Analysis of References across Wikipedia Languages *

Włodzimierz Lewoniewski, Krzysztof Węcel, Witold Abramowicz

Poznań University of Economics and Business, Poland
{wlodzimierz.lewoniewski, krzysztof.wecel, witold.abramowicz}@ue.poznan.pl

Abstract. Reliable information sources are important to assess content quality in Wikipedia. Using references readers can verify facts or find more details about described topic. Each Wikipedia article can have over 290 language versions. As articles can be edited independently in any language, even by anonymous users, the information about the same topic may be inconsistent. This also applies to sources that can be found in various language versions of particular article, so the same statement can have different sources. In some cases, Wikipedia users, which speak two or more languages, can transfer information with references between language versions. This paper presents an analysis of using common references in over 10 million articles in several Wikipedia language editions: English, German, French, Russian, Polish, Ukrainian, Belarussian. Also, the study shows the use of similar sources and their number in language sensitive topics.

Keywords: Wikipedia, reference, source, citation.

1 Introduction

Wikipedia is a popular large collection of human knowledge. In April, 2017 this free online encyclopedia was the fifth most visited website in the world.[1] Nowadays there are over 44 million articles in almost 300 language versions of Wikipedia. The biggest language version is English, which has more than 5 million articles.

Wikipedia offers an innovative way to read and edit the information online for people around the world. Even anonymous users without confirming their skills and experience can collaborate in articles creation in this community knowledge base.

Despite the fact that Wikipedia is often criticized for poor quality of information, for the last 10 years its articles have been cited in over 80 thousands scientific publications.[2] This is almost 10 times more than number articles citing Encyclopaedia Britannica in scientific publications in the same period.

One of the most important quality measures for Wikipedia is verifiability. Different language versions of the same topic in Wikipedia can be created and edited independently. Therefore, there are often differences in quality between various language version of the same article. Wikipedia users who speak several languages, try

[1] http://www.alexa.com/siteinfo/wikipedia.org

[2] Information about the number of scientific publications is taken from https://www.scopus.com where search query was *REF(wikipedia.org/wiki)* in works published in 2008-2017

* This is a preprint version. The final publication is available at Springer via https://doi.org/10.1007/978-3-319-67642-5_47

to translate some content between more and less developed language versions. Often along with the content, users also transfer information about references. Referencing verifiable resources enhances the quality of Wikipedia articles [10].

In this paper we analyze number of references included in Wikipedia articles in various languages, the most popular information sources, number of common references in different pairs of Wikipedia language editions. In order to compare the same references with different description we used the unification method based on special identifiers. In this study we analyze all articles with references from some of the most the developed Wikipedia editions and some less developed ones: English (EN), German (DE), French (FR), Russian (RU), Polish (PL), Ukrainian (UK), and Belarussian (BE).

2 Sources in Wikipedia

Wikipedia articles with high quality must be well-researched and have representative survey of the relevant literature.[3] When adding or editing article content, authors must also add reliable and published sources. As a result, people using the encyclopedia can check where the information comes from and verify the facts described in it.

A large number of Wikipedia articles are unassessed or have low quality grade [1]. Differences between language versions about same topic cause an additional difficulty in assessing the quality of articles.

There is a series of studies that use references for assessing quality of Wikipedia articles. One group of scientific works examined how references affected the articles quality. Experiments showed that number of references and derivatives (e.g. references and articles length ratio) were one of the most important predictors in article quality models [2,3]. Online service WikiRank[4] together with other features uses the number of references to assess and compare the quality of Wikipedia articles in different languages.

Second group of studies focused on quality of references in Wikipedia. One of the first studies in this direction suggested that Wikipedia articles tend to cite articles in high impact journals such as New England Journal of Medicine, Nature, Science [8]. At the same time number of peer reviewed academic papers in the health sciences which are citing Wikipedia is increasing [4]. References can cover a wide range of subjects, but particularly focused on articles from ecology, evolution and other topics that can enrich the encyclopedia with scholarly sources [6]. More than half of the references used in the history articles of the encyclopedia are internet sources, such as news, media, government websites [7]. If users add references connected with academic publications, then they prefer to use book as a source rather than articles [5]. So, Wikipedia is especially valuable due to the potential direct linkages to other primary sources through special identifier such as DOIs or PubMed IDs [9]. Additionally, academic status of work is the most important predictor of its appearance in Wikipedia references [12].

[3] https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

[4] http://wikirank.net

Wikipedia has also developed a set of templates for flagging articles that have not enough references or there are no references at all.[5] That template is the most frequent in English Wikipedia from the over 300 specific quality flaw templates [11]. So, we can conclude that Wikipedia community pays special attention to availability of references in articles.

3 Extraction of References

Using Wikipedia dumps from May, 2017, we have extracted all references from over 10 million articles in 7 language editions (BE, DE, EN, FR, PL, RU, UK).

In wiki-code references are usually placed between special tags <*ref*>...</*ref*>.[6] In general, we can divide this references into two groups: with special template and without it. In the case of references without special template they usually have URL of source and some optional description (e.g. title).

References with special templates can have different data describing the source. Here in separate fields we can add information about author(s), title, URL, format, access date, publisher and others. Additionally, these templates can contain special identifiers such as DOI, JSTOR, PMC, PMID, arXiv, ISBN, ISSN, and OCLC. The set of possible parameters depends on the type of templates, which can describe web source, book, journal, news, conference, act and others. It is important to note that each language version of Wikipedia can use own group of templates with own names and set of parameters that describe information sources.

Table 1. Articles and references count in different language versions of Wikipedia in May 2017.

Lang.	Number of articles	Articles with ref.	Number of references	Unique ref.	Ref. with template	Unique ref. domains
BE	143,023	31,522	111,961	82,295	54,456	22,042
DE	2,057,871	874,370	3,777,825	2,988,443	1,275,773	500,560
EN	5,396,615	3,540,201	25,534,467	18,470,122	19,942,239	1,588,692
FR	1,866,412	818,909	4,510,703	3,364,408	2,789,431	389,588
PL	1,219,709	611,247	2,468,167	1,548,696	2,045,508	184,909
RU	1,391,120	714,599	3,852,470	2,873,069	2,184,470	356,896
UK	693,969	260,913	1,010,965	635,149	567,615	114,109
Total	**12,768,719**	**6,851,761**	**41,266,558**	**29,962,182**	**28,859,492**	**3,156,796**

Source: own calculation based on Wikipedia dumps.

In order to extract information about sources we created own parser, which takes into the account different names of references templates and parameters in each

[5] https://en.wikipedia.org/wiki/Template:Unreferenced

[6] Also can be <*ref name="..."*>...</*ref*> or <*ref name="..." /*>

Wikipedia language edition. We investigated about 12,7 million articles (which are not redirects to other articles) and found over 42 million references from over 3 million website domains in 7 language versions. More detailed statistics are placed in Table 1.

Zipfian distribution of domains frequency of sources in each language is shown in Figure 1.

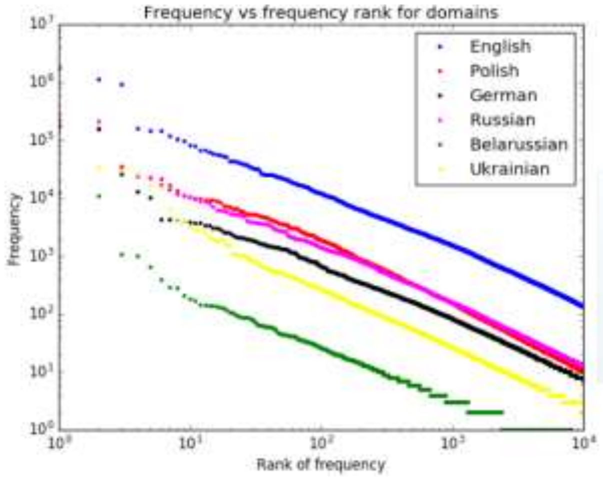

Figure 1. Zipflaw frequency vs. frequency rank for domains in each language version of Wikipedia

It is important to note that for references with the same special identifiers we can determine equivalency even though they have different parameters in description (e.g. titles in another languages). We can also unify their URL. For example if reference have ISBN number “978-3-319-46254-7”, we give it URL “books.google.com/books?vid=ISBN9783319462547”. More detailed information about identifiers which we used to unifying the references is shown in Table 2.

Table 2. Identifiers that used for URL unification of references.

Identifier	Description	New URL
arXiv	arXiv repository identifier	*http://arxiv.org/abs/...*
DOI	Digital object identifier	*http://doi.org/...*
ISBN	International Standard Book Number	*http://books.google.com/books?vid=ISBN...*
ISSN	International Standard Serial Number	*https://worldcat.org/ISSN/...*
JSTOR	Journal Storage number	*https://jstor.org/stable/...*
PMC	PubMed Central	*https://ncbi.nlm.nih.gov/pmc/articles/PMC...*
PMID	PubMed	*https://ncbi.nlm.nih.gov/pubmed/...*
OCLC	WorldCat's Online Computer Library Center	*https://worldcat.org/oclc/...*

Table 3 present number of unique references with particular identifier in each language version of Wikipedia.

Table 3. Number of references with particular identifier in Wikipedia articles

lang.	arXiv	DOI	ISBN	ISSN	JSTOR	PMC	PMID	OCLC
BE	90	1,185	13,656	78	28	53	198	19
DE	2,416	31,014	171,073	12,696	1,591	1,022	3,481	2,671
EN	4,226	1,014,602	1,670,495	79,442	35,709	16,384	52,387	54,995
FR	842	50,381	332,593	25,297	2,045	782	7,406	7,598
PL	577	41,796	245,833	23,319	781	338	11,157	1,131
RU	1,577	33,956	232,427	3,045	785	1,236	5,164	977
UK	301	2,562	37,628	618	96	160	313	400
Total	**10,029**	**1,175,496**	**27,03,705**	**144,495**	**41,035**	**19,975**	**80,106**	**67,791**

Source: own calculations.

Unification of URLs based on identifiers was used for counting the number of unique references and will be used for comparison of similarity of references in different language versions of Wikipedia articles.

4 Similarity of Sources

In order to examine similarity of sources across different Wikipedia language versions we create three datasets with articles covering different topics (Wiki, Wiki7, Wiki5) and three datasets with language sensitive topics (LST). All data extracted from Wikipedia dumps from May 2017.

4.1. Wiki

We first chose all the articles (about 6.9 million) with references in 7 considered languages. After extraction we had almost 30 million references. Table 4 presents results in a number of common sources on each language intersection.

Table 4. Number of common references used in Wikipedia language versions in Wiki dataset.

lang.	BE	DE	EN	FR	PL	RU	UK
BE	**82,295**	3,522	19,116	6,127	5,043	47,931	13,100
DE	-	**2,988,443**	345,202	81,572	41,558	69,634	21,097
EN	-	-	**18,470,130**	584,037	244,120	635,546	160,408
FR	-	-	-	**3,364,409**	61,104	118,700	32,470
PL	-	-	-	-	**1,548,696**	71,221	26,022
RU	-	-	-	-	-	**2,873,070**	185,473
UK	-	-	-	-	-	-	**635,149**

Source: own calculations.

The largest number of references in the English Wikipedia can be explained by the largest number of articles in it. In the next datasets we take equal number of articles in each language. We show unique references overlaps between selected language versions in Figure 1.

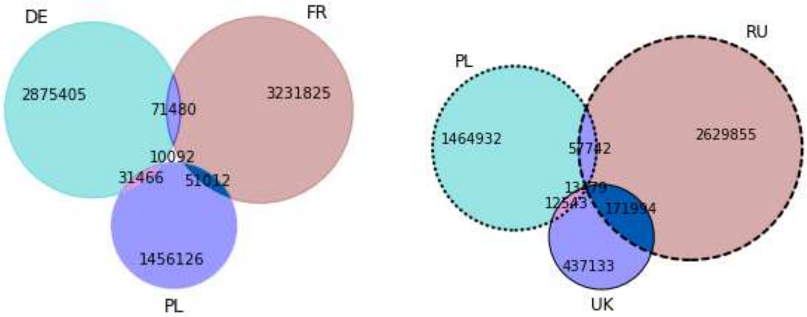

Figure 2. Unique references overlap between selected language version of Wikipedia. Source: own calculations.

It is noticeable that there are more common sources among Slavic language versions (PL, RU, UK).

Table 5. Top 10 most popular reference domains in various Wikipedia language versions in Wiki dataset[7]

BE	DE	EN	FR
books.google.com	books.google.com	books.google.com	books.google.com
pravo.by	books.google.de	doi.org	doi.org
football.by	spiegel.de	nytimes.com	books.google.fr
doi.org	doi.org	news.bbc.co.uk	worldcat.org
cuetracker.net	welt.de	bbc.co.uk	lemonde.fr
naviny.org	zeit.de	theguardian.com	legifrance.gouv.fr
by.tribuna.com	faz.net	worldcat.org	lefigaro.fr
worldsnooker.com	worldcat.org	news.google.com	insee.fr
web.archive.org	youtube.com	youtube.com	gallica.bnf.fr
gks.ru	sueddeutsche.de	census.gov	interieur.gouv.fr

PL	RU	UK
books.google.com	books.google.com	insee.fr
web.archive.org	doi.org	books.google.com
doi.org	insee.fr	kia.hu
sports-reference.com	billboard.com	w1.c1.rada.gov.ua
archive.is	textual.ru	demo.istat.it

[7] Top 100 popular references domains with the number of references in each language version of Wikipedia can be found on page: http://en.lewoniewski.info/2017/top-100-domains-in-wikipedia-references/

worldcat.org stat.gov.pl discogs.com allmusic.com getamap.ordnancesurvey.co.uk	int.soccerway.com lenta.ru web.archive.org youtube.com kommersant.ru	nsi.bg cvk.gov.ua pravda.com.ua youtube.com web.archive.org

Source: own calculations.

Table 6. Number of common references' domains used in Wikipedia language versions in Wiki dataset.

lang.	BE	DE	EN	FR	PL	RU	UK
BE	**22,042**	10,563	15,393	10,475	9,783	19,030	12,485
DE	-	**500,560**	219,536	104,212	62,595	90,361	41,407
EN	-	-	**1,588,692**	201,601	101,495	183,234	69,437
FR	-	-	-	**389,588**	56,693	86,071	39,426
PL	-	-	-	-	**184,909**	60,130	32,382
RU	-	-	-	-	-	**356,896**	73,254
UK	-	-	-	-	-	-	**114,109**

Source: own calculations.

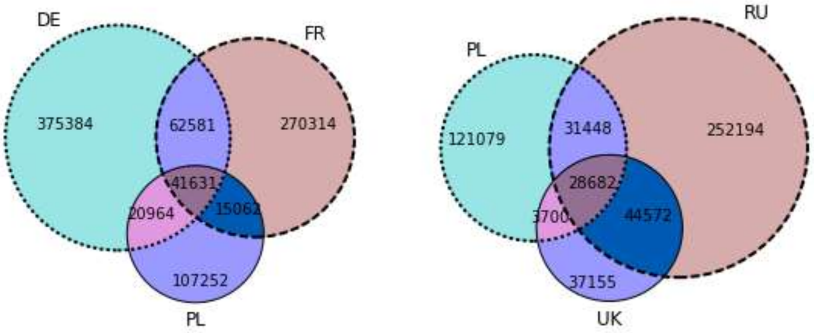

Figure 3. References' domains overlap between selected language version of Wikipedia. Source: own calculations.

Comparing figure 2 and 3, we can find that references domains are more international - there are relatively more common across language versions of Wikipedia.

4.2. Wiki5

In this dataset there are 273,878 articles, that are written in five language versions: DE, EN, FR, PL, RU. Number of articles and extracted references are shown in Table 7.

Table 7. Articles and references count in different language versions of Wikipedia in Wiki5 dataset.

Lang.	Number of articles	Articles with ref.	Number of references	Ref. with template	Unique ref.	Unique ref. domains
DE	273,878	149,664	917,936	326,514	792,077	155,869
EN	273,878	205,503	3,897,533	3,232,357	3,261,656	383,766
FR	273,878	147,655	1,276,342	821,887	1,056,169	148,614
PL	273,878	129,118	745,196	615,556	561,213	83,519
RU	273,878	154,936	1,154,815	712,284	963,545	151,549
Total	**1,369,390**	**786,876**	**7,991,822**	**5,708,598**	**6,634,660**	**923,317**

Source: own calculations.

Table 8. Number of common references used in Wikipedia language versions in Wiki5 dataset.

lang.	DE	EN	FR	PL	RU
DE	**792,077**	90,863	26,797	19,345	31,043
EN	-	**3,261,658**	170,200	104,595	236,229
FR	-	-	**1,056,170**	29,015	49,156
PL	-	-	-	**561,213**	36,239
RU	-	-	-	-	**963,546**

Source: own calculations.

4.3. Wiki7

In this dataset there are 46,957 articles, that are written in all seven analyzed languages: BE, DE, EN, FR, PL, RU, UK. Number of articles and extracted references are shown on table 8.

Table 9. Articles and references count in different language versions of Wikipedia in Wiki7 dataset.

Lang.	Number of articles	Articles with ref.	Number of references	Ref. with template	Unique ref.	Unique ref. domains
BE	46,957	10,538	51,387	28,016	43,778	13,497
DE	46,957	27,278	239,520	86,902	217,236	54,640
EN	46,957	37,884	1,089,035	918,726	955,305	152,324
FR	46,957	33,589	415,599	272,618	354,607	61,427
PL	46,957	24,493	203,567	169,139	159,002	31,853
RU	46,957	27,959	353,592	202,034	308,499	65,567
UK	46,957	20,431	111,213	60,023	91,191	26,268
Total	**328,699**	**182,172**	**2,463,913**	**1,737,458**	**2,129,618**	**405,576**

Source: own calculations.

Table 10. Number of common references used in Wikipedia language versions in Wiki7 dataset.

lang.	BE	DE	EN	FR	PL	RU	UK
BE	**43,778**	1,378	9,733	2,757	2,637	27,378	6,794
DE	-	**217,236**	17,768	5,467	3,572	5,377	2,585
EN	-	-	**955,305**	44,528	26,139	47,782	21,066
FR	-	-	-	**354,607**	7,262	11,134	4,532
PL	-	-	-	-	**159,002**	8,320	3,711
RU	-	-	-	-	-	**308,500**	28,619
UK	-	-	-	-	-	-	**91,191**

Source: own calculations.

4.4. LST

Additionally to the above analyses, we decided to carry out additional analysis concerning "nationality" of sources. We chose three sub datasets, which described cities in particular country: Poland, Germany, and France. So, these datasets are Language Sensitive. We further chose cities, which were described at least in five languages: DE, EN, FR, PL, RU. As a result we obtained a dataset with articles about 10516 German cities, 10092 French cities, and 904 Polish cities.

German cities (LST DE)

Similarly to the previous datasets, Table 11 presents number of articles with references and number of references in each language. It is noticeable that German Wikipedia have the highest number of articles with references and the highest total number of references. So, information about German cities is the most verifiable in German Wikipedia.

Table 11. Articles and references count in different language versions of Wikipedia in LST DE dataset.

Lang.	Number of articles	Articles with ref.	Number of references	Ref. with template	Unique ref.	Unique ref. domains
DE	10,516	9,532	64,305	18,893	49,436	16,541
EN	10,516	2,540	11,744	3,168	7,936	3,359
FR	10,516	1,129	2,752	484	1,719	956
PL	10,516	2,805	5,087	1,204	1,572	1,155
RU	10,516	8,820	9,875	292	961	607
Total	**52,580**	**24,826**	**93,763**	**24,041**	**61,624**	**22,618**

Source: own calculations.

From Table 12 we can argue that more common sources have German end English Wikipedia when describing German cities.

Table 12. Number of common references used in Wikipedia language versions in LST DE dataset.

lang.	DE	EN	FR	PL	RU
DE	**49,436**	1,045	234	80	90
EN	-	**7,936**	77	49	75
FR	-	-	**1,719**	16	24
PL	-	-	-	**1,572**	25
RU	-	-	-	-	**961**

Source: own calculations.

French cities (LST FR)

Based on tables1 13 and 14 we can make a similar conclusion, that French cities have the most verifiable description in French Wikipedia, and more common references have this language version with English Wikipedia.

Table 13. Articles and references count in different language versions of Wikipedia in LST FR dataset.

Lang.	Number of articles	Articles with ref.	Number of references	Ref. with template	Unique ref.	Unique ref. domains
DE	10,092	2,568	8,167	3,460	6,959	1,902
EN	10,092	1,738	11,896	5,830	9,652	3,342
FR	10,092	8,763	101,325	52,003	70,817	15,700
PL	10,092	643	1,144	954	497	179
RU	10,092	8,157	38,007	34,844	21,930	1,103
Total	**50,460**	**21,869**	**160,539**	**97,091**	**109,855**	**22,226**

Source: own calculations.

Table 14. Number of common references used in Wikipedia language versions in LST FR dataset.

lang.	DE	EN	FR	PL	RU
DE	**6,959**	128	368	14	408
EN	-	**9,652**	2,076	10	87
FR	-	-	**70,817**	27	683
PL	-	-	-	**497**	6
RU	-	-	-	-	**21,930**

Source: own calculations.

Polish cities (LST PL)

Finally, in the case of Polish cities, Table 15 demonstrates similar tendency – Polish Wikipedia have the highest number of references, and therefore is the most prominent for this dataset. However, Table 16 shows that pair EN&PL does not have the biggest number of common references (99) – a little more have EN&FR language version (101).

Table 15. Articles and references count in different language versions of Wikipedia in LST PL dataset.

Lang.	Number of articles	Articles with ref.	Number of references	Ref. with template	Unique ref.	Unique ref. domains
DE	904	608	2,439	387	2,116	932
EN	904	476	2,747	1,930	2,382	1,320
FR	904	253	541	179	472	350
PL	904	904	14,804	9,471	11,098	4,451
RU	904	158	394	151	339	235
Total	**4,520**	**2,399**	**20,925**	**12,118**	**16,407**	**7,288**

Source: own calculations.

Table 16. Number of common references used in Wikipedia language versions in LST PL dataset.

lang.	DE	EN	FR	PL	RU
DE	**2,116**	81	13	58	9
EN	-	**2,382**	101	99	53
FR	-	-	**472**	37	10
PL	-	-	-	**11,098**	40
RU	-	-	-	-	**339**

Source: own calculations.

We can see that in each language sensitive datasets the total number of references is always the biggest in own language. If we look to the biggest number of common sources between two languages, always English version is the first. This could mean that most users that translate content from one language to another often choose English version as a source or a destination.

5 Conclusions and Future Work

Wikipedia community puts great emphasis on verifiability of information contained in the articles. Using special identifiers we can unify the same references that are present in various Wikipedia editions.

This study shows that different language versions of Wikipedia use common sources in different manner depends on a topic. The biggest number of common references have English and German versions – 345,202. However, we need to take into account total number of articles in these languages – they are the biggest Wikipedia editions. If we consider only articles that are represented in at least 5 considered languages, than the biggest number of common references have Russian and English Wikipedia editions.

For language sensitive topics we always get the same results – the most verifiable information is available in the respective language. In this case, often this topics have more common references with the biggest language version of Wikipedia – English.

Our future work will be devoted to more in-depth researches about similarity of references. We plan to use some external open citation databases (e.g. WorldCat[8], Google Schoolar[9], Microsoft Academic[10]) to find different data about same sources (URLs, titles, identifiers, etc.). This databases can be also helpful to find information about importance of particular source (e.g. citation index, impact factor). We plan include this analysis to assess the quality of articles and parameters in special templates – infoboxes. This can help to improve the articles quality in less developed language versions of Wikipedia and also enrich other popular open knowledge databases such as DBpedia[11], Wikidata[12], YAGO, Freebase and others.

References

1. Węcel, K., Lewoniewski, W., (2015), *Modelling the Quality of Attributes in Wikipedia Infoboxes.* Business Information Systems Workshops. Volume 228 of Lecture Notes in Business Information Processing. Springer International Publishing, pp. 308–320
2. Warncke-Wang, M., Cosley, D., Riedl, J., (2013), *Tell me more: an actionable quality model for Wikipedia*, Proceedings of the 9th International Symposium on Open Collaboration.
3. Lewoniewski,W., Węcel, K., Abramowicz,W., (2016), *Quality and importance of Wikipedia articles in different languages*, Information and Software Technologies: 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016, Proceedings. Springer International Publishing, Cham, pp. 613–624.
4. Bould, M. D., Hladkowicz, E. S., Pigford, A. A. E., Ufholz, L. A., Postonogova, T., Shin, E., Boet, S. (2014), *References that anyone can edit: review of Wikipedia citations in peer reviewed health science literature*, BMJ, 348.
5. Kousha, K., Thelwall, M., (2017), *Are Wikipedia citations important evidence of the impact of scholarly articles and books?*, Journal of the Association for Information Science and Technology, 68(3), pp. 762-779.
6. Lin, J., Fenner, M. (2014), *An analysis of Wikipedia references across PLOS publications*, 14: Expanding impacts and metrics, An ACM Web Science Conference 2014 Workshop, pp. 23-26.
7. Luyt, B., & Tan, D. (2010). Improving Wikipedia's credibility: References and citations in a sample of history articles. Journal of the American Society for Information Science and Technology, 61(4), 715-722.
8. Nielsen, F. Å., (2007), *Scientific citations in Wikipedia*, First Monday, 12(8)
9. Page, R. D., (2010), *Wikipedia as an encyclopaedia of life*, Organisms Diversity & Evolution, 10(4), pp. 343-349.
10. Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., & Lanamäki, A., (2015), *"The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia*, Journal of the Association for Information Science and Technology, 66(2), pp. 219-245.
11. Anderka, M., (2013), *Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia*, Doctoral dissertation, Bauhaus-Universität Weimar Germany.
12. Teplitskiy, M., Lu, G., Duede, E., (2016), *Amplifying the impact of Open Access: Wikipedia and the diffusion of science*, Journal of the Association for Information Science and Technology.

[8] http://www.worldcat.org
[9] https://scholar.google.com
[10] https://academic.microsoft.com/
[11] http://www.dbpedia.org
[12] https://www.wikidata.org