Dead Ends of Datafication. The Problem of Wikipedia Authorship

Krzysztof Gajewski
Institute of Literary Research
Polish Academy of Science

Wrocław, 15 November 2019

Table of Contents

The Problem of Authorship of Wikipedia

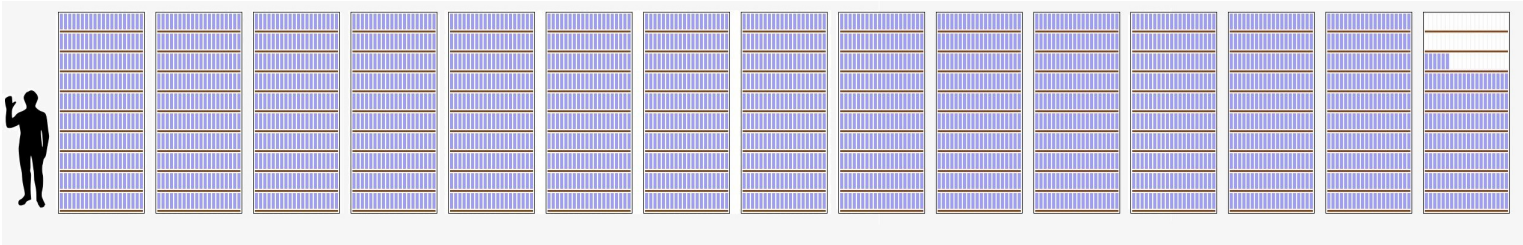
The Gang of 500 vs. The Anonymous Horde

Dead Ends of Datafication

Alternativ Metrics: Persistent Word View and Persistent Word Revision

The Problem of Authorship of Wikipedia

English Wikipedia Size in Volumes



Wikipedia pure text version in print (multimedia excluded)[1]

- 15 stacks
- 2946 volumes (Britannica size)
- 5.95 million articles (as of October 2019)
- 3.068 billion words (as of July 2016)
- Who wrote it?

[1] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

Problems with the Quality of the Wikipedia Content

- the bias of communication (Harold Innis)
- "geographical imbalance": "Global North" vs. "Global South"
- "racial bias"
- gender bias (Reagle, Rhue, 2011)
- political bias (Greenstein, Zhu, 2012; Greenstein, Zhu, Gu, 2016)
- systemic bias
- a way to make Wikipedia content more objective (Hansen, Berente, Lyytinen 2009).

An General Map of Biases of Polish Wikipedia

1. persons
   - history (Dynastia Pahlawi, Józef Piłsudski, Józef Stalin, Kult Józefa Piłsudskiego, Stalin's cult of personality, Stalinizm)
   - race (Biała odmiana człowieka, Czarna odmiana człowieka, Indianie, Rasa człowieka, White people, Żółta odmiana człowieka)
   - gender (Hipoteza o wyższości mężczyzn, Kobieta, LGBT)
2. events
   - colonialism (Historia odkryć geograficznych, Powstanie mahdystów)
   - holocaust (Pogrom w Jedwabnem)
   - others (II wojna światowa)
3. notions
   - communism (Komunizm, Związek Socjalistycznych Republik Radzieckich)
   - others (Sanacja, Teoria literatury)

Motivations of Cotributors

- "opposites attract" — editors are slightly more likely to contribute to articles which exhibit an opposite slant to their own[2]

[2] Shane Greenstein, Feng Zhu, and Yuan Gu, Ideological Segregation among Online Collaborators: Evidence from Wikipedians, Working Paper, Harvard Business School 2016.

The Gang of 500 vs. The Anonymous Horde

Jimmy Wales: The Gang of 500

Jimmy Wales:

- 73.4% of all the edits — 2% of the users (1400 people)
- 50% of all the edits — 0.7% of the users (524 people)[3]

[3] Swartz 2006

What Is an Edit?

Two types of edits:

1. uploading textual content
2. „wikisation" of a content already uploaded

Aaron Swartz: The Anonymous Horde

Top 10 contributors to „Alan Alda" entry

1. by edits: 7 registered vs. 3 anonymous
2. by letters added: 8 anonymous vs. 2 registered

Aaron Swartz: The Anonymous Horde

Apparent exceptions

1. translations from other language version of Wikipedia
2. plagiarism[4]

[4] Aaron Swartz, False Outliers, `http://www.aaronsw.com/weblog/writefp`

Dead Ends of Datafication

Dead Ends of Datafication

I am adding more exceptions

1. translations from other language version of Wikipedia ('Alkane')
2. plagiarism ('Atlas Shrugged')
3. content revitalizing ('Krupy (opad atmosferyczny)', en. Graupel)
4. paraphrase ('Wojna z terroryzmem', en. 'War on terror')

Who writes Wikipedia?

Three types of editors:

1. regular (long term)
2. occasional (short term)
3. bot (non-human)

Translations

- (cur | prev) 15:44, 21 October 2005 Physchim62 (talk | contribs) . . (32,560 bytes) (+498) . . (→Molecular geometry: insert images) (undo)
- (cur | prev) 15:26, 21 October 2005 Physchim62 (talk | contribs) . . (32,062 bytes) **(-1,810)** . . (→Higher alkanes: translation cleanup) (undo)
- (cur | prev) 15:02, 21 October 2005 Physchim62 (talk | contribs) m . . (33,872 bytes) (+3) . . (→Ethane: boldface edit) (undo)
- (cur | prev) 14:53, 21 October 2005 Physchim62 (talk | contribs) . . (33,869 bytes) (-251) . . (→Ethane: translation cleanup) (undo)
- (cur | prev) 13:59, 21 October 2005 Physchim62 (talk | contribs) . . (34,120 bytes) (-446) . . (→Molecular geometry: translation cleanup) (undo)
- (cur | prev) 13:41, 21 October 2005 Physchim62 (talk | contribs) . . (34,566 bytes) **(+7,876)** . . (→Molecular geometry: translation from de:Alkane) (undo)
- (cur | prev) 13:32, 21 October 2005 Physchim62 (talk | contribs) . . (26,690 bytes) **(+995)** . . (→Properties: Molecular geometry) (undo)
- (cur | prev) 21:02, 16 October 2005 196.203.36.148 (talk) . . (25,695 bytes) (-1) . . (corrected Propane (C4H10 -> C3H8)) (undo)
- (cur | prev) 11:16, 16 October 2005 Physchim62 (talk | contribs) . . (25,696 bytes) **(-627)** . . (→Alkanes in nature: translation cleanup) (undo)
- (cur | prev) 10:53, 16 October 2005 Physchim62 (talk | contribs) . . (26,323 bytes) (-55) . . (→Animals: translation cleanup) (undo)
- (cur | prev) 10:24, 16 October 2005 Physchim62 (talk | contribs) . . (26,378 bytes) (+25) . . (→Alkanes in nature: translation cleanup) (undo)
- (cur | prev) 09:59, 16 October 2005 Physchim62 (talk | contribs) . . (26,353 bytes) (-136) . . (→Alkanes in nature: translation cleanup) (undo)
- (cur | prev) 09:20, 16 October 2005 Physchim62 (talk | contribs) . . (26,489 bytes) **(-1,024)** . . (→Purification and use: translation cleanup) (undo)
- (cur | prev) 20:15, 15 October 2005 Physchim62 (talk | contribs) . . (27,513 bytes) (+113) . . (→Fungi and plants: image fix) (undo)
- (cur | prev) 19:50, 15 October 2005 Physchim62 (talk | contribs) . . (27,400 bytes) **(+5,724)** . . (→Alkanes in nature: from de:Alkane) (undo)
- (cur | prev) 19:23, 15 October 2005 Physchim62 (talk | contribs) . . (21,676 bytes) (+49) . . (→Purification and use: fix image) (undo)
- (cur | prev) 19:21, 15 October 2005 Physchim62 (talk | contribs) . . (21,627 bytes) **(+502)** . . (→Alkanes in nature: from de:Alkane) (undo)
- (cur | prev) 19:17, 15 October 2005 Physchim62 (talk | contribs) . . (21,125 bytes) **(+4,417)** . . (→Purification and use: from de:Alkane) (undo)
- (cur | prev) 18:54, 15 October 2005 Physchim62 (talk | contribs) . . (16,708 bytes) (+140) . . (→Occurance: adding headings for further translation) (undo)
- (cur | prev) 18:28, 15 October 2005 Physchim62 (talk | contribs) . . (16,568 bytes) (+52) . . (→Physical properties: add image) (undo)
- (cur | prev) 18:25, 15 October 2005 Physchim62 (talk | contribs) . . (16,516 bytes) (+176) . . (→Occurance: add images) (undo)
- (cur | prev) 17:57, 15 October 2005 Physchim62 (talk | contribs) . . (16,340 bytes) (-19) . . (→Occurance: Translation from de:Alkane) (undo)
- (cur | prev) 17:26, 15 October 2005 Physchim62 (talk | contribs) . . (16,359 bytes) **(+2,597)** . . (→[[Combustion]]: adding text) (undo)
- (cur | prev) 17:14, 15 October 2005 Physchim62 (talk | contribs) . . (13,762 bytes) **(+917)** . . (→Properties: translation from de:Alkane) (undo)
- (cur | prev) 16:53, 15 October 2005 Physchim62 (talk | contribs) . . (12,845 bytes) **(+2,692)** . . (→Physical properties: translation from de:Alkane) (undo)
- (cur | prev) 12:16, 15 October 2005 Physchim62 (talk | contribs) . . (10,153 bytes) (+2) . . (→Alkanes with branched carbon chains: fixed header) (undo)

The author

An exceprt from the history of the entry 'Alkane'

- comment on tranlation is not obligatory

Plagiarism

(undo)

- (cur | prev) 05:48, 26 May 2003 Vroman (talk | contribs) . . (26,153 bytes) (+171) . . *(added character description of Midas mulligan)* (undo)
- (cur | prev) 22:41, 21 March 2003 Ams80 (talk | contribs) m . . (25,982 bytes) (-60) . . *(Fixed links)* (undo)
- (cur | prev) 22:38, 21 March 2003 Ams80 (talk | contribs) m . . (26,042 bytes) (-23) . . *(Fixed links)* (undo)
- (cur | prev) 22:23, 21 March 2003 Ams80 (talk | contribs) m . . (26,065 bytes) (-136) . . *(Fixed links)* (undo)
- (cur | prev) 21:49, 21 March 2003 Ams80 (talk | contribs) m . . (26,201 bytes) (-91) . . *(Fixing links)* (undo)
- (cur | prev) 21:41, 21 March 2003 Ams80 (talk | contribs) m . . (26,292 bytes) (+10) . . *(Fixed link)* (undo)
- (cur | prev) 14:20, 21 March 2003 Ams80 (talk | contribs) . . (26,282 bytes) **(-1,698)** . . *(Fixing links)* (undo)
- (cur | prev) 20:19, 19 March 2003 CatherineMunro (talk | contribs) m . . (27,980 bytes) (+4) . . (undo)
- (cur | prev) 02:06, 15 March 2003 Ams80 (talk | contribs) m . . (27,976 bytes) (+30) . . *(Little formatting and grammar changes)* (undo)
- (cur | prev) 20:42, 13 March 2003 MartinHarper (talk | contribs) m . . (27,946 bytes) (-135) . . *(Merge Gwen Ives into bit on Hank Rearden)* (undo)
- (cur | prev) 17:54, 13 March 2003 CatherineMunro (talk | contribs) . . (28,081 bytes) **(-7,356)** . . *(more characters consolidated -- done now!)* (undo)
- (cur | prev) 09:10, 13 March 2003 CatherineMunro (talk | contribs) m . . (35,437 bytes) **(+8,035)** . . (undo)
- (cur | prev) 08:10, 13 March 2003 CatherineMunro (talk | contribs) m . . (27,402 bytes) **(+5,828)** . . *(more characters consolidated; work in progress)* (undo)
- (cur | prev) 05:45, 13 March 2003 CatherineMunro (talk | contribs) . . (21,574 bytes) **(+8,322)** . . *(more characters consolidated; work in progress)* (undo)
- (cur | prev) 00:34, 13 March 2003 CatherineMunro (talk | contribs) m . . (13,252 bytes) **(+1,484)** . . *(more characters consolidated; work in progress)* (undo)
- (cur | prev) 22:48, 13 March 2003 CatherineMunro (talk | contribs) m . . (11,768 bytes) **(+741)** . . (undo)
- (cur | prev) 22:37, 12 March 2003 CatherineMunro (talk | contribs) . . (11,027 bytes) **(+5,031)** . . *(more characters consolidated; work in progress)* (undo)
- (cur | prev) 21:39, 12 March 2003 CatherineMunro (talk | contribs) m . . (5,996 bytes) **(+1,793)** . . (undo)
- (cur | prev) 20:18, 12 March 2003 CatherineMunro (talk | contribs) m . . (4,203 bytes) (+460) . . (undo)
- (cur | prev) 15:51, 25 February 2002 Conversion script (talk | contribs) m . . (3,743 bytes) **(+1,760)** . . *(Automated conversion)* (undo)
- (cur | prev) 03:28, 27 February 2001 TimShell (talk | contribs) m . . (1,983 bytes) **(+1,983)**

Compare selected revisions

(newest | oldest) View (newer 500 | older 500) (20 | 50 | 100 | 250 | 500)

The author

An exceprt from the history of the entry 'Atlas Shrugged'

- the fact of plagiarism is not evident at all (the content has been copied from 3rd party sites)

Content revitalizing

- (bież. | poprz.) 00:24, 24 lip 2009 Xqbot (dyskusja | edycje) m . . (1205 bajtów) (+23) . . *(robot dodaje: no:Graupel; zmiany kosmetyczne)* (anuluj edycję)
- (bież. | poprz.) 04:46, 20 lut 2009 Xqbot (dyskusja | edycje) m . . (1182 bajty) (+20) . . *(robot dodaje: he:גראופל)* (anuluj edycję)
- (bież. | poprz.) 16:37, 11 sty 2009 DragonBot (dyskusja | edycje) m . . (1162 bajty) (+17) . . *(robot dodaje: sv:Trindsnö)* (anuluj edycję)
- (bież. | poprz.) 13:18, 6 lis 2008 Alexbot (dyskusja | edycje) m . . (1145 bajtów) (+17) . . *(robot dodaje: ko:싸락눈)* (anuluj edycję)
- (bież. | poprz.) 11:18, 26 sie 2008 84.223.92.146 (dyskusja) . . (1128 bajtów) (-7) . . *(sl.wiki)* (anuluj edycję)
- (bież. | poprz.) 03:30, 22 sie 2008 SpBot (dyskusja | edycje) m . . (1135 bajtów) (+20) . . *(robot dodaje: sl:Babje pšeno)* (anuluj edycję)
- (bież. | poprz.) 23:30, 11 sty 2008 AI~plwiki (dyskusja | edycje) m . . (1115 bajtów) (+121) . . *(Dodanie grafiki z commons)* (anuluj edycję)
- (bież. | poprz.) 13:55, 15 sie 2007 Loveless (dyskusja | edycje) m . . (994 bajty) (+11) . . *(robot dodaje: ja:霰)* (anuluj edycję)
- (bież. | poprz.) 14:42, 2 cze 2007 Micgryga (dyskusja | edycje) m . . (983 bajty) (+285) . . *(stub, +info)* (anuluj edycję)
- (bież. | poprz.) 23:19, 22 maj 2007 Holek.Bot (dyskusja | edycje) m . . (698 bajtów) (-6) . . *(Szablony ujednoznaczniające - poprawa)* (anuluj edycję)
- (bież. | poprz.) 13:58, 5 maj 2007 59.120.106.104 (dyskusja) . . (704 bajty) (+11) . . (anuluj edycję)
- (bież. | poprz.) 03:42, 1 mar 2007 Thijs!bot (dyskusja | edycje) m . . (693 bajty) (-7) . . *(robot poprawia: en:Graupel)* (anuluj edycję)
- (bież. | poprz.) 03:13, 25 sty 2007 K.J.Bot (dyskusja | edycje) m . . (700 bajtów) (+108) . . *(robot dodaje: eo, fi, fr, it, ru poprawia: en)* (anuluj edycję)
- (bież. | poprz.) 22:10, 9 gru 2006 83.24.121.149 (dyskusja) . . (592 bajty) (+114) . . *(→Zobacz też)* (anuluj edycję)
- (bież. | poprz.) 06:33, 29 mar 2006 Pcirrus (dyskusja | edycje) . . (478 bajtów) (-1) . . *(do podka opady)* (anuluj edycję)
- (bież. | poprz.) 08:07, 20 lut 2006 Roo72 (dyskusja | edycje) m . . (479 bajtów) (-46) . . (anuluj edycję)
- (bież. | poprz.) 07:53, 20 lut 2006 Pcirrus (dyskusja | edycje) m . . (525 bajtów) (+21) . . *(podkategoria chmury)* (anuluj edycję)
- (bież. | poprz.) 19:42, 25 sty 2006 Tsca.bot (dyskusja | edycje) . . (504 bajty) (+19) . . *(⇆ zmiana Kategorii:Meteorologia na Kategoria:Meteorologia i fizyka atmosfery)* (anuluj edycję)
- (bież. | poprz.) 09:31, 22 sty 2006 Beno (dyskusja | edycje) m . . (485 bajtów) (+9) . . (anuluj edycję)
- (bież. | poprz.) 03:33, 22 sty 2006 Roo72 (dyskusja | edycje) m . . (476 bajtów) (-29) . . (anuluj edycję)
- (bież. | poprz.) 03:26, 22 sty 2006 Kimbar (dyskusja | edycje) . . (505 bajtów) **(+505)** . . *(z historii krupy)*

The author

Porównaj wybrane wersje

An exceprt from the history of the entry 'Krupy'

- The link to the source is dead, since the source has already changed its name (the content has been copied from deleted Wikipedia entry)

Paraphrase

On 5th August of 2005 an Anonymous A added a sentence to the article ('Wojna z terroryzmem', en. 'War on terror')

> *"The 'war on terror' understood in this way has already caused over 25,000 civilian casualties."*

> *"Tak rozumiana „wojna z terroryzmem" pochłonęła już ponad 25 000 ofiar cywilnych"*[5]

[5] https://pl.wikipedia.org/w/index.php?diff=1321892&oldid=1173134&title=Wojna_z_terroryzmem&diffmode=visual

Paraphrase

On 29th August of 2005 an Anonymous B removed this sentence and added its paraphrase instead':

> *"They [scil. critics of US policy] also note the numerous civilian casualties (several thousand) caused by warfare"*
>
> *"Zwracają oni [scil. krytycy polityki USA] także uwagę na liczne ofiary cywilne (kilkanaście tysięcy) spowodowane działaniami wojennymi"*[6]

Quite falsely, the anonymous B will be granted authorhip of this sentence, Anonymous A will not appear on the list of authors of the entry.

[6] https://pl.wikipedia.org/w/index.php?diff=1321892&oldid=1173134&title=Wojna_z_terroryzmem&diffmode=visual

Alternativ Metrics: Persistent Word View and Persistent Word Revision

Persistent Word View

Persistent Word View (PWV) is based on

1. number of letters she input
2. the popularity of the content[7]

Top 10% most active editors generated 86% of Persistent Word View (Feb. 2006)

[7] Priedhorsky 2007

Persistent Word Revision

Persistent Word Revision (PWR) is

> *The sum total of subsequent revisions persisted by the words in a revision.*[8]

[8] Research:Content persistence

Persistent Word View vs. Persistent Word Revision

- PWV stresses a role of a reader
- PWR — editors decide of the value of the content

Works Cited

1. Content persistence, Wikimedia. Meta-wiki, https://meta.wikimedia.org/wiki/Research:Content_persistence
2. Shane Greenstein, Feng Zhu, and Yuan Gu, Ideological Segregation among Online Collaborators: Evidence from Wikipedians, Working Paper, Harvard Business School 2016.
3. Shane Greenstein, Feng Zhu, Is Wikipedia Biased?, Ąmerican Economic Review: Papers and Proceedings", 102, no. 3 (May 2012)
4. Sean Hansen, Nicholas Berente, Kalle Lyytinen Wikipedia, Critical Social Theory, and the Possibility of Rational Discourse The Information Society, Volume 25, Number 1, January 2009
5. Reid Priedhorsky, Jilin Chen, Shyong Lam, Katherine Panciera, Loren Terveen, John Riedl, Creating, Destroying, and Restoring Value in Wikipedia, w: Proceedings of the 2007 International ACM Conference on Supporting Group Work. red. Tom Gross, Kori Inkpen, Nowy Jork 2007.
6. Joseph Reagle, Lauren Rhue, Gender Bias in Wikipedia and Britannica, „International Journal of Communication" 5 (2011).
7. Aaron Swartz, False Outliers, http://www.aaronsw.com/weblog/writefp, 2006.
8. Aaron Swartz, Who Writes Wikipedia, http://www.aaronsw.com/weblog/whowriteswikipedia, 2006