Review

From Who and What to How and Why – The Future of Online Encyclopaedias

Nenad Prelog
University of Zadar
dr. Franje Tuđmana 24i, Zadar, Croatia
nenad@edemokracija.hr

Domagoj Bebić
Faculty of Political Science, University of Zagreb
Lepušićeva 6, Zagreb, Croatia
domagoj@edemokracija.hr

Summary

This paper discusses changes in the structure of knowledge, the increasing demand for abilities of search and retrieval, assessment and evaluation, organization and creative use of relevant information. It is a review of the topic intended to serve as a basis for further research aimed at answering the many questions that arise from this paper. Today, users are finding content through search engines. This requires a different approach to the organization of encyclopaedias and other lexicographical issues. All can be found, but it is also important to know where and how to look for it.

The conducted research centered around the quality of "coverage" of some, in Croatia well known, lexical units, in different Wikipedias: four regional languages and editions (Croatian, Bosnian, Serbian, Serbo-Croatian), and four world languages (English, French, German and Spanish). We observed the presence of writers and athletes in those languages. None of the selected writers had an article in every language/edition. For instance our writers are almost non-existent in Spanish and French editions, and the situation is only slightly better when it comes to German. The representation of athletes was much better, almost all of them occurring in the selected world languages.

At the end we come to a few questions, such as whether it is more important to write "for yourself", i.e. to work on creating the best possible encyclopaedia intended for audiences in Croatia, and all those who use the Croatian language, or should we systematically work on the presentation of "our" issues in the publications in other languages; whether it be persons or texts from the sphere of politics, history, or (most widely understood) culture and art. If we agree that we will get relatively successful answers to questions starting with who, what, where, when, and much less successful to those beginning with how or why, then this answers the question on the role of the online encyclopaedia in *the transfer of knowledge. The value of the knowledge contained in the answer to the question posed with how or why is "value added" to the encyclopaedia; it is what distinguishes it from a dictionary or a search engine.*

Key words: online, encyclopedia, Wikipedia, web, search engine

Introduction: Information is a search

The digitalization of all aspects of life and work has had a huge impact in the media, because it changed the former (more or less one-way) model of production and transfer of information from the source (author, publisher) to the user (audience), at the same time overcoming the increasing number of (emerging) information, by using new methods and techniques of search and retrieval, where often the simplest act of connecting concepts (sending, linking) is becoming more important than the information content itself. [1]

This is particularly evident in the organization and the presentation of knowledge. Because of the constant changes in its structure, the relationships between facts, definitions, understanding, context, the history and future of an event or phenomenon, of a person and its environment, etc. are changing. At the same time the importance of learning the facts is declining, and increasingly is being replaced with the ability to search and find, assess and evaluate, organize and creatively use relevant information. Since the form of learning is changing – it is less about learning the facts and more and more about learning the ways to use them - organized facts, available to everyone are essential for the functioning of the educational system.

Likewise we need to differentiate the public knowledge accessible to all and free of charge from the so-called private knowledge, knowledge that is someone's property, knowledge that is somehow protected, by copyright or other rights (e.g. commercial law). Nowadays, the content itself is often not protected, but the way in which the content is organized is protected [2] (e.g. data from the phone book are not protected, but the directory as a database is protected) or the reproduction of a public content is protected etc.

In a society that is increasingly based on the use of universally available, mostly free sources of information on the Internet, it becomes necessary to fundamentally redefine the role of those who "produce knowledge". Today the emphasis is on searching and finding information, the emphasis is not only on the content but even more on links, or references between the content, expanding the origi-

[1] Prelog, N: *Elektroničko (interaktivno) nakladništvo danas – Kako je mreža poosobila masovne medije (II),* Medijska istraživanja, no. 1, 1999. http://www.mediaresearch.cro.net/clanak.aspx?l=hr&id=126

[2] Smedinghoff, T: *Legal Guide to Multimedia*. Adison-Wesley Publishing, Reading, 1994.

nal context in which the requested concept appears. Finally, what is asked for is a dynamic rather than a static approach, constant updating and alignment with the actual situation. If the documents (texts or pictures) are not updated for a longer time the frequency of their use reduces, and the number of users and the influence of the site in the process of learning and exchanging knowledge rapidly decrease.

Electronic editions of the most famous encyclopaedias appeared at the end of the twentieth century. Among the first was the Academic American Encyclopedia, as early as 1985 and in the early nineties Compton's Encyclopedia - in 1992, first on CD, and later on DVD. Online editions followed a few years later, leading with Britannica Online in 1994. In the beginning the role of the Internet came down to only the possibility to update content/releases on a permanent, yet computer-readable medium. It was only later followed by "real" online encyclopaedias, first as copies of electronic editions with the addition of interactivity. In the early 21st century complex knowledge databases begin to dominate, which - building on the multimedia capabilities and the comprehensiveness of the web - provide more information and a large number of services to its users in this respect surpassing in many ways the conventional editions, both on paper as well as those on DVD.

Today, most online editions of encyclopaedias include a number of additional content and possibilities of getting the requested information, including dictionaries, thesauruses, atlases, lists of important quotations, different Happened on this day categories, historical overviews of events or the developments of some ideas, lists of acronyms, abbreviations and idioms, current news, statistics and a range of other content such as important historical documents, famous speeches, excerpts from literary works, automatic translators, quizzes, blogs and so on and so forth.

With the increased importance of new information technologies and the use of new media to store and transfer information, the structure of information products (this includes all artistic, scientific and other works), and the way of presentation of such content changed. It's not just about the conversion of data from one medium to another; an electronic encyclopaedia assumes a different organization of information and interactivity in work (the set of information found is adjusted to the current user request). Encyclopaedias online strive to the total access to knowledge, allowing the exchange of ideas, they also protect cultural heritage and particularities, and constantly explore new possibilities for connecting (both existing and new) content.

The limitations of search engines

With the appearance and the growing importance of information search engines (primarily Google) the mode of access to information contained on the Web quickly began to change. While a few years ago it was common when reading online editions of newspapers or magazines to start from the first, front page, and then - though rarely - to advance page by page, imitating the conventional edition or - much more frequently – to look up in the index, or the review of sections those titles which we are interested in and then select the article for reading or viewing or listening, it is no longer the dominant approach. Part of the users will reach the information they need (text, image or video) with the help of a mediator, a portal devoted to news services, or dissemination of information to defined areas of interest (RSS). However, today the vast majority of users come to that which they are interested in with the help of search engines, not taking into account the context in which this information is published, often not even noticing the title of the newspaper or the magazine that brings them. Thus it is clear that encyclopaedias on the Web cannot behave as if they are self-sufficient, assuming that the future users will start searching from the home page of the publication or the publisher. The tools offered to users in order to facilitate finding the required information are necessary, but their importance comes into play only when the user begins his journey to knowledge from the home page - which happens less often - or when he is already on the pages of the encyclopaedia/lexicon/vocabulary, etc., regardless of the way he got there.

On the Web, almost everything is possible to find but it is completely different if you have to know where and how to look, while today we are accustomed to simply type a few words into some of the browsers. This naturally results in certain mental laziness, but few people are concerned about this because the goal, in this case the answer to these questions, often justifies the means. It has been shown many times, that even Google cannot answer all his questions. If we take as a criterion of quality those 5 eternal journalistic questions (who, what, where, when, why) we easily find that the answers to the questions asked with the first four interrogative pronouns are always relatively accessible, even adequate, accurate and sufficient, while things mostly get stuck when you ask a question with **why**.

Demands set on the encyclopaedia may have more ambiguity, unexpectedness and lack of adaptation to the system of finding information, than its developers would expect. The encyclopaedia is not a directory in which the amount of questionability is predictable, where the name, address, the name of company/organization and/or branch of industry unambiguously determines the required number. The amount of ambiguity in any remotely ambitious collection of knowledge is far greater and therefore the description of individual units (tag), as well as the linguistic accuracy of the names of concepts and explanations must be subordinated to it, but it also requires maximum flexibility regarding the ways in which a concept and all the connections surrounding the term can be accessed. Encyclopaedias that evolved from conventional, paper editions to computer-readable i.e. online editions often have within them the immanent structure of the "old" information organization. So no matter how many different new ways to search / find are offered to readers / users, they feel that the links are not natural, that all the possibilities of interactivity and the logic of the functioning of the new media have not been completely used.

On the contrary, those lexicographic and encyclopaedia publications that started on the web are not "suffering" from distinct solutions not appropriate for the new medium. So as the greatest, inimitable success in bookselling was accomplished by a virtual bookstore that never operated in the real world (Amazon), and in sales through classified ads Craiglist which also did not have a paper start, among encyclopaedias by far the greatest importance has had the Wikipedia, the encyclopaedia that in only 10 years (appeared in 2001) grew to over 12 million articles in 279 languages [3]. No need to remind that only an online edition exists – if Wikipedia would be published in books the size of 25 cm and with thickness of 5 cm (about 6 MB per volume), the English edition alone would have 750 volumes.

Another question that arises when we analyze the representation of people or concepts is: what are the criteria upon which it is decided whether a term will be included? When after five years the figure of one million articles was exceeded, people began wonder whether each and every possible person, geographic concept, event or idea belongs in the encyclopaedia. If the intention was to cover every aspect of human knowledge then we have to raise the question of triviality of some concepts. If more stringent editorial policies are introduced to ensure the reputation of a source that has credibility, there remains still the question of inclusion, what has been left out and why? [4]

There is a lot of discussion about this problem both among the authors and editors of Wikipedia, as well as in the academic and professional community. The question that hovers in the air can be reduced to this: Could a possible criterion for the "publication" or "remaining" of a concept or term in the Wikipedia be his popularity, i.e. the number of people who using a search engine look for exactly that term? If this were the sole or even only the most important criterion, then Wikipedia would soon turn - at least by content - into a "yellow" publication, with all the negative connotations of that colour. It is clear that the inclusion of a concept that will never be sought burdens the encyclopaedia. Besides in every country (or in every language) there must be thousands of those who would like to prepare a text about themselves or a favourite character, theme, or the like, not thinking at all whether anyone will ever (other than themselves) ask the question to which such a text could serve as an answer.

[3] http://www.wikipedia.org/

[4] Hellweg, E: *The Wikipedia War; A recent high-profile dispute over the user-written encyclopedia's veracity has the site rethinking some of its rules*. MIT Technology review, Friday, December 16, 2005

3 Wiki – how much do we know each other

As we speak today about the "googlization" of information (what is not on Google may as well not have happened), the "powerpointization" of the presentation of facts or opinions (because the structure of the presentation adapts to the logic of concise entries), or the "facebookization" of personality (if you do not have a profile on that social network it is as if you do not even exist), thus we can safely say that today we encounter the "wikization" of education. There is almost no student who - to prepare a seminar, an essay or just homework - will not at least in part use the definitions, explanations and/or references listed in this online encyclopaedia. The cause is - apart from an already formed habit, not to say addiction - also the fact that the text from Wikipedia, in the vast majority of cases will be found on the first page of the search results display, regardless of the browser used. If the knowledge mentioned there is used only as a starting point for further research, this in itself not a bad thing, but mostly it is simply taking over finished units, including illustrations and other supplemental material.

In this study, we centered around the quality of "coverage" (taking into account only the formal elements, i.e. the number of lines) of some close to us and we might say self-explanatory lexical units in different Wikipedias. We conducted an analysis of the representation of prominent individuals in the above-mentioned four languages or editions (whatever you want to call it), and four world languages (English/EN, French/FR, German (NJ) and Spanish (SP)). We observed the presence of writers and athletes in those languages. At the very beginning we chose the subjects: seven prominent members of the Department of Literature of HAZU (Croatian Academy of Arts and Sciences), and 5 well-known writers who are not members of the Academy. We took into account only living persons, active (or at least recently active) individuals. Previous testing showed that these are the most often represented people in the Wikipedia (e.g., scientists or artists are far less common). What was the case?

Table 1

Members of HAZU – Department of Literature	**(HR)**	**(BS)**	**(SR)**	**(SH)**	**(EN)**	**(FR)**	**(NJ)**	**(SP)**
Ivan Aralica	114	35			55			
Nedjeljko Fabrio	40							
Ivan Kušan	40	15			15			
Slobodan Novak	55							
Luko Paljetak	80							
Pavao Pavličić	75	80	60		5	6	115	
Goran Tribuson	70			35	140			

Table 2

Writers who are not members of HAZU	**(HR)**	**(BS)**	**(SR)**	**(SH)**	**(EN)**	**(FR)**	**(NJ)**	**(SP)**
Slavenka Drakulić		50		42	55		30	
Miro Gavran	50			50		5	70	
Miljenko Jergović	31	55			7	16	47	
Predrag Matvejević	80	55			45	75		
Dubravka Ugrešić	80	75	75	75	55		70	

First, we analyzed members of the HAZU: Ivan Aralica, and Ivan Kušan are represented in the Croatian, Bosnian and English edition; Nedjeljko Fabrio, Luko Paljetak and Slobodan Novak only in the Croatian; Pavao Pavličić is represented with articles in the Croatian, Bosnian, Serbian, English, French and German edition, and Goran Tribuson in the Croatian, Serbo-Croatian and English. For authors who are not members of the HAZU the situation is somewhat different, as we can see from the table.

The results are just calling for a comment! Not wanting to engage in comparisons or evaluations how legitimate one's representation or under-representation is, we can not avoid a few important conclusions. In the first place it is striking that none of the authors has an article in all the languages. Pavao Pavličić and Dubravka Ugrešić have articles in the most editions and they have the most lines: Ugrešić 430, Pavličić 330. At the same time it should be noted that Pavličić is the only one represented in all the world languages we took into account other than Spanish. On the other hand, only Slavenka Drakulić has no article in the Croatian edition, and two authors, Slobodan Novak and Luko Paljetak have an article only in the Croatian edition. Looking at the languages particularly, it is evident that our authors virtually do not exist in the French edition, and the situation is only slightly better with Germany. At the same time one fact should be taken into consideration; the German edition covers parts of Switzerland and Austria, where - or it is at least presumed so - the cultural connections, not only because of a shared history, are far greater than in the case of other countries and languages. On the other hand, the English edition is by far the largest, and was first launched, so in part this may account for the fact that it covers by far the greatest number of authors - as many as eight. Finally one should also say that in Spain Croatian writers do not exist - if we take as a criterion their representation in the Spanish edition of Wikipedia.

Concerning how familiar we are with each other in the territory of former Yugoslavia, seen in relation to the languages we have no problem communicating with, and not even mentioning the common history, it is absolutely startling that out of the 12 monitored authors only 2 are represented in the Serbian edition, 4 in Serbo-Croatian (?), and 7 in the Bosnian edition. The evaluation of who is relevant for the culture and education in a particular environment in the case of Wikipedia, of course, is not made by a ministry, commission or any administrative authority, but by an assumed interest of those who write and edit the articles, as well as those who appear as users of these publications.

Of course, the analysis of individual qualifications that are given to different authors in different editions would take us far, so for example, Dubravka Ugresic is a Croatian (HR), (EN), (NJ) or Yugoslav, Croatian, Dutch and international (BO), (SR) (SH) writer. Predrag Matvejevic is a Bosnian and Croatian (HR) and Yugoslavian (FR) author, while the other editions (wisely?) do not take sides in this respect.

There is most information on Miljenko Jergović. It is the result of the popularity of his works which were translated in a number of foreign languages. He is also the only writer we included in both countries; considering his long-term work in Bosnia and Herzegovina, and the last 15 years of living in Croatia. Aleksandar Hemon is another writer for whom there is also plenty of information. He is not only a writer but also works in the film industry, he lives and works in Chicago, so this is another reason for a relatively large number of lines. Abdulah Sidran is mentioned in four Wikipedias (HR, BiH, EN and FR), which is probably the result of his numerous awards for screenplays and published works in many countries. In the Serbian Wikipedia there is only information on Hemon, and the German Wikipedia has only information on Miljenko Jergović whose many works were translated to that language.

The writer Dobrica Ćosić has the largest number of lines, 421 respectively, which is probably the result of his great political involvement. From the 4 world editions of Wikipedia, the one in English has the largest number of Serbian writers represented - 5 out of 7, while the Spanish edition has the least number. The Bosnian edition offered a surprisingly small number of results, from the 6 authors offered there are articles only on two writers. One of them is Milorad Pavić, who has texts in most of the Wikipedias. But although all four editions of the world Wikipedias write about him, the Serbian, Serbo-Croatian, and even the Bosnian edition, the Croatian edition of Wikipedia does not have a single line on him. It should be noted that Milorad Pavić was still alive at the time of this research (he died on November 30th 2009 and the survey was conducted during the summer of 2009). Dragan Velikić is mentioned in the Serbian Wikipedia, then in the Croatian, because he was born in Pula, as well as in the German, because he was the ambassador of Serbia in Vienna.

Besides writers some athletes have been selected, all very well known, starting from the fact that the boundaries of individual cultures or languages are much easier crossed in a world where sports and what we commonly call "show business" have long since become globalized. This proved to be correct.

Representation of athletes was as expected far better, almost all of them appearing in the selected world languages, with an absolute record in the representation of Goran Ivanišević, with ~ 1450 rows, Janica Kostelić with ~ 1030, Blanka Vlašić with ~ 740, Toni Kukoč ~ 700, and Davor Šuker with ~ 560 lines. Janica Kostelić and Blanka Vlašić are the only ones represented in all the languages/editions, and Goran Ivanišević still has far greater significance to other (international) languages, than in his own region. It is simply unbelievable that there is not even a word on Ivanisevic in two editions (Serbo-Croatian and Bosnian), but the number of lines dedicated to him in the English edition (by far the biggest article!) only confirms that the glory of winning Wimbledon is slow to fade. Toni Kukoč and Ivano Balić are in every edition except Serbo-Croatian while Davor Šuker does not exist in the Serbian and Serbo-Croatian edition.

The Bosnian websites write relatively little about the athletes of their own country. Regarding the "non-football" athletes there are only one karate athlete, and a chess player who have accomplished internationally relevant sports results. Bosnian football players are playing for German and Italian clubs, so they are mentioned in almost all the Wikipedias. Exceptions are the "Serbo-Croatian", which did not offer in its results any of the Bosnian athletes we have included, and the Serbian, which has a few words only on Džeko and Salihamidzic. The Spanish Wikipedia has completely forgotten about Edin Džeko, but the English and German Wikipedia have offered the most extensive information on Bosnian athletes.

In all of the Wikipedias there is the most information on Novak Đoković. It is interesting that some of the foreign Wikipedias contain more information than the Serbian one. The Bosnian Wikipedia has information only on one famous athlete from their neighboring country. The Spanish Wikipedia dedicated the most lines to Ana Ivanović, as a result of her former relationship with a famous person from that country. Jasna Šekarić was declared the world's best shooter of the 20th century by the World Shooting Federation, and she is mentioned in only four of the Wikipedias. The English Wikipedia has the largest number of lines on Nemanja Vidić because he plays for Manchester United. Milorad Čavić is one of the world's most famous swimmers, but by the number of lines in the Wikipedias this might not be concluded.

The purpose of this analysis naturally was not to lament over how much others know about us because then we would then have to start with ourselves first and analyze how much we know about others, and whether it is enough. Instead several questions should be asked: Is it more important to write "for ourselves", that is, to work on creating the best possible encyclopaedia designed for the audience (mostly students who are the most frequent users, but also others) in Croatia and for all who use the Croatian language, or should we systematically work on the presentation of "our" issues in the editions in other languages; whether it be persons or texts from the sphere of politics, history, or (most widely understood) culture and art. In the globalized world of today, is it more important to have all the answers "at hand" and who or what will be written about us left to the good will of the authors in the world (and perhaps some of the more ambitious ones at home)? Should someone be taking care about this?

In the endless debate about the quality and accuracy of what is published in the Wikipedia (the bibliography of articles on this topic has already reached a few hundred units)[5] a new factor recently appeared: the more relevant the concept is i.e. the greater the interest for it is the greater the chance is that it will be corrected i.e. that a large number of readers/reviewers/authors will not allow inaccurate or biased information. Accordingly, if there are mistakes – they are in those places that are not much visited. The quality of the texts is thus proportional to the interest, and the number of reviewers (reader response) is what ensures accuracy.

Finally we return to the issues raised in the debate on the possibilities of finding relevant information by search engines. If we agree that we will get relatively successful answers to questions starting with who, what, where, when, and much less successful to those beginning with how or why, then this answers the question on the role of the online encyclopaedia in the transfer of knowledge and can even suggest an answer to the ever present dilemma: what to give for free, and what to charge (either by a subscription system, or as a single answer)? The value of the knowledge contained in the answer to the question posed with how or why is "value added" to the encyclopaedia; it is what distinguishes it from a dictionary or a search engine.

All these questions are definitely matter for further research; therefore this paper can be seen only as a foundation review of the topic intended to serve as a basis for research.

References

Berners-Lee, Tim: Weaving The Web: The Past, Present And Future Of The World Wide Web. Orion, London, 1999.

Britannica reaches out to the web. http://news.bbc.co.uk/1/hi/technology/7846986.stm

Cohen, Noah: Wikipedia Tries Approval System to Reduce Vandalism on Pages. http://bits.blogs.nytimes.com/2008/07/17/wikipedia-tries-approval-system-to-reduce-vandalism-on-pages/

Giles, Jim: Jimmy Wales' Wikipedia comes close to Britannica in terms of the accuracy of its science entries, a Nature investigation finds. Nature 438, 900-901 (15 December 2005).

Hefferman, V: Lexicographical Longing. May 11, 2008, http://www.nytimes.com/2008/05/11/magazine/11wwln-medium-t.html?_r=1

Kolbitsch, J. Et. Al: Dynamic Adaptation of Content and Structure in Electronic Encyclopaedias Texas Digital Library, Vol 8, No 3 (2007), http://journals.tdl.org/jodi/issue/view/35

Kung, Lucy; Picard, Robert; Towse, Ruth: The Internet and the Mass Media. Sage, London, 2008.

Lee Rainie; Tancer, Bill: 36% of online American adults consult Wikipedia. Pew Internet & American Life Project, April 2007

Manovich. Lev: The Language of New Media. The MIT Press, Cambridge, 2001.

Tapscott, Don: Growing Up Digital – How the Net Generation is Changing Your World. McGraw-Hill, New York, 2008.

The battle for Wikipedia's soul. The Economist, Mar 6th 2008. http://www.economist.com/printedition/displaystory.cfm?STORY_ID=10789354

Viégas, F. B. et. al: The Hidden Order of Wikipedia. http://www.research.ibm.com/visual/papers/hidden_order_wikipedia.pdf

[5] Giles, Jim: *Jimmy Wales' Wikipedia comes close to Britannica in terms of the accuracy of its science entries, a Nature investigation finds*. Nature 438, 900-901 (15 December 2005).