Dead Ends of Datafication.
The Problem of Wikipedia Authorship

Krzysztof Gajewski
Institute of Literary Research
Polish Academy of Science

Wrocław, 15 November 2019
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?

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 Contributors

• "opposites attract" — editors are slightly more likely to contribute to articles which exhibit an opposite slant to their own\(^2\)

---
The Gang of 500 vs. The Anonymous Horde
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
Two types of edits:

1. uploading textual content
2. „wikisation” of a content already uploaded
Top 10 contributors to „Alan Alda” entry

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

1. translations from other language version of Wikipedia
2. plagiarism\textsuperscript{4}

\textsuperscript{4}Aaron Swartz, False Outliers, http://www.aaronsw.com/weblog/writefp
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)
An excerpt from the history of the entry 'Alkane'

- comment on translation is not obligatory
An excerpt 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

An excerpt 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)
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”

---

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. krypcy polityki USA] także uwagę na liczne ofiary cywilne (kilkanaście tysięcy) spowodowane działaniami wojennymi"\(^6\)

Quite falsely, the anonymous B will be granted authorship 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 (PWV) is based on

1. number of letters she input
2. the popularity of the content\textsuperscript{7}

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

\textsuperscript{7}Priedhorsky 2007
Persistent Word Revision (PWR) is

*The sum total of subsequent revisions persisted by the words in a revision.*

---

8Research: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
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