



# 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

“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



**Scholar** About 800 results (0.03 sec)

Articles

Legal documents

Any time

Since 2012

Since 2011

Since 2008

Custom range...

2008 — 2009

Search

[\[book\] 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

“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

The screenshot shows a Google Scholar search interface. At the top, the Google logo is on the left, and a search bar contains the text "allintitle: wikipedia". Below the search bar, the word "Scholar" is displayed in red. To the right of "Scholar", the text "About 800 results (0.03 sec)" is circled in red. On the left side, there are filters for "Articles", "Legal documents", "Any time", "Since 2012", "Since 2011", "Since 2008", and "Custom range...". Below these filters, there are two input boxes for the years "2008" and "2009", with a minus sign between them, and a "Search" button. The main content area on the right displays search results. The first result is titled "[book] Blogs, Wikipedia, Second Life, and beyond: From production to produsage" by A. Bruns, dated 2008, with a link to books.google.com. The abstract mentions TIME's Person of the Year 2006 and AdAge's Advertising Agency of the Year 2007. The second result is titled "Learning to link with wikipedia" by D. Milne, dated 2008, with a link to dl.acm.org. The abstract describes a machine learning approach to cross-reference documents with Wikipedia. Both results include citation counts and links to related articles and all versions.

Google

allintitle: wikipedia

Scholar

About 800 results (0.03 sec)

Articles

Legal documents

Any time

Since 2012

Since 2011

Since 2008

Custom range...

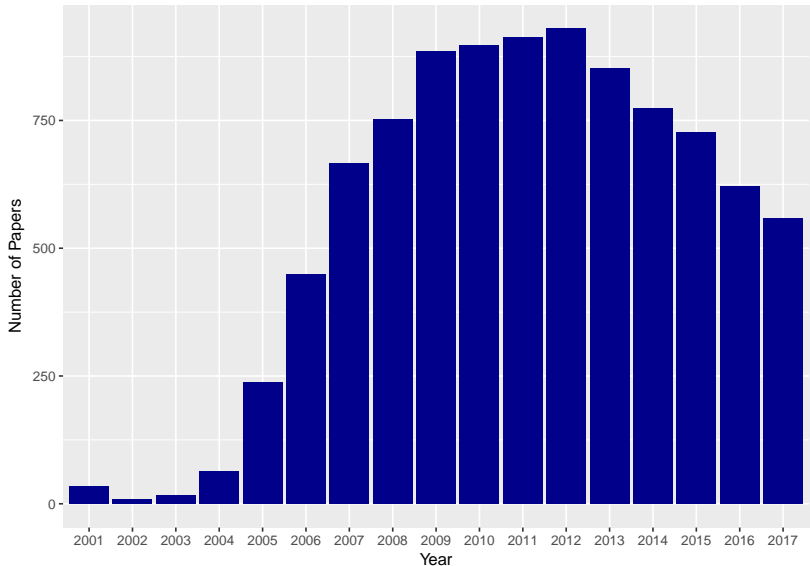
2008 — 2009

Search

[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



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

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

- 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!)



**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:

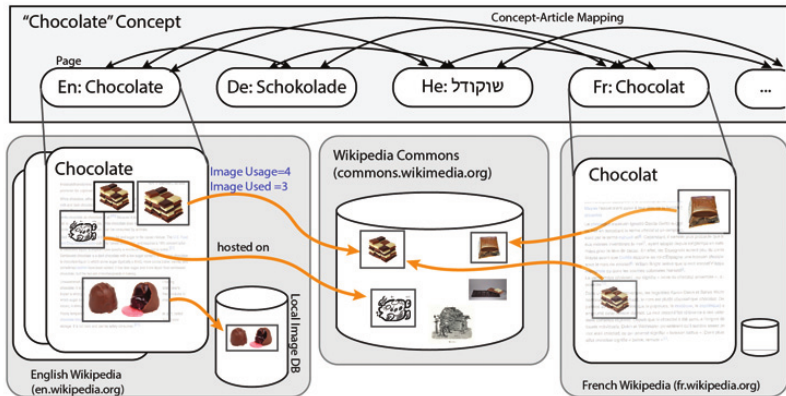
- Represent **important themes** from Wikipedia in the last year.
- Research that is likely to be of **interest** to Wikimedians.
- Research by people who are **not at Wikimania**.
- ...with a bias towards **peer-reviewed** publications

# Images & Media



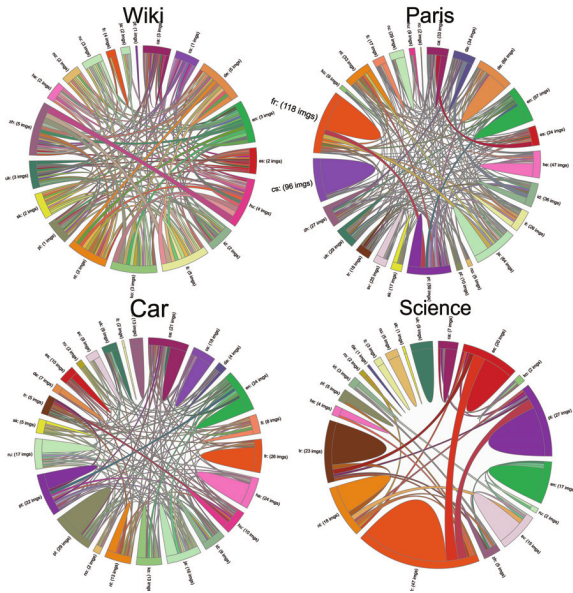
He, Shiqing, Allen Yilun Lin, Eytan Adar, and Brent Hecht. 2018. "The\_Tower\_of\_Babel.Jpg: Diversity of Visual Encyclopedic Knowledge across Wikipedia Language Editions." In *Proceedings of the Twelfth International AAI Conference on Web and Social Media (ICWSM 2018)*. Palo Alto, California: AAI. <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17903>.

# He et al. 2018: Image diversity

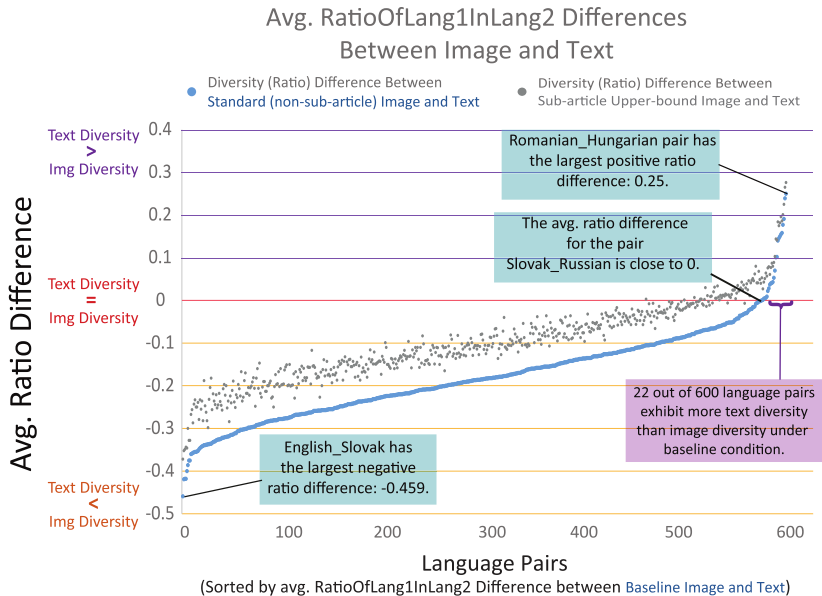




## He et al. 2018: Variance in image diversity across concepts



# He et al. 2018: Diversity in text and images



# 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>.



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

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

# 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>

## Lewoniewski et al.: Multilingual quality and popularity

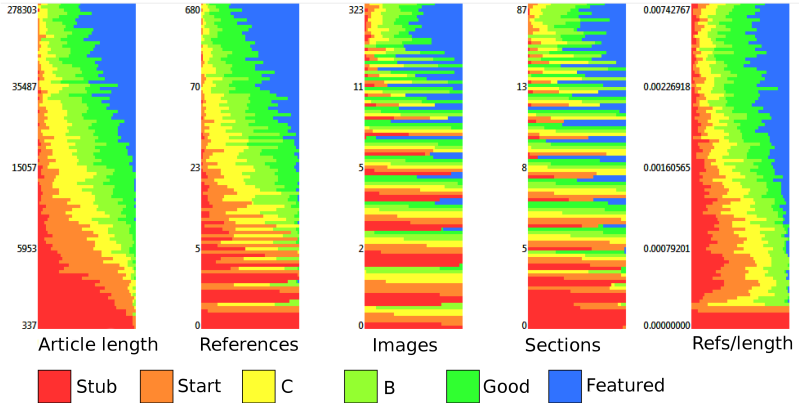
Construct a common quality metric to compare over 28 million articles in 44 language Wikipedias, based on:

- 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



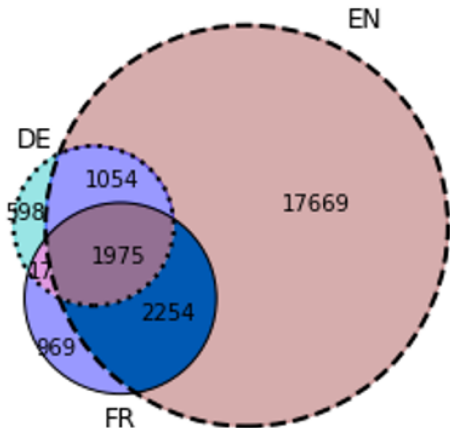
# Lewoniewski et al.: Multilingual quality and popularity comparison

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/>)





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

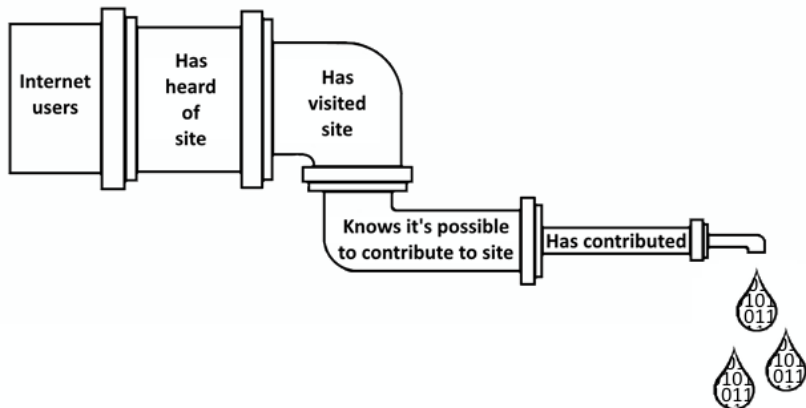
- The German Wikipedia's articles about albums and video games have the highest average quality score (among the 44 languages).
- However, its footballer biographies only rank 10 in quality.
- Quality and popularity (measured via pageviews) correlate positively - but more strongly for some topics and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

# **Nonparticipation: Who is not contributing?**

## 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



Nationally representative survey of 1512 US adults.

# 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 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.

# 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>.



**Table 1**

Corpus categories and number of studies in each sub-category.

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.: 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
<b>Information retrieval</b>													
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			
<b>Natural language processing</b>													
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			
<b>Ontology building</b>	1			12	1				2	6			
<b>Other corpus topics</b>				3					2	4			
<b>Total number of distinct studies</b>	<b>2</b>	<b>4</b>	<b>1</b>	<b>76</b>	<b>3</b>	<b>6</b>	<b>3</b>	<b>1</b>	<b>8</b>	<b>48</b>	<b>1</b>	<b>1</b>	<b>2</b>

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>)

## 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

