewoniewski.info/informatics2017/vn/`)

Categories were deliberately not used, and Wikidata isn’t mentioned in the paper at all.

Lewoniewski et al.: Multilingual quality and popularity

This results in a detailed comparison of average quality and popularity across 12 topics and 44 languages. E.g.:

- The German Wikipedia’s articles about albums and video games have the highest average quality score (among the 44 languages).
- However, its footballer biographies only rank 10 in quality.
- Quality and popularity (measured via pageviews) correlate positively - but more strongly for some topics and languages than for others. Most strongly for the topic ”company”, most weakly for the topic ”settlements”.



Lewoniewski et al.: Multilingual quality and popularity

This results in a detailed comparison of average quality and popularity across 12 topics and 44 languages. E.g.:

- The German Wikipedia’s articles about albums and video games have the highest average quality score (among the 44 languages).
- However, its footballer biographies only rank 10 in quality.
- Quality and popularity (measured via pageviews) correlate positively - but more strongly for some topics and languages than for others. Most strongly for the topic ”company”, most weakly for the topic ”settlements”.

NB: This result does not necessarily mean that the German Wikipedia has the best experts about albums and video games among its editors. More likely, this is because its overage of these topics is much more limited due to stricter notability criteria. (Some quick comparisons via the Venn diagram tool seem to confirm that other major languages have many more articles about these topics.)

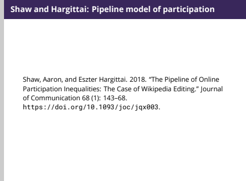
The authors wisely refrain from calculating an overall quality score for each Wikipedia. I myself was less prudent and couldn’t resist playing around with their data to (rather unscientifically) calculate the average of all topic averages for each language. By that measure, the German Wikipedia would come out on top - but only narrowly, closely followed by the English, Greek, Hindi and Chinese Wikipedia ;)

Nonparticipation:
Who is not contributing?

Once again, an important theme this year—related to knowledge equity—is Why do internet users from different social groups contribute differently to Wikipedia?

Shaw and Hargittai: Pipeline model of participation

Shaw, Aaron, and Eszter Hargittai. 2018. “The Pipeline of Online Participation Inequalities: The Case of Wikipedia Editing.” Journal of Communication 68 (1): 143–68. `https://doi.org/10.1093/joc/jqx003`.

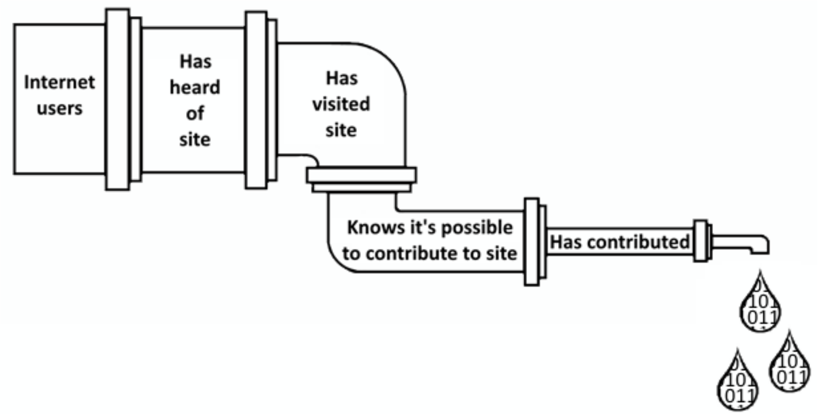


Shaw and Hargittai: Pipeline model of participation

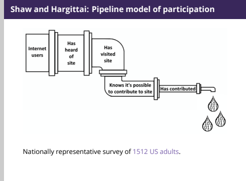
Shaw, Aaron, and Eszter Hargittai. 2018. “The Pipeline of Online Participation Inequalities: The Case of Wikipedia Editing.” Journal of Communication 68 (1): 143–68. `https://doi.org/10.1093/joc/jqx003`.

This paper explored the factors and processes that influence these ‘participation gaps.’

Shaw and Hargittai: Pipeline model of participation



Nationally representative survey of 1512 US adults.


Shaw and Hargittai: Pipeline model of participation

Nationally representative survey of 1512 US adults.

Analysing survey data collected from 1512 adults in the US in 2016, the authors used logistic regression to model the activity of online knowledge production as a step-by-step process that internet users who contribute to Wikipedia go through.

They conceptualized a pipeline that anticipates leaks at the different stages of the knowledge production process so that fewer contributors remain at each subsequent step, beginning from a cohort of internet users.

Most work on the participation gap has focused on the final stage about whether or not people contribute. The authors of this paper show that there are gaps at many earlier stages such as whether or not people know that Wikipedia is editable, whether they have been on the site, or whether they know it even exists.

Shaw and Hargittai: Pipeline model of participation

Participation increased at all stages of the pipeline when respondents'

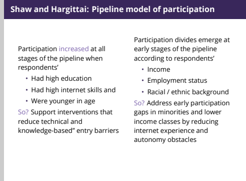
- Had high education
- Had high internet skills and
- Were younger in age

So? Support interventions that reduce technical and knowledge-based" entry barriers

Participation divides emerge at early stages of the pipeline according to respondents'

- Income
- Employment status
- Racial / ethnic background

So? Address early participation gaps in minorities and lower income classes by reducing internet experience and autonomy obstacles



Shaw and Hargittai: Pipeline model of participation

Participation increased at all stages of the pipeline when respondents'

- Had high education
- Had high internet skills and
- Were younger in age

So? Support interventions that reduce technical and knowledge-based" entry barriers

Participation divides emerge at early stages of the pipeline according to respondents'

- Income
- Employment status
- Racial / ethnic background

So? Address early participation gaps in minorities and lower income classes by reducing internet experience and autonomy obstacles

The results showed that: (At all stages of the pipeline): Education levels, internet literacy levels, and age; significantly influenced levels of activity at each step of the pipeline.

(Recommendation): With this information, the authors recommend the "support to interventions that reduce technical and knowledge-based" entry barriers as a means to increase participation at all the levels of knowledge production.

(At the early stages of the pipeline): Income, employment and race are significant factors that influence levels of activity in that stage of knowledge production.

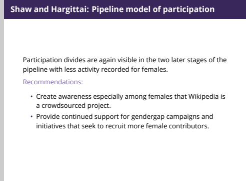
(Recommendation): "This suggests the need for interventions addressing early participation gaps in minorities and lower income classes by reducing internet experience and autonomy obstacles".

Shaw and Hargittai: Pipeline model of participation

Participation divides are again visible in the two later stages of the pipeline with less activity recorded for females.

Recommendations:

- Create awareness especially among females that Wikipedia is a crowdsourced project.
- Provide continued support for gendergap campaigns and initiatives that seek to recruit more female contributors.



Shaw and Hargittai: Pipeline model of participation

Participation divides are again visible in the two later stages of the pipeline with less activity recorded for females.

Recommendations:

- Create awareness especially among females that Wikipedia is a crowdsourced project.
- Provide continued support for gendergap campaigns and initiatives that seek to recruit more female contributors.

(At the later stage of the pipeline): Gender played a role to determine that, compared with males, fewer people who identify as female know that "Wikipedia is editable" and actually go beyond that awareness to contribute to Wikipedia. (Recommendation): The results therefore suggests two things; the need to 1. Create awareness among females that Wikipedia is a crowdsourced project that anybody can edit. 2. To continue support for gendergap campaigns and initiatives that seek to recruit more female contributors.

Wikipedia as a Source of Data

Wikipedia as a Source of Data

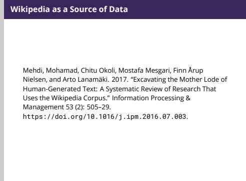
Mako

Perhaps the only topic that we’ve covered ever year is studies that use Wikipedia as source of data because there are loads and loads of these papers—every year.

Once again, this year saw a new crop of these.

Wikipedia as a Source of Data

Mehdi, Mohamad, Chitu Okoli, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. 2017. “Excavating the Mother Lode of Human-Generated Text: A Systematic Review of Research That Uses the Wikipedia Corpus.” Information Processing & Management 53 (2): 505–29. `https://doi.org/10.1016/j.ipm.2016.07.003`.



Wikipedia as a Source of Data

Mehdi, Mohamad, Chitu Okoli, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. 2017. “Excavating the Mother Lode of Human-Generated Text: A Systematic Review of Research That Uses the Wikipedia Corpus.” Information Processing & Management 53 (2): 505–29. `https://doi.org/10.1016/j.ipm.2016.07.003`.

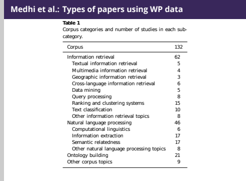
One of these papers was a paper led by Mohamed Medhi at Concordia University in Montréal that uses papers that use Wikipedia as a source of data as… wait for it… a source of data.

This paper is a systematic review of work meaning that it doesn’t present new work. It presents a summary of a large body of other work. In this case, 132 papers that use Wikipedia as a data source

Medhi et al.: Types of papers using WP data

Table 1
Corpus categories and number of studies in each subcategory.

Corpus	132
Information retrieval	62
Textual information retrieval	5
Multimedia information retrieval	4
Geographic information retrieval	3
Cross-language information retrieval	6
Data mining	5
Query processing	8
Ranking and clustering systems	15
Text classification	10
Other information retrieval topics	8
Natural language processing	46
Computational linguistics	6
Information extraction	17
Semantic relatedness	17
Other natural language processing topics	8
Ontology building	21
Other corpus topics	9



Medhi et al.: Types of papers using WP data

Table 1
Corpus categories and number of studies in each subcategory.

Corpus	132
Information retrieval	62
Textual information retrieval	5
Multimedia information retrieval	4
Geographic information retrieval	3
Cross-language information retrieval	6
Data mining	5
Query processing	8
Ranking and clustering systems	15
Text classification	10
Other information retrieval topics	8
Natural language processing	46
Computational linguistics	6
Information extraction	17
Semantic relatedness	17
Other natural language processing topics	8
Ontology building	21
Other corpus topics	9

In addition summarizing papers. they break things down very systematically into 10 tables that categorize papers along a set of dimensions.

For example, they categorize most papers in this space as in the broad area of information retrieval a body of computer and information science focused around giving people good answers to queries.

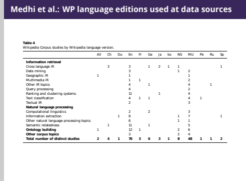
The other big area is natural language processing. In this case, Wikipedia contains data which can help systems that seek to understand language. This might include studies that use wikidata and inter-language links as a source of translation data.

Each has a bunch of subareas.

Medhi et al.: WP language editions used at data sources

Table 4
Wikipedia Corpus studies by Wikipedia language version.

	All	Ch	Du	En	Fr	Ge	Ja	ko	NS	MU	Pe	Ru	Sp
Information retrieval													
Cross-language IR		3		3		1	2	1	1				1
Data mining				3					1	2			
Geographic IR	1			1						1			
Multimedia IR				1	1					2			
Other IR topics				4		1				4		1	
Query processing				4						2			
Ranking and clustering systems				11			1			4			
Text classification				4	1	1				4	1		
Textual IR				2						3			
Natural language processing													
Computational linguistics				2		2				3			
Information extraction			1	9					1	7			1
Other natural language processing topics				6					1	1			
Semantic relatedness		1		11		1				5			
Ontology building	1			12	1				2	6			
Other corpus topics				3					2	4			
Total number of distinct studies	**2**	**4**	**1**	**76**	**3**	**6**	**3**	**1**	**8**	**48**	**1**	**1**	**2**



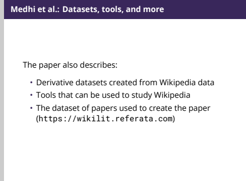
They break things down in lots of ways. And much of what they show is holes. The vast majority of studies that use WP as a data source are focused on English WP. Some use multiple languages.

The vast majority look at article data (and increasingly at WikiData) but not other sources.

Medhi et al.: Datasets, tools, and more

The paper also describes:

- Derivative datasets created from Wikipedia data
- Tools that can be used to study Wikipedia
- The dataset of papers used to create the paper (`https://wikilit.referata.com`)



Medhi et al.: Datasets, tools, and more

The paper also describes:

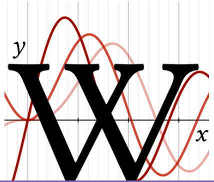
- Derivative datasets created from Wikipedia data
- Tools that can be used to study Wikipedia
- The dataset of papers used to create the paper (`https://wikilit.referata.com`)

Published in 2017 but the paper has been a long time coming. The first version of this paper was submitted in 2014! The speed of academic publishing.

The big change has been a push toward WikiDAta

More Resources

- Wikimedia Research Newsletter [[:meta:Research:Newsletter]] / @WikiResearch
- WikiSym/OpenSym (Next month in France!)
- Wiki Workshop at the Web Conference
- [[:meta:Research:Events]]
- WMF Research Showcase
- Much More



More Resources

- Wikimedia Research Newsletter [[:meta:Research:Newsletter]] / @WikiResearch
- WikiSym/OpenSym (Next month in France!)
- Wiki Workshop at the Web Conference
- [[:meta:Research:Events]]
- WMF Research Showcase
- Much More

Those are our eight exemplary studies from the past year.

There has been just tons and tons of work in this area. Trying to talk about this in 40 minutes strikes me as increasingly crazy every year we try to do it.

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 Dario Taraborelli and User:Masssly for finding and cataloguing new publications throughout the year!)

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