WIKIPEDIA AND THE PROBLEM OF AUTHORSHIP: AARON SWARTZ'S HYPOTHESIS

KRZYSZTOF GAJEWSKI
Polish Academy of Science, Poland

A traditional approach to the question of who is responsible for the most of the content of the Wikipedia, as represented by Jimmy Wales, one of the founders of the project, points out to the core of Wikipedia community — most active and most frequent users with the longest Wikipedia experience. However, an alternative proposal was formulated by Aaron Swartz who analyzed the amount of contribution to an article in terms of words added. He came to a conclusion that most of the content has been produced by occasional, fortuitous users that deliver long pieces of a text or even entire, developed entries.

The paper discusses some methods that could allow to an empirical verification of the Swartz's thesis, continuing the line of research he elaborated. In the final, theoretical part of the paper the conflict between two opposite authorship models, the Wales' and the Swartz's, will be interpreted in frame of communication theory. Wales's point will be confronted with James W. Carey's ritual view of communication, whereas Swartz's approach will lead to a new communication model. It will be called a conflict view of communication, following results of investigations of Dariusz Jemielniak who utters that the main motivation to contribute to the Wikipedia is a conflict between our believes and what we read in the Wikipedia.

KEYWORDS: *articipatory culture, Wikipedia, authorship, new media, communication theory, ritual view of communication*

A traditional model of communication contains three elements: a sender, a message, and a receiver. The message this paper is talking about is Wikipedia, an internet encyclopedia, controversial, but popular and widely used as a source of information. The receiver then is us, Wikipedia users. This includes not only occasional users accessing it during their free time for entertainment, but also professionals dealing in their job with information processing, such as journalists or researchers, from private companies, as well as state owned institutions. What about the third element of this communication process? Who is a sender?

An answer to this question may seem very simple to someone who has some knowledge about the mechanism of Wikipedia functioning. Wikipedia is created by thousands of internet users, since this site, as it declares, is "the free encyclopedia that anyone can edit." It is what we can read on the front page.

The Wiki-text is a logical consequence of the palimpsest technology, which develops it almost to perfection. It allows adding supplements, removing old content, inputting a new one, and at the same time preserving all the past versions of an article. Nothing is lost in Wikipedia, all the interventions of every contributor are saved. As one can read in the Wikipedia tutorials, you cannot change the content of Wikipedia, you can only supply a new one.

Nevertheless, a small percentage of Wikipedia readers have ever edited Wikipedia, or even know someone who has done it. Where are they? Who is writing Wikipedia?

The Gang of 500 versus the Anonymous Horde

This question attracted the attention of Aaron Swartz, an Internet activist and a devoted Wikipedian. He contrasted two theories. The first one, called "The Gang of 500" theory is supported by a lot of Wikipedians, among them Jimmy Wales, one of the founders of Wikipedia. Swartz quotes passages from Wales' researches stating that

> „over 50% of all the edits are done by just 0.7% of the users … 524 people. … And in fact the most active 2%, which is 1400 people, have done 73.4% of all the edits." (Swartz 2006)

One edit is equal to one click of „Save" button, what makes sending to Wikipedia servers a new version of an article. The interventions of an editor, however, can be of various sorts. We can divide them all into two fairly well separated types:

1. uploading textual content;
2. „wikisation" of a content already uploaded. The term of wikisation refers to every edit that is aimed at making a content already present in the article comply in a possible strict way with all the standards of the Wikipedia project.

An actual intervention can be the mix of the two types above mentioned. Still, it is always possible to determine a precise amount of both types of interventions in one edit submitted by a Wikipedia editor. One can add some new facts to an existing entry, but at the same time make some grammatical and editorial corrections. Then, if she clicks "Save," all the input provided during this session is sent to servers as a one edit, independently of its being a 10 000 chars text or an action of removing one coma. One click on the "Save" button counts as one edit.

Following this line of thought Swartz advanced an alternative way to determine the contribution of users in a more precise and adequate way. He decided to count the amount of text each user contributed to the final version of an entry. He examined a random article, which happened to be an entry about Alan Alda, an American actor. Swartz compared the results obtained in both ways: Wales' and his own. When he counted the edits, he found that in top 10 of users contributing to this article 7 were registered users, 2 of them very active Wikipedians. Nevertheless, when he took as a unit of added value a letter in the final version of an entry, the ranking list has dramatically changed: only 2 of 10 top users were registered ones, all the others were occasional users who didn't contribute to the project much, apart from the entry in question. Therefore, a hypothesis arose suggesting that the real authors of Wikipedia are an anonymous mass of fortuitous, occasional editors. Swartz names this hypothesis "The Anonymous Horde" theory.

An analysis of one arbitrarily chosen entry cannot prove or disprove any thesis. Swartz executed the algorithm on 200 articles (Swartz). The thesis expressed above got confirmed, save for a few exceptions, which, as it sometimes happens, finally, at a closer look, proved the rule as well. Nonetheless, it would be useful to describe them, since they seem to illustrate general tendencies. To put it briefly, he came across a few cases that apparently contradicted the thesis of the anonymous horde, but all of them fell into one of the two groups:

1. translations
2. plagiarism

In the case of translation an article happened to be a translated version of an article from another language version of Wikipedia. Plagiarism includes cases where an article is compiled out of the copy-and-pasted content of third-part websites. Whereas a translation from other language Wikipedia versions is accepted, even if not appreciated (and must always be clearly stated in the history of articles), plagiarism is illegal, unless the works belong to the public domain.

In his research sample Swartz hasn't found any article that was created mostly by one user. The 'horde' means that the authors are not only occasional and anonymous, but also very numerous, that every article, even a very long and complicated one, is the result of the work of dozens of people supplying few sentences each. Here we can apply the notion of crowdsourcing, a wisdom of crowds, a smart mob, swarm intelligence (SI), or collective intelligence (Lévy 1994), as far as a kind of organizational aspect is concerned, or a long tail (Anderson 2006), so as to grasp the phenomenon from the point of view of economy and organization. Contrarily, according to Jimbo Wales' view, Wikipedia would function in a much more traditional way, as a centralized organization, with a core team producing most of the added value of the site — these are the conclusions Swartz is drawing in his publication.

In my opinion, the difference between Wales' and Swartz's results derives from the very definition of authorship. According to Swartz, counting should be done by the letters/words added to the final version of the text. Wales is counting all the interventions. One could probably lean towards Swartz's definition, as far as authorship is concerned, understood as a value added by a user. However, one may suppose that occasional authors of texts wouldn't be encouraged to deliver their contribution unless articles are kept clean, readable, and encyclopedic by the Wikipedia editors.

Alternatives: Persistent Word View and Persistent Word Revision

The difference between the value of incidental authors' and editors' input becomes less evident when one takes into consideration Reid Priedhorsky and his team's research. As a unit of value added by a user they propose the persistent word view (PWV). This notion is based on the idea of a web site view, but it is more precise and it concerns not the whole site, but each single word separately. The contribution of an editor is measured not just by the number of letters she inputs, but also by the popularity of its content (Priedhorsky 2007). This methodological approach is more sophisticated than Swartz's, based only on the amount of the text introduced. Priedhorsky's parameter exemplifies the perspective of a reader, since the value of a content is proportional to their use for the audience. The more popular and viewed the word is, the bigger its value grows, according to Priedhorsky's approach.

Priedhorsky's results confirmed the thesis of Jimmy Wales about The Gang of 500. As of February 2006, top 10% most active editors generated 86% of Persistent Word Views. Nonetheless, one must remember that it doesn't mean they contributed to 86% of the whole textual content. Their textual input was probably smaller, but it covered the most popular topics.

A similar logic stands behind another wiki research methodological proposal, which is aimed at measuring the quality of a content. Persistent Word Revision (PWR):

> The sum total of subsequent revisions persisted by the words
> in a revision. (Research: Content persistence)

In this case the value of a word is increasing after every edition of an entry, providing the word hasn't been removed. This approach was inspired by examinations that showed that low-quality input doesn't last much time and is always sooner or later removed. The whole idea is implied by one of the principles of Wikipedia, which is a maxim: "publish first, edit later."

While the Persistent Word View concept stresses the role of a reader, since the value of a content is based on its popularity with the public, the Persistent Word Revision factor appeals to the importance of editors' interventions, and it is their activity that determines and corroborates the value of a contribution. View approach and Revision approach make a perfect symmetry in exactly the same way as a reader and an author.

However, both of these methodological tools do not seem to hit the point of our research problem. There is no doubt that Persistent Word View and Persistent Word Revision are more precise as far as the estimation of the general use and quality of a content is concerned. Nevertheless, the topic of our investigation is a question of authorship, which is the simple formulation of an expression, or a sentence, and storing it in a written form. The question of quality or popularity of a content must be distinguished from the question of its authorship.

There are several other works on the model of authorship that we confront in the case of Wikipedia, explaining it with a notion of an aggregate author (Jordan 2007) or situating it against the background of the European literary tradition of a romantic author figure (Chon 2012).

Jordan starts his analysis with a description of the Seigenthaller controversy. A Wikipedia article on an American journalist contained almost only false information and actually was a libel created by a Wiki vandal. However, it stayed online four months, despite Seigenthaller's efforts to remove it. In spite of that Jordan is not blaming Wikipedia for irresponsibly circulating false information. He takes this controversy as the cost for the unprecedented opportunity that Wikipedia provides to us. The online encyclopedia allows for tracing processes of meaning construction that theorists of semiotics and deconstruction have been describing for the last few decades. The electronic platform makes it more visible and accessible to empirical verification. It proposes a new model of authorship, an aggregate author that consists of

> „the interaction between writer and reader" (Jordan 2007: 165)

Margaret Chon describes "the author effect," which consists of two aspects of authorship: a genius creating new ideas, and an arbiter, an authority, who is able to translate the individual experience of a genius to the mass, "authorizing" them to participate in it. She investigates how a collective type of authorship inscribes itself into a romantic model. One should remark that in Wikipedia we deal with exactly the same two types of contributors: creators (authors) and editors (arbiters).

Verification of Swartz's thesis and Problems Arising

Swartz formulated his thesis 10 years ago and since then it has got neither confirmation, nor refutation. It may seem surprising, however, as I am going to show, it is not so much, given the complexity of the problem.

As we have already seen, it is not clear what authorship is at all. We don't agree whether we should found it on the amount of editions, a textual contribution, or a view of contribution. One shouldn't forget that Wikipedia is not only a text, but also links, bibliography, data formatted in tables, sound files, images and other sorts of graphical materials, like diagrams and schemes. When talking about the authorship of Wikipedia, one shouldn't be confined to textual bias, ignoring other kinds of creation. Therefore, my research proposal is to count contribution to Wikipedia not just with the bare text added, thus following in

Swartz' footsteps, or simply regressing to Wales' "brutal edits" metrics, but to try to take into consideration both the authors' and editors' activities. The idea is very simple and it is supported by the Wikipedia interface, providing the value of the size of every article and every user contribution in bytes. This is definitely not the perfect way of measuring article value for a reader, since there are several "invisible" chars. Most of them are aimed at formatting a layout: adding styles, creating headings, lists, tables, etc. Some serve for including interlinks, images, or other audiovisual content. There is also a templates system, some of them only temporary, serving as an annotation for future editors. Actually, the template caused a serious problem as regards the question of counting user contribution. It is based on the idea of transclusion, i. e. inclusion of the content of one site (a template site) to another site (for instance, a Wikipedia entry; see Wikipedia: Transclusion). It means that an editor writes just the name of a template along with some parameters in the input box and this text in substituted by the whole content of a template. Therefore a software amplifies the input of the editor by way of a mechanism of macro-definition.

Shortly, every Wiki entry contains a smaller or bigger quantity of Wiki MarkUp language designed not for a human reader of Wikipedia, but for the internal engine of MediaWiki, the software every language version of Wikipedia is based on. In my opinion, the Wiki MarkUp syntax should also be counted as a contribution, for it helps a human reader to read and better understand the article. Taking into consideration a contribution of the editors, I still differ from Wales in that he counts not the size of input provided by way of entry edits, but the number of edits. The difference is like that between calculating the total weight of all the parcels or counting the number of parcels.

The Double Face of Wiki Text

For the need of this paper I put together a sample of 30 entries from the Polish Wikipedia (by using an option "Random article" from the site navigation). For each entry I noted its total size in the Wikipedia data base such as it is displayed in the history of an entry. This figure should comprise the raw code of an article in the Wiki Markup syntax. This is an author/editor contribution in the original form, before it is parsed by the Media Wiki engine. Throughout this paper I will call it an input version of the Wikipedia article (IVA). The volume of this output was also recorded. The output is the final form of an input having undergone parsing according to the Wiki syntax rules. This is what a Wikipedia reader sees. I will refer to this entity as an output version of article (OVA)[1]. They are the two faces of a Wiki-text and they resemble a couple of notions from IT terminology: back-end and front-end. Back-end is the side of a site editor while front-end is the end user side, the final effect of editor work.

It is worth mentioning that this situation is unusual in human language communication. In oral speech, in writing or in print, what the sender of a message is creating as material correlate of the message is an entity perceived directly by the receiver of the message, This is due to a fact that all language communication before the age of automatic computing was intended for a human listener/reader, while in Computer Mediated Communication (CMC) there is some language content directed to a non-personal receiver, which is a Turing machine, popularly known as a computer. This is why some interfaces of CMC are now WYSIWYG type. WYSIWYG is an abbreviation of the sentence "What you see is what you get" and characterizes such media as writing and the printing press. What a writer writes is what a reader sees. There is no writer version that is separate from the reader version of a message. The same idea stays behind such text processors as Open Office, Libre Office, or Microsoft Office. They all are WYSIWYG, since what a user sees on the screen she will see on the printed form of a document as well. WYSIWYG interfaces are much easier to operate than those that are not WYSIWYG, but they are limited in functionalities. The Wiki-text, like HTML and LaTeX, is not WYSIWYG, and that is why it has both IVA and OVA[2].

The whole thing is, however, more complex. A lot of Wikipedia entries enclose pictures and other multimedia content. This is mentioned in IVA in the form of a link to a file, whereas in OVA we basically see or hear the content of the multimedia file. Consequently, we count it in IVA not as a content, but as a link to a content[3]. In the case of OVA, we ignore it for the sake of simplicity. On the one hand, there is no doubt that

[1] For the need of this paper, I presumed that OVA doesn't include the title of a Wikipedia article, nor a categorisation list, nor an automatically generated table of contents. It contains captions of images and bibliography.

[2] Even though a visual editor that allows editing in the WYSIWYG mode appeared in 2013 in the Polish Wikipedia (Kronika polskojęzycznej Wikipedii). Still, it is only a friendly graphical editor of a code which is not WYSIWYG by itself.

[3] A link to an image is about 100 bytes, while an image can easily be a thousand times bigger in terms of bytes.

if a photo is used in an article, the author of the photo is a co-author of the article, even though she doesn't have any idea about this fact. But, on the other hand, it would be difficult to compare a textual contribution and a visual contribution. For this reason I have decided to ignore the audiovisual contribution and to stick to a strictly textual one.

But even with this constraint, it is often not obvious what should be counted to OVA, among others, because of a text generated by templates. Every Wikipedia template was created by a human being, but every template can be and is applied many times. On the other hand, the text generated by a template makes integral part of OVA from the point of view of a Wikipedia reader — its receiver. For the sake of this experiment, I take for granted any text produced in this way.

Concise as a Wikipedia Article

The first surprise was related to the question of entries size. They seemed unexpectedly tiny in comparison to what we figure out as the size of a typical Wikipedia article. Most of them were smaller than four lines in a standard full-screen browser window. It was sure then that there was really not much space for several authors. The average volume of an entry (IVA) was 3960 characters, but as for clean textual content (OVA), it was only about 1961 chars, which is about 50% of IVA. It is not a lot, as it may seem, since OVA still contains a lot of 'empty' or almost 'empty' words like table labels (present in the form of infobox in a great bulk of Wikipedia articles), chapter titles (eg. "Bibliography") and other kinds of metatextuality, to use Gérard Genette's term (Genette 1997).

As for the actual size of articles I have identified three stages of the article volume, according to the amount of lines in a standard, full-screen window size:

Table 1. The size of articles from a research sample

Size of articles		Number of articles	Percentage of articles
I. developed	more than 10 lines	6	20%
II. somehow developed	4-10 lines	12	40%
III. not developed	less than 4 lines	12	40%

The average (IVA) size of the Polish Wikipedia article was, as of February 2014, 2718 bytes. 2718 chars is about 1 kB less than our result, 3960. Accordingly, the average number of edits for a Polish Wikipedia article was 28.2. The average article size has been growing steadily since the beginning of the Polish Wikipedia, starting from 537 bytes in October 2001 (Wikipedia Statistics Polish). Current statistics are not accessible yet, but it is very probable that this research sample mirrors the real Wikipedia characteristics to a fairly accurate degree.

The Authors

Only 1 out of 5 articles (6 of them) from the sample exceeded four lines of text (in default screen resolution, default browsers settings). The modest length of an article from the sample implied a limited number of contributors. In most of the cases the list of authors didn't exceed two. I then distinguished two groups of authors insofar as the size of their contribution in OVA is concerned.

The 1st Author

11 articles, that is more than 1/3 of the whole sample, have a short history of editions and only 1 human author responsible for the whole content of the article, amplified in 9 of the 11 cases by a bot, a software that edits Wikipedia pages in automated way, making standard, repetitive corrections or supplements.

In two other cases ("Anthaxia attenuata", "Klaffer am Hochficht") the entry history consisted exactly of 1 edition, and the creator of the entry was at the same time its only editor. In both cases there were registered active, keen Wikipedians. The latter entry, however, consisted of just one line (179 chars), if we do not count the info-box that added 372 chars to the whole size of OVA. By coincidence, the author of this one-line entry, Cojan, happened to be responsible for another article in the research sample.

The set of all 1st authors in our sample of 30 articles has only 29 members, since one of them happened to be the author of 2 articles. Apart from "Klaffer am Hochficht" Cojan created "Oberwiera", also a one-line entry that he was the only human contributor of. Cojan is one of the most active Polish Wikipedians and he is ranked among Wikipedians on the 21st position insofar as the amount of edits is concerned (Wikipedia: Najpłodniejsi wikipedyści/2016-01-01/bajty).

At a first glance one must admit that, contrary to the Anonymous Horde hypothesis, only two of the most prolific contributors to the articles of the sample happened to be anonymous, what in Wikipedia software is signed with an IP number.

Also, two of the 1st authors happened to be a bot. Tsca.bot was the author of two sentences entry "Wyganki," referring to a village in Poland. This entry was generated in an automated way by a user operating Tsca.bot. This kind of bot imports data from an external database, for instance, that provided by the Central Statistical Office, and in a mechanical way, and, according to a specific template, it creates a series of entries. Tsca.bot is specialized in geographical content and it has already generated series of articles, such as the municipalities in Denmark or the cities in Italy. The results articles created by bot are called stubs — small entries, containing few basic facts. Another one was MalarzBOT, the most active Polish Wikipedian, as for the number of editions. MalarzBOT happened to be the 2nd author too in a few cases.

The practice of creating new entries with the help of a bot have become popular in Wikipedia. One of the Wikibots, Lsjbot, operated by Swedish Wikipedian Sverker Johansson, brought about 2.7 million entries as of 2014 [Lsjbot]. Two thirds of these entries belong to Cebuano Wikipedia, one third to a Swedish one, making them, respectively, rank 3 and 2 of world Wikipedias' ranking based on the number of articles [List of Wikipedias 2016]. The Lsjbot, according to his operator, is devoted to creating articles on all living beings, especially birds and fungal species.

The 2nd Author

Only in 8 out of 30 articles the second author and the other ones added some factual content or bibliographical sources. All the other 20 cases involved only redaction and/or wikisation, such as attributing categories, including info-boxes, etc., with no textual contribution.

Of 8 factual contributions 2 were due to unregistered, anonymous users, 6 to registered Wikipedians.

Of these 20 cases of contributions half was made by bots, mostly by MalarzBOT.

Among the 2nd authors one human user appeared twice — Lowdown, also one of the most active Polish Wikipedians. One of his contributions was just a wikisation, but another one involved a factual amplification.

Table 2. The characteristics of 1st and 2nd authors of articles from the research sample

	all	registered	anonymous	bot
1st authors	29	25 (86%)	2 (7%)	2 (7%)
2nd authors	21	14 (67%)	4 (19%)	3 (14%)

Copied Content Problem

Several problems arise when trying to examine the authorship of a Wikipedia entry. In particular, we cannot display the contribution of a user in a simple way. We have to infer about it on the basis of other facts. That is not always evident. What we can access directly is a list of succeeding versions or revisions of an article and the nick of a person who saved each of them. There is also the option of comparing every couple of revisions in two columns showing differences on the level of a single paragraph. But even when we finally attribute a contribution to a user, that is not the end of the job. As we know from Swartz, a text input by a Wikipedian is not always her or his own original content. Swartz enumerated two exceptions from this case. The first one was a translation form another language version of Wikipedia. This is completely legal and in accordance with the rules of the project as long as it is overtly stated in the history of an article. The second exception was a copied content, which mostly involves plagiarism, sometimes unwitting and involuntary, and it is off course illegal.

During the examination of the research sample another class of exception emerged. This was a case of copying content from a Wikipedia article whose content was split into two articles or served as the basis for a new article. Let us take a look at this option in the following example. The entry, "Powstanie Kantonalistów" ("Cantonal Revolution") seemed to be created by a user, Diogenes2007, who made it out at 19:02 on 14 Feb 2008 as a simple redirection to another entry, "Rewolucja naftowa" ("Petrol Revolution"). Nonetheless, when we follow the history of "Rewolucja naftowa", we realize that the whole thing looks somewhat different. The entry entitled "Powstanie Kantonalistów" was brought about by a user Seksa on 17 July 2005. Three years later, on 14 Feb 2008, at 19:02, the user Diogenes2007 basically changed its name to "Rewolucja naftowa" and created another brand-new article "Powstanie Kantonalistów" and put into it a redirection to the old entry. So the article "Powstanie Kantonalistów" simply forked into two new articles, which started their own, independent life. Thus, the question about the identity of a Wikipedia article arises. We cannot rely simply on the name of an article, since it can undergo changes.

As we have seen in the case of "Powstanie Kantonalistów," almost the whole content seems to be created by a single user, Diogenes2007. However, one can figure out a case when the content of an old entry is used for the creation of a new one, and a user that creates the new entry and copies the content from the old one will appear as the creator of this content. Hence, to the list of situations enumerated by Swartz, which apparently contradicts the Anonymous Horde thesis, such as translation or plagiarism, one must add a case of a text copied from an older version of a Wikipedia entry, when changing the name of an entry and moving its content (partly or wholly) to a new one.

Such was the fascinating story of an entry "Krupy (powiat sokołowski)," referring to a village in Sokolow county. This entry was originally created at 8:51am, on September 11, 2002, most probably under the title "Krupy." A common noun *krupy* refers in Polish to the weather phenomenon of a graupel (soft hail, snow pellets). In its first version, created integrally by Sławojar, this entry contained a two line definition of what graupel is and a list of other similar phenomena, such as rain, drizzle, snow, or hail. When following the history of the entry, we see at a moment that one revision amplifies this entry with a new, parasitic entry in the same article. At 1:20 pm, Jan 16 2006, below the definition of a graupel, an anonymous user included a four sentence description of Krupy, a village in Sokolow powiat. In the next revision, 6 minutes later, she/he removed the definition of Krupy as a weather phenomenon. This content is now lost forever, because six days later, at 1:26 am, Jan 22, 2006, a new entry was brought about by Kimbar under the title "Krupy (opad atmosferyczny)" (Graupel (atmospheric precipitation)). The entry contained the content added by Kimbar, and he would be gratified as the author of a content by a research software. We saw, however, that this text was the contribution of Sławojar.

Another similar exception is probably common, namely it happens when an editor B paraphrases a sentence added by a previous editor A, correcting style or reordering an article. In such a case the contribution of the editor A will not be counted to his account, but all his work will be attributed to the editor B, whose role was just proofreading. Let's take an example. At 10:34 am, on 5 August 2005, an anonymous user A added a sentence to the entry "Wojna z terroryzmem" ("War on terror"). The sentence was:

> "The so-called 'war on terrorism' caused 25 000 civilian deaths"[4]
> (Wojna z terroryzmem)

Two weeks later, at 6:01pm, on 29 August 2005 another anonymous user B removed this whole sentence from the entry, and in another place of the article added a sentence:

> "They [opponents of USA policy] also point out the numerous civilian casualties (several thousands) due to military operations."[5]

User B described her/his contribution as NPOV-sation, which means making it follow the rule of Neutral Point of View. User B removed an undocumented figure and put a more general expression ("several thousands" instead of 25 000). B also left a note with a request for a source for the figure quoted. How did both edits contribute to the final version of the entry in terms of a text belonging to the latest version of an entry? Well, user A will be excluded from the group of authors. It doesn't, however, make much sense, since B will be counted with the contribution of a sentence of which she/he was only a redactor, a co-author at best.

This is one of the reasons why the authorship problem is not easy to solve with automated software and it needs human control.

Conclusions. Ritual vs. Conflict View of Communication

The results of the investigation undertaken have not confirmed Aaron Swartz's hypothesis, that is, finding the authors of Wikipedia content in a "long tail" of dispersed, anonymous users. On the contrary, among the main authors of articles from the research sample only 2 out of 29 (7%) happened to be anonymous. Limiting statistics to main authors, the contribution of registered users in terms of OVA exceeded 90%, whereas anonymous contributions was about 7% (the rest was the work of bots).

In my opinion, these conclusions by no means contradict Swartz' hypothesis. The main limit of the research undertaken is the small size

[4] „Tak rozumiana "wojna z terroryzmem" pochłonęła już ponad 25 000 ofiar cywilnych."

[5] „Zwracają oni [krytycy polityki USA] także uwagę na liczne ofiary cywilne (kilkanaście tysięcy) spowodowane działaniami wojennymi."

of the research sample. This is due to the amount of human work that is necessary to precisely follow the history of a Wikipedia article revisions. As I have shown, this examination cannot be fully automatized and human intervention is often necessary. Then, the answer to the question of authorship of Wikipedia is not resolvable by way of the "brutal force" of quick calculations. Probably the solution lies in researching larger and more developed articles, with a longer history, leaving place for more intensive creative contribution. Only 4 (13%) articles from the research sample had more than 50 revisions. Most probably in the case of thousands of uniform articles, like the ones referring to cities, sportsmen, or biological species, the cooperation model is more centralized than in articles that, thanks to their volume, allow for more advanced contributors cooperation.

The Anonymous Horde theory inscribes itself into the popular knowledge about crowdsourcing or wikinomy: a new network organization that lacks a central management and reveals self-organizational capabilities. Quite the contrary, Gang of 500 theory seems to come back to the traditional concept of a centralized system and the neoliberal Pareto principle. We can find an analogy to this opposition in the theory of communication, Jimmy Wales Gang of 500 theory stresses community values. Participants share the results of their unpaid job in exchange of all the profits they can have out of participating in the community they identify with. This picture reveals some similarities to James Carey's ritual model of communication:

> "In a ritual definition, communication is linked to terms such as 'sharing,' 'participation,' 'association,' 'fellowship,' and 'the possession of a common faith.' This definition exploits the ancient identity and common roots of the terms 'commonness,' 'communion,' 'community,' and 'communication.' A ritual view of communication is directed not toward the extension of messages in space but toward the maintenance of society in time; not the act of imparting information but the representation of shared beliefs." (Carey 2009: 15)

Swartz's hypothesis of the Anonymous Horde doesn't fit to this vision, since an anonymous, occasional user who doesn't belong to the Wikipedia community will not identify with it and will not be looking for community values. The motivation of this kind of participant seems be better described by the proposal of Dariusz Jemielniak, who advocates a view one could call a conflict model of communication. According to him, one of the strongest impulses to add and edit the content of Wikipedia is a disagreement between what one can read on Wikipedia and what one thinks (Jemielniak 2006: 124). A fundamental rule of human life in the epoch of social networks is: I cannot go to bed because someone is wrong on the Internet.

Research Sample

Alex MacDowall[6]
Andrzej Ekiert
Anthaxia attenuata
Aurora (telenowela)
Bitwa pod Olszanicą
Bradford (Ohio)
Chelmsley Wood
Droga krajowa B5 (Niemcy)
Droga krajowa nr 471 (Węgry)
FSO Polonez Analog
(Get A) Grip (On Yourself)
Gmina Czarnylas
Hrabstwo White (Tennessee)
Jerzy Panek (polityk)
Klaffer am Hochficht
Kościół św. Mikołaja w Wilnie
Krupy (powiat sokołowski)
Most Królowej Jadwigi w Poznaniu
Oberwiera
Park Narodowy Ałtaj-Tawanbogd
Podróż na Tajemniczą Wyspę
Powstanie Kantonalistów
Praszywe (Dolina Łatana)
Rodrigo Oliveira de Bittencourt
San Roque (Mariany Północne)
Scottish Premier League (2002/2003)
Strange Frontier
Walenty Foryś
Wojna z terroryzmem
Wyganki

[6] All the articles have been researched in their versions of May 30, 2016.

References

Anderson Ch. 2006, *The Long Tail: Why the Future of Business Is Selling Less of More*. New York: Hyperion.

Carey J. W. 2009. *Communication as Culture: Essays on Media and Society*. New York: Routledge.

Chon M. 2012. The Romantic Collective Author, *Vanderbilt Journal of Entertainment & Technology Law.* Summer, Volume 14, Issue 4, 829-849.

Genette, G. 1997. *Palimpsests : literature in the second degree,* Lincoln: University of Nebraska Press.

Jemielniak, D. 2013. *Życie wirtualnych dzikich Netnografia Wikipedii, największego projektu współtworzonego przez ludzi,* Warsaw: Poltext.

Jordan, S. T. 2007. The Problem of the Aggregate Author, *International Journal of the Book.* Volume 4, Issue 4, 161-167.

Lévy P. 1994. *L'intelligence collective. Pour une anthropologie du cyberespace*. Paris : La Découverte.

List of Wikipedias. 2016. http://wikistats.wmflabs.org/display.php?t=wp

Lsjbot. (n.d.) https://en.wikipedia.org/wiki/Lsjbot

Kronika polskojęzycznej Wikipedii. (n.d.) https://pl.wikipedia.org/wiki/Wikipedia: Kronika_polskoj%C4%99zycznej_Wikipedii.

Priedhorsky R. et al. 2007. *Creating, Destroying, and Restoring Value in Wikipedia*. In: Gross T. (ed.) *Proceedings of the 2007 International ACM Conference on Supporting Group Work,* New York, N.Y. : Association for Computing Machinery.

Research:Content persistence. (n.d.) https://meta.wikimedia.org/wiki/Research: Content_persistence

Swartz, A. 2006. Who Writes Wikipedia. (n.d.) http://www.aaronsw.com/weblog/whowriteswikipedia

Swartz, A. nd. "Who Writes Wikipedia?" (Swartz 2006), http://www.aaronsw.com/2002/whowriteswikipedia/swartz2006

Wikipedia: Najpłodniejsi wikipedyści/2016-01-01/bajty. (n.d.) https://pl.wikipedia.org/wiki/Wikipedia:Najp%C5%82odniejsi_wikipedy%C5%9Bci/2016-01-01/bajty

Wikipedia Statistics Polish. (n.d.) https://stats.wikimedia.org/EN/TablesWikipediaPL.htm

Wikipedia: Transclusion. (n.d.) https://en.wikipedia.org/wiki/Wikipedia: Transclusion

Wojna z terroryzmem, (n.d.) https://pl.wikipedia.org/wiki/Wojna_z_terroryzmem