Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics

Włodzimierz Lewoniewski[1,†,‡] *, Krzysztof Węcel[1,‡] and Witold Abramowicz[1,‡]

[1]   Poznań University of Economics and Business;
     {wlodzimierz.lewoniewski,krzysztof.wecel,witold.abramowicz}@ue.poznan.pl
*   Correspondence: wlodzimierz.lewoniewski@ue.poznan.pl; Tel.: +48 (61) 639-27-93
†   Current address: al. Niepodległości 10, 61-875 Poznań, Poland
‡   These authors contributed equally to this work.

Abstract: In Wikipedia, articles about various topics can be created and edited independently in each language version. Therefore, quality of information about the same topic depends on language. Any interested user can improve an article and that improvement may depend on popularity of the article. The goal of this study is to show what topics are best represented in different language versions of Wikipedia using results of quality assessment for over 39 million articles in 55 languages. In this paper we also analyze how popular are selected topics among readers and authors in various languages. We used two approaches to assign articles to various topics. First, we divided articles into 27 main topics based on information extracted from over 10 million categories in 55 language versions and analyzed about 400 million links from articles to over 10 million categories and over 26 million links between categories. In the second approach we used data from DBpedia and Wikidata. We also showed how the results of the study can be used to build local and global rankings of the Wikipedia content.

Keywords: Wikipedia; Information quality; Popularity; Topics identification; Wikidata; DBpedia; WikiRank

1. Introduction

Nowadays, in order to make the right economic decisions, one needs to analyze and interpret vast amount of information. The quantity and quality of information in a large degree determine the quality of decisions in various branches of the economy. On the one hand, one must take care of access to proper sources of information. On the other hand, the quality of information determined by various characteristics (such as relevance, accuracy) is also important. High-quality information is essential for effective operation and decision-making in the organization [1]. Inaccurate and incomplete information may have a negative impact on the company's competitive edge [2].

The Internet enables cooperation and exchange of information on a global scale. Useful information can be found both in specialized sources as well as in general online resources. Nowadays, everyone can also contribute to the development of common human knowledge on the Internet. One of the best examples of such online repositories is Wikipedia, in which content can be created from the level of a web browser. This online encyclopedia has been available for approximately 20 years as a freely available resource, and anyone willing can co-create content. Wikipedia relatively quickly became an important source of information around the world. It contains over 50 million articles in over 300 different languages [3]. The English language version is the largest and contains over 5.8 million articles. Currently, Wikipedia is placed on the fifth place in the ranking of the most visited websites on the Internet [4], giving way only to Google, YouTube, Facebook and Baidu. Such popularity even led to the fact that Wikipedia can shape the language that scientists use in their works [5].

Despite its popularity, Wikipedia is often criticized for the low quality of content [6]. Articles on a specific topic can be created independently in each language version. Therefore, often the quality of information about the same subject may vary depending on the language [7–10]. It should also be noted that the topic described in one language version can be translated into other languages. However, a relatively small number of users with knowledge of two or more languages take up such an initiative by transferring content between different language versions [11].

In order to discern the quality of content, the Wikipedia community created a grading system for articles. However, each language version can use its own standards and grading scale [12,13]. For example, in English Wikipedia, articles can get one of 7 grades (from highest to lowest): Featured Articles (FA), Good Article (GA), A-class, B-class, C-class, Start, Stub. Russian Wikipedia has also 7 quality grades but with other names and criteria: Izbrannaja Stat'ja (similar to FA), Horoshaja Stat'ja (similar to GA), Dobrotnaja Stat'ja, I, II, III, IV (similar to Stub). German Wikipedia used only two quality grades (Exzellente Artikel and Lesenswerte Artikel) which has similar criteria to FA and GA grades in English version. Polish Wikipedia uses 5 quality grades: Artykuły na Medal (similar to FA), Dobre Artykuł (similar to GA), Czwórka, Start, Załążek (similar to Stub).

Depending on language version, we can have different quantity of the quality grades and rules of assessments of the articles quality. There is also an additional problem – a large number of unassessed articles. For example, German and Polish Wikipedia has less than 1% of articles with quality grades. These reasons not only pose problems for comparing the quality of articles in the same language version but also lead to challenging task to evaluate and compare different language versions of articles on the same topic.

Automatic quality assessment of Wikipedia articles is a known challenge in the scientific community. However, most of the approaches focused on only the biggest edition (English) or a few language versions of Wikipedia. Using machine learning techniques it is possible to solve the problem of quality assessment of Wikipedia articles as a classification task. To build such models, various features can be taken into the account, for example length of the articles, number of the references, number of the images, sections to mention just the simplest ones [14–19]. These approaches are usually limited to the number of the grades in each languages and often the grading schemes are different between languages.

One of the universal approaches for quality assessment of multilingual articles is Objective Revision Evaluation Service (ORES) [20]. This service automates tasks like detection of vandalism and removal of edits made in bad faith [21]. Additionally the service can evaluate articles on a scale between 0 and 1 in some language versions. However, currently, automatic quality assessment of an article by the ORES is only limited to 9 language version of the Wikipedia and it does not include such developed language chapters as German, Spanish, Italian, Polish, Japanese or Chinese.

In these study we used the synthetic measure [12,22] to combine several article features to assess the quality of over 39 million Wikipedia articles in different languages on a scale between 0 to 100. Additional focus of this work was to analyze demand for information about various topics and languages from the point of view of readers, as well as from the creators of Wikipedia content.

Our previous study [23] showed that using different SEO metrics can be also useful to assess the quality of the Wikipedia articles. Such indicators as social signals from Facebook, Twitter, Pinterest, Youtube and others can help to determine also the popularity the content in multilingual encyclopaedia from the external sources. In this paper we decided to use popularity features that are available in Wikipedia database – page views and number of unique authors of the articles.

For the purposes of the study we selected 55 language versions of Wikipedia that in 2018 and 2019 had at least 100 thousand articles and the depth indicator was at least 5. The depth (or editing depth) shows how frequently articles are updated in selected language version of Wikipedia [24]. Table 1 presents basic statistics about 55 language versions of Wikipedia that were considered in the study.

Table 1. 55 language versions of Wikipedia with articles count, views from unique devices and total page views.

No.	Language version	Abbr.	Articles	Authors	Total page views	Unique devices
1	English	en	5 835 946	36 031 942	7 846 676 922	866 456 515
2	Swedish	sv	3 748 546	664 601	102 423 252	12 597 043
3	German	de	2 288 148	3 158 210	975 590 897	114 380 633
4	French	fr	2 094 723	3 405 365	742 709 055	96 553 550
5	Dutch	nl	1 962 531	986 565	155 136 113	23 873 475
6	Russian	ru	1 539 411	2 500 221	896 358 323	96 537 026
7	Italian	it	1 518 702	1 803 513	544 481 445	53 459 817
8	Spanish	es	1 514 431	5 375 409	1 090 438 930	180 071 200
9	Polish	pl	1 329 622	949 766	278 226 329	29 262 659
10	Vietnamese	vi	1 205 176	660 020	68 454 735	16 396 173
11	Japanese	ja	1 145 838	1 462 052	1 043 323 322	98 636 732
12	Chinese	zh	1 051 874	2 709 195	412 676 457	52 328 429
13	Portuguese	pt	1 007 942	2 230 598	352 570 671	69 605 320
14	Ukrainian	uk	896 476	448 345	62 906 361	10 849 975
15	Arabic	ar	715 850	1 643 146	188 230 435	39 994 487
16	Persian	fa	671 576	812 855	142 075 761	21 993 488
17	Serbian	sr	618 230	240 802	27 054 615	4 776 849
18	Catalan	ca	610 217	319 681	21 121 481	3 439 969
19	Norwegian (Bokmål)	no	506 510	457 767	36 974 998	6 017 919
20	Indonesian	id	458 034	1 047 391	146 481 271	33 774 831
21	Finnish	fi	454 859	413 533	65 437 832	7 372 105
22	Korean	ko	450 896	559 608	83 623 819	19 933 158
23	Hungarian	hu	448 744	133 232	54 741 921	8 298 454
24	Serbo-Croatian	sh	447 790	409 910	5 900 087	2 372 396
25	Czech	cs	425 852	448 816	73 574 810	9 338 114
26	Romanian	ro	393 439	470 902	39 466 674	7 711 157
27	Basque	eu	332 997	98 920	9 067 706	446 209
28	Turkish	tr	325 627	233 118	25 389 323	3 076 606
29	Malay	ms	325 592	1 028 128	12 291 727	3 960 414
30	Esperanto	eo	256 487	156 711	1 981 767	263 084
31	Bulgarian	bg	254 272	84 451	27 272 998	4 093 761
32	Danish	da	250 890	249 638	30 667 722	5 190 512
33	Armenian	hy	248 278	349 917	6 013 622	918 474
34	Hebrew	he	240 943	507 618	58 213 949	6 344 428
35	Slovak	sk	229 146	171 238	16 854 614	3 117 661
36	Min Nan	zh-min-nan	228 102	37 919	572 773	84 788
37	Kazakh	kk	223 881	85 934	11 562 925	2 142 268
38	Croatian	hr	204 240	216 016	21 779 929	4 497 371
39	Lithuanian	lt	194 537	131 095	12 276 882	1 984 922
40	Estonian	et	189 742	125 754	11 502 319	1 187 671
41	Belarusian	be	166 775	84 971	1 711 658	253 243
42	Slovenian	sl	164 036	178 042	8 497 867	1 491 437
43	Greek	el	160 482	271 125	34 866 919	6 330 938
44	Galician	gl	155 573	96 617	2 533 863	512 368
45	Azerbaijani	az	145 060	172 093	12 826 807	1 748 834
46	Urdu	ur	144 942	93 377	2 916 140	506 414
47	Simple English	simple	144 053	823 355	19 179 047	9 071 802
48	Norwegian (Nynorsk)	nn	142 635	95 945	1 733 721	563 079
49	Uzbek	uz	130 990	44 264	3 256 673	569 355
50	Thai	th	130 723	349 695	63 983 646	14 758 190
51	Hindi	hi	130 443	444 004	56 017 398	17 087 729
52	Latin	la	130 327	117 110	1 086 052	173 591
53	Georgian	ka	127 899	109 531	8 642 199	1 147 871
54	Volapük	vo	122 757	26 048	266 020	38 888
55	Tamil	ta	121 501	152 024	8 357 708	2 295 703

Each subject (a thing, a man, an event etc.) can be described in each language version independently. Even the largest English Wikipedia does not contain information about all subjects. As we can see in Figure 1, there are over 15 million unique subjects among 55 considered language versions. This can be explained by the fact that some issues may be more common in smaller geographical areas, hence the probability of finding more information on a given topic in the relevant language versions (other than English). So, we can find topics that are not covered in English at all and appear in less-developed versions of Wikipedia [7,25].
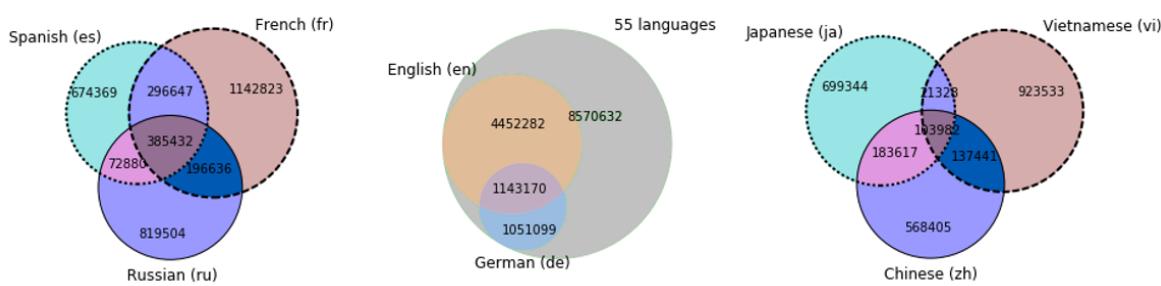
Figure 1. Topic overlaps of articles in various language versions of Wikipedia. Source: own calculation based on Wikipedia dumps in April, 2019. Over 175 thousand of interactive combinations of these Venn diagrams can be found on the Web page: http://data.lewoniewski.info/computers/vn1/

2. Quality Measures

Diverse approaches to defining information by researchers lead also to inconsistencies in defining the notion of its quality. In general, quality of information can be defined as fitness for use [26,27].

In order to define the quality dimensions in Wikipedia, one should take into account the similarity of this website with traditional encyclopedias and Web 2.0 services. On the one hand, content in Wikipedia is created to be a reference point, in an encyclopedic style. According to various studies it has comparable accuracy to other traditional encyclopedias [28,29]. The quality of an article in a traditional encyclopedia can be defined by 7 dimensions: authority, completeness, format, objectivity, style, timeliness, uniqueness [30,31]. On the other hand, Wikipedia is built in a way to allow collaboration between users. It it therefore based on Web 2.0 technologies, which have the following quality dimensions: accessibility, completeness, credibility, involvement, objectivity, readability, relevance, reputation, style, timeliness, uniqueness, usefulness [31,32].

Considering the quality criteria adopted by the Wikipedia community and previously described characteristics of traditional encyclopedia and Web 2.0 documents, we can choose the following quality dimensions for the Wikipedia articles: completeness, credibility, objectivity, readability, relevance, style, timeliness. Figure 2 shows coverage between quality dimensions of the Web 2.0, traditional encyclopedia and Wikipedia.
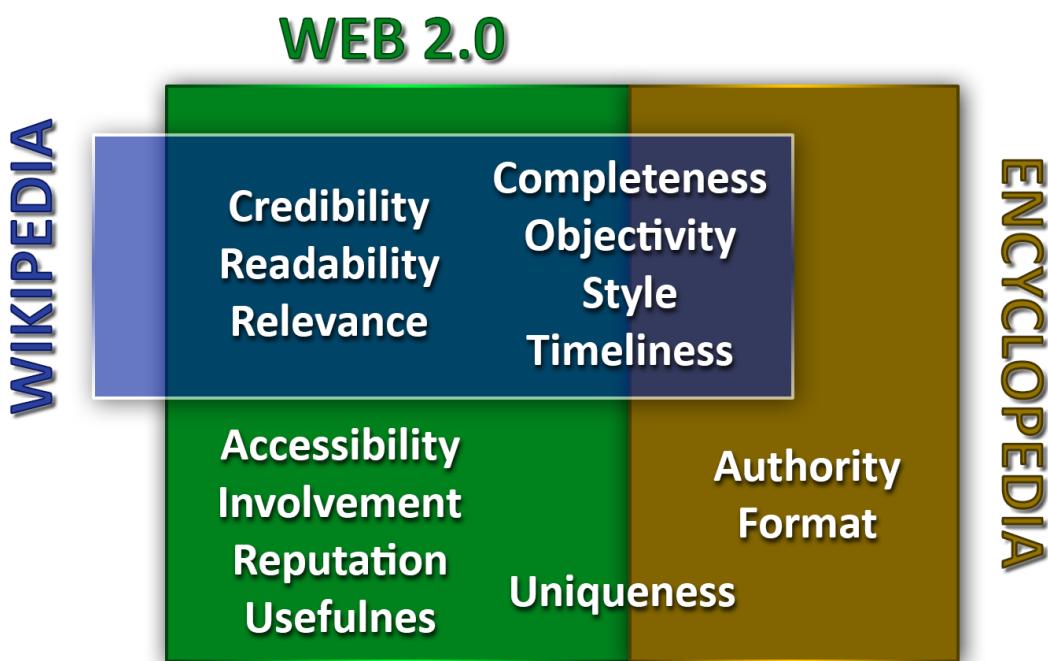
Figure 2. Quality dimensions of the Web 2.0, encyclopedias and Wikipedia. Source: own work based on [31]

Each quality dimension contains a specific set of features (measures). Some features can be related to multiple quality dimensions. There are different ways to define and extract features of the Wikipedia articles. Based on the literature and own experiments, we focused on one of the important features, which can show quality of Wikipedia article from different dimensions.

The length of the text can be measured in various ways - most often it is represented by the length in bytes, the number of letters or words [12,22,33–46]. Length of the article is related to completeness and may indicate the presence of relevant facts and details in its articles.

High-quality articles are expected to use reliable sources [47]. Readers of encyclopedias must be able to check where the information comes from [48]. Therefore, one of the most commonly used reliability measures is the number of references in Wikipedia articles [12,18,22,34–36,44,46,49–53]. References are related to credibility of the article. Our previous research has shown that there are also ways to analyze not only the quantity but also the quality of the references [23].

The length of text can be positively correlated with the number of references but it is important that all relevant facts in Wikipedia should be supported by reliable sources. For this purpose, the reference density can be calculated as the number of references divided by the length of text.

Wikipedia articles must provide information in a fair and impartial manner. In this case, we can take into account information, presented graphically - images [12,18,22,33,36,41,44–46,49,50,54–56]. On the one hand, pictures can help to assess the objectivity of the presented material. On the other hand we can also measure completeness (because articles on a specific topic should contain images) and style (because the authors decided to add more photos instead of writing long text).

High quality content must be prepared in accordance with the guidelines of Wikipedia regarding the style that applies to, among others, organization and structure of the article. Therefore, one of the most-simplest and most popular measures of this dimension is the number of sections in the article [12,16,18,36,38,44,46,49–51].

Quality measures mentioned before can be combined to build synthetic measure for evaluation of Wikipedia articles. Unlike most methods in this area, a synthetic measure can assess the quality of Wikipedia articles on a scale from 0 to 100 [22]. So, we can compare quality of articles between different language chapters, which can have own quality grading scheme (see 1).

Synthetic measure encompasses normalized values of each of 5 features: length, number of references, reference density, number of images, and number sections. Normalization of these features depends on language chapter of Wikipedia, since it uses thresholds, which depend on the best articles in the considered language version[22]. Every considered language of Wikipedia has a special distinction for articles of the highest quality – equivalents to FA and GA grades in English version.

For each language version of Wikipedia, each feature could play an important role in assessing the quality; therefore we first counted the normalized metrics average (NMA) by the following formula:

$$NMA = \frac{1}{c} \sum_{i=1}^{c} \hat{m}_i, \tag{1}$$

where $\hat{m}_i$ is a normalized measure m_i and c is the number of measures.

Next we took into account the number of quality flaw templates (*QFT*) in the considered article (if they existed) and our final formula for the quality measure reads as follows:

$$QualityScore = NMA - NMA \cdot 0.05 \cdot QFT \tag{2}$$

3. Popularity Measures

Popularity of an article can be determined with measures reflecting the demand for information contained in it by the readers and Wikipedia authors. Popularity can play an important role in quality estimation in specific language versions of Wikipedia [13,18]. Greater number of users reading an article can affect the speed of noticing and correcting errors, therefore changes can be made more often (including update of the information).

The popularity of the article can be measured based on the number of visits on the page [18,22]. For example, one of the studies compared reptiles species' page view numbers across languages and in relationship to their spatial distribution and various biological attributes [57]. For each page of Wikipedia, daily page views statistics are available on special online service [58] and Wikimedia dumps [59]. We used dumps to analyze popularity of over 39 million articles in considered language versions of Wikipedia.

Authors' interest can be measured with number of unique authors of the Wikipedia articles. Each user editing articles on Wikipedia can have their own experience, level of knowledge and can adhere to a certain world view. In this regard, it can be assumed that greater number of authors can positively influence the objectivity of the article, since it may contain different points of view on a particular question. On the other hand, the number of authors of the article also can indicates the level of relevance of the article among the Wikipedia community. To sum up, articles created by a larger number of people may be more objective, hence one of the measures leverages in our research is the number of unique authors [12,18,33,41,43–46,51,52,54,55,60–63].

4. Article Categories and Semantic Classification

4.1. Main Topic Classifications

Wikipedia has extensive category network and each article can be annotated with multiple categories, organized into an "ontology of topics" [64]. Each language chapter can define own structure and hierarchy of categories. As a result in some language versions that structure is often too fine-grained to be directly analyzed [65], which may make it difficult to determine the number of possible topics to deal with.

Category structure and alignment of articles to each category can be analyzed based on files from Wikipedia dumps. For these purpose we selected three files from each language version, such as for example for English Wikipedia:

174 - **enwiki-latest-category.sql.gz** – category information; here we use category identifiers and their
175   names;
176 - **en-latest-categorylinks.sql.gz** – wiki category membership link records; here we use information
177   about source page ID and destination category name;
178 - **en-latest-page.sql.gz** – base per-page data; here we use pages ID, title and information about
179   namespaces to identify articles (ns 0) and category (ns 14) pages.

180   For further research we extracted information about over 10 million articles in 55 language
181 versions and analyzed about 400 million links from articles to categories and over 26 million links
182 between categories. General statistics about categories is presented in table 2. Category density shows
183 how number of unique categories per number of articles in particular language version. The highest
184 value of this indicator has Urdu Wikipedia - 1.23. This means that at least one unique article can be
185 found for each unique category. The largest English Wikipedia is in the middle in the ranking of the
186 value of this indicator.

Table 2. Number of categories, number of links from articles to categories and between categories in 55 lanquage versions of Wikipedia. Source: own calculations in April, 2019

Wikipedia language	Number of categories		Category density	Number of links		Average number of categories per article
	All	Without page		From articles to categories	Between categories	
Arabic (ar)	576872	6368	0.806	21548319	1982157	30.102
Azerbaijani (az)	65627	2104	0.452	906108	127144	6.246
Belarusian (be)	115205	33807	0.691	1182398	193168	7.090
Bulgarian (bg)	68898	2624	0.271	1291378	150452	5.079
Catalan (ca)	75951	168	0.124	2672097	179483	4.379
Czech (cs)	140757	665	0.331	2730698	333870	6.412
Danish (da)	62490	5005	0.249	1861533	156608	7.420
German (de)	354701	29	0.155	12255563	886269	5.356
Greek (el)	60056	3826	0.374	1218241	156199	7.591
English (en)	1711545	97	0.293	127118195	5545938	21.782
Esperanto (eo)	83331	15727	0.325	1136030	184428	4.429
Spanish (es)	398828	23074	0.263	9103226	903999	6.011
Estonian (et)	29889	441	0.158	553027	53933	2.915
Basque (eu)	73827	19206	0.222	1497904	170504	4.498
Persian (fa)	499231	37	0.743	9748824	1568018	14.516
Finnish (fi)	72006	280	0.158	2707673	157913	5.953
French (fr)	425707	76	0.203	38654880	2583394	18.453
Galician (gl)	62109	577	0.399	689762	120190	4.434
Hebrew (he)	71150	25	0.295	2310076	170736	9.588
Hindi (hi)	54785	30507	0.420	593496	50673	4.550
Croatian (hr)	19065	53	0.093	503920	32903	2.467
Hungarian (hu)	60203	30	0.134	2895750	111067	6.453
Armenian (hy)	87522	25729	0.353	1601227	136013	6.449
Indonesian (id)	186977	102406	0.408	5279994	185266	11.528
Italian (it)	348216	32	0.229	14715516	847583	9.690
Japanese (ja)	232881	20231	0.203	8060212	551980	7.034
Georgian (ka)	65047	15317	0.509	435646	103973	3.406
Kazakh (kk)	45512	23083	0.203	1660294	41958	7.416
Korean (ko)	268761	20773	0.596	4462341	652764	9.897
Latin (la)	38187	89	0.293	628280	76726	4.821
Lithuanian (lt)	24721	316	0.127	541911	45874	2.786
Malay (ms)	91578	62870	0.281	1393588	59264	4.280
Dutch (nl)	114899	10	0.059	10060345	320354	5.126
Norwegian (Nynorsk) (nn)	88804	18156	0.623	789450	158280	5.535
Norwegian (Bokmål) (no)	148816	6509	0.294	4182237	340251	8.257
Polish (pl)	205391	206	0.154	5310093	399299	3.994
Portuguese (pt)	316318	11293	0.314	9346482	751718	9.273
Romanian (ro)	115325	26231	0.293	3398779	274858	8.639
Russian (ru)	469180	53068	0.305	17351449	929165	11.271
Serbo-Croatian (sh)	45527	374	0.102	1520947	101515	3.397
Simple English (simple)	40052	477	0.278	778386	101112	5.403
Slovak (sk)	70586	76	0.308	919689	199717	4.014
Slovenian (sl)	77146	21649	0.470	1078180	119567	6.573
Serbian (sr)	59254	10899	0.096	4355457	106286	7.045
Swedish (sv)	354075	16	0.094	20002023	639059	5.336
Tamil (ta)	30477	7661	0.251	483546	41080	3.980
Thai (th)	73106	25130	0.559	922356	118369	7.056
Turkish (tr)	226145	10383	0.694	2322792	542366	7.133
Ukrainian (uk)	248614	46181	0.277	7008669	538437	7.818
Urdu (ur)	178271	8836	1.230	1048967	775590	7.237
Uzbek (uz)	12026	4001	0.092	832321	12758	6.354
Vietnamese (vi)	276936	101173	0.230	7745566	476364	6.427
Volapük (vo)	2440	269	0.020	353343	2878	2.878
Chinese (zh)	395448	101111	0.376	12793208	716798	12.162
Min Nan (zh-min-nan)	32592	14516	0.143	608969	46280	2.670

187   Another indicator that can be useful to analyze how often Wikipedia users assign different
188 categories to describe each article is the average number of categories per article. Based on data from
189 table 2 we can define top 3 leaders of the highest value of this indicator: Arabic with 30, English with
190 21 and French with 18 categories per article.
191   We can also notice that in some language versions of Wikipedia there is a large number of
192 categories that do not have own page that describes these categories and point to the parent category.
193 The highest values has Vietnamese, Chinese and Indonesian Wikipedia - about 100 thousand categories
194 without pages. For first two languages with about 1 million articles this is one fourth and one third of all
195 categories respectively. In Indonesian with about 460 thousand articles it is about half of all categories.
196 For comparison, the largest English version with over 5 million articles has only 97 categories without
197 page.
198   The so called main categories are present in majority of considered languages. This applies
199 mainly to those categories that are at highest levels in the polyhierarchy. One of the main categories
200 are presented at special page "Category:Main topic classifications" [66]. Based on this page, we can
201 identify 38 categories on specific topics in the English Wikipedia. Table 3 shows names of these
202 categories with number of the considered language versions.

Table 3. List of the categories in „Category:Main topic classifications" in English Wikipedia with number of the considered language versions.

No.	Category name	Number of considered language versions
1	Education	55
2	Geography	55
3	History	55
4	Mathematics	55
5	Music	55
6	Philosophy	55
7	Religion	55
8	Science	55
9	Society	55
10	Sports	55
11	Arts	54
12	Organizations	54
13	People	54
14	Politics	54
15	Culture	53
16	Law	53
17	Technology	53
18	Health	52
19	Military	52
20	Entertainment	51
21	Events	51
22	Food and drink	51
23	Government	49
24	Nature	49
25	Crime	48
26	Business	47
27	Life	47
28	Academic disciplines	45
29	Human behavior	44
30	Knowledge	44
31	Concepts	43
32	Language	39
33	Objects	37
34	Mind	28
35	Humanities	27
36	World	27
37	Economy	17
38	Universe	5

Due to the fact that some topics may be not available in other language chapters of Wikipedia, we decided to select only those categories that have appropriate category in more than 80 percent of the considered languages (over 44 language versions of Wikipedia). We also did not consider category "Academic disciplines" due to the large coverage of the subcategories with main categories. As a result, 27 main categories were selected for the research.

As mentioned before, the category structure is a complex and ever-changing, as it can be edited by any person – users can add or change a category assignment to other category. The resulting category structure is noisy [64], sparse and it contains duplications and oversights [65]. So, we can also face the situation that categories are repeated at different levels of the tree, in which the root can be a different other main categories (one of the 27 considered). In order to avoid such situations, we cut off those branches that were found at higher levels. Figure 3 shows example of such procedure, when subcategory „Food and Drink" is found at different levels of the tree and only one remains, which is at the highest level.
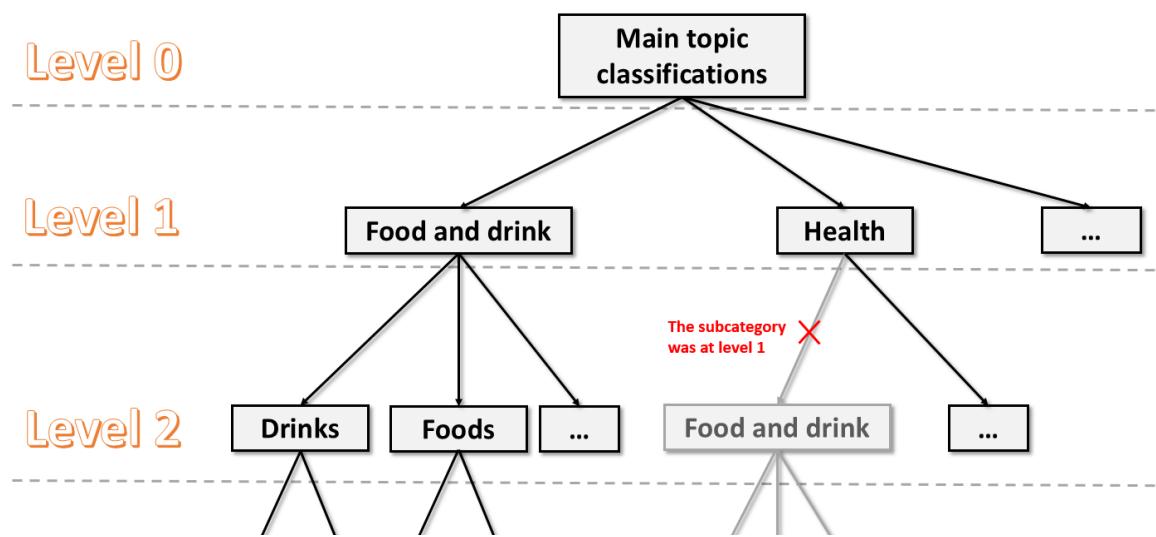
Figure 3. Occurrence of similar sub-categories in the English Wikipedia category polyhierarchy. Source: own work based on Wikipedia dumps from April 2019.

216   If we count articles in English Wikipedia in each of considered main categories, we will see that
217   almost 15% are written about people. Pie chart 4 shows shares of articles in English Wikipedia in 27
218   considered categories.
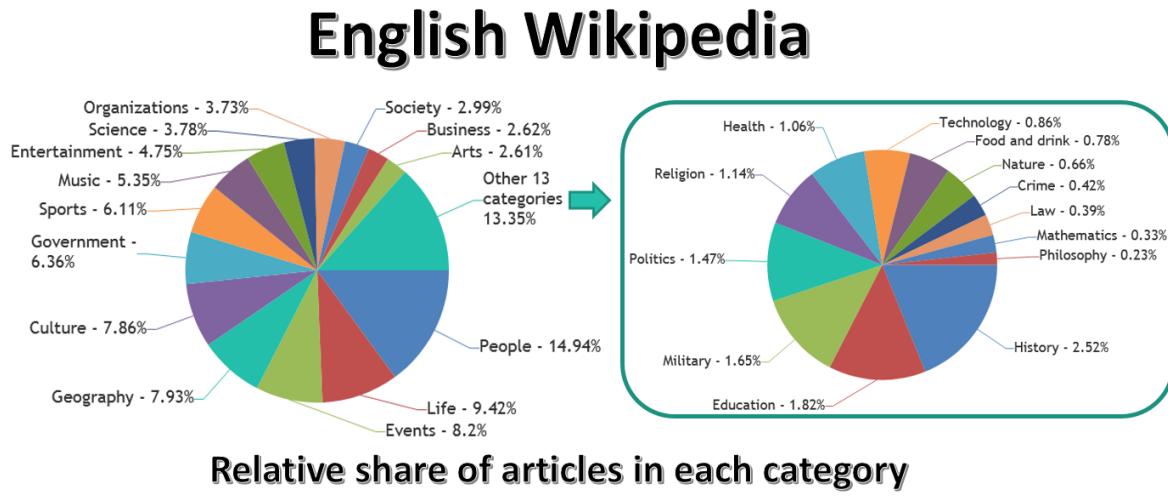
Figure 4. Shares of articles in each category in English Wikipedia. Source: own calculation based on Wikipedia dumps in April, 2019.

219   Heatmap 5 shows distribution of the articles in each topic within each considered language
220   version of Wikipedia.
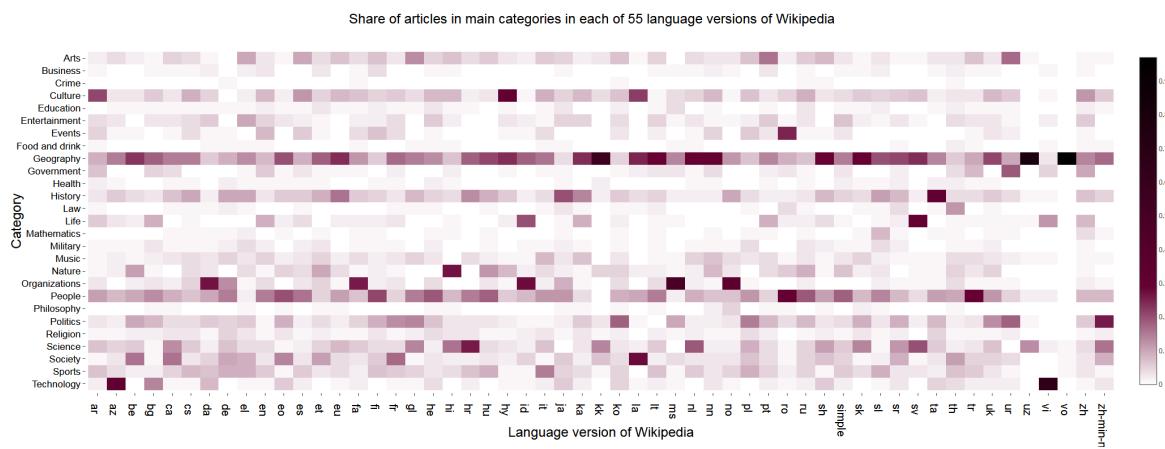
Figure 5. Share of articles in main categories within each of 55 language versions of Wikipedia. Source: own calculation based on Wikipedia dumps in April, 2019. More detailed and interactive chart can be found on the Web page: http://data.lewoniewski.info/computers/heatmap-cat-art

After combining articles from all considered language versions to particular category we concluded that the largest number of articles are in one of two categories: Geography (12.68%) and People (11.48%). Pie chart 6 shows which part of the articles in all considered Wikipedia languages belongs to each of 27 main categories.
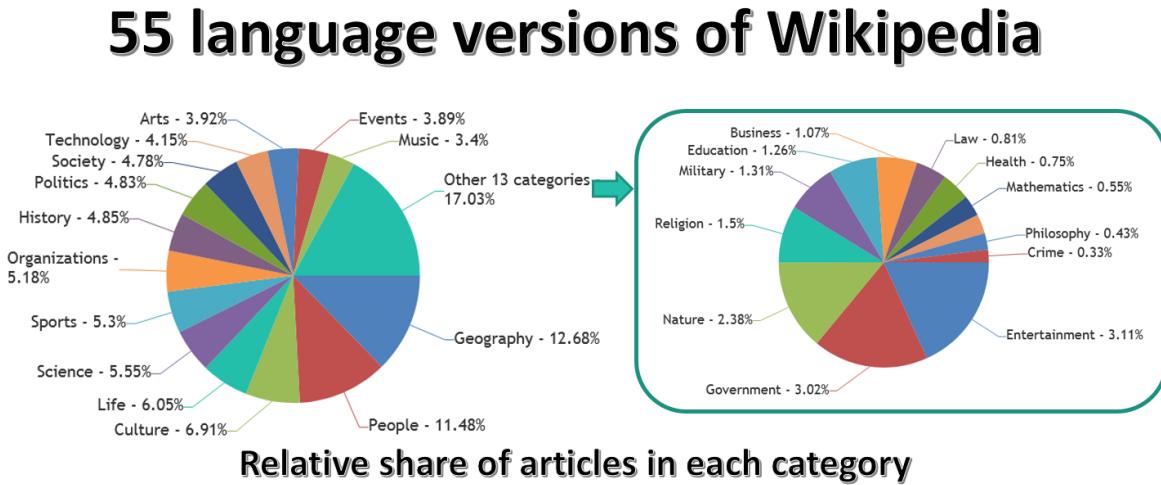
Figure 6. Shares of articles in each category in 55 language versions of Wikipedia. Source: own calculation based on Wikipedia dumps in April, 2019.

As we mentioned before, in some language versions there is a relative high average number of articles assigned to each article. This may increase the possibility of an article falling into more than one main category. Therefore, we decided to show how it looks like in leading language versions (Arabic, English, French) with the highest average number of categories in articles in Figure 7.
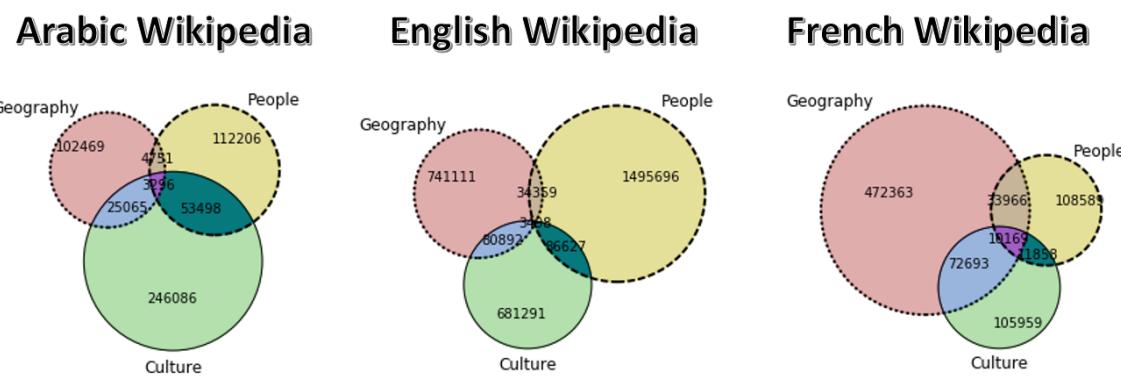
Figure 7. Coverage of articles between selected main topics in Arabic, English and French Wikipedia. Source: own calculation based on Wikipedia dumps in April, 2019. Over million of interactive combinations of these Venn diagrams (each main categories and language versions) can be found on the Web page: http://data.lewoniewski.info/computers/vn2/.

4.2. Semantic Classification

The second approach to topic assignment of the Wikipedia articles used Wikidata and DBpedia. Wikidata is a collaboratively edited knowledge base [67]. DBpedia is the semantic database that extracts structured, multilingual knowledge from Wikipedia [68,69]. The data from this open databases are widely used in a number of domains: web search, life sciences, maritime domain, art market, digital libraries, business networks and others [70–73]

DBpedia uses own ontology with defined properties ans classes organized into a hierarchy. DBpedia gives names in English to each class, such as „Place", „Species", „Person" etc. Wikidata gives unique identifier to each class, for example class „city" is marked as Q515, „human" as Q5, „Organization" as Q43229. Another difference between these databases lies in the number of classes and placing these classes in an ontology. Wikidata has over 300 thousand classes [74], at the same time DBpedia ontology consist about 800 classes [75].

A significantly larger number of classes in Wikidata can lead to difficulties in finding a list of objects on a particular topic. For example, if we want to find all cities, it is not enough to take into account only one class Q515 (city), because city can also be described by Q1637706 (city with millions of inhabitants), Q5119 (capital), Q2264924 (port city), Q58339717 (city of India), Q174844 (megacity) and other identifiers. This variety of classes leads to significantly fewer instances in each class in Wikidata than in DBpedia [74].

We should consider also way of assigning a class to objects in these semantic databases. DBpedia extracts information from Wikipedia infoboxes, and identifies classes based on name of the infobox and values of some special parameters. Thus, articles with the same infobox name often go to the same class. In Wikidata, items can be edited by everyone, so the different classes can be attached to similar objects.

There are some papers that study differences between DBpedia and Wikidata [74,76,77]. Each has own advantages, so we decided to use combined data to divide articles into separate classes: actor, automobile, business, city, film, footballplayer, human, programming, university, videogame, and website. One of the advantages of such a classification approach by topic is that we are dealing here with more explicit assigning of articles to specific classes and each language version has at least several representatives of each class.

5. Quality and Popularity Assessment

Following the procedures described in previous sections, we extracted over 100 million values of features characterizing articles in all analyzed languages. These values were then used to calculate the synthetic measure that assesses quality of the content. Next we grouped articles by 27 main categories

and 55 languages. Within each of obtained groups (almost 1500) we calculated sum of all synthetic measure values and divided it by the number of articles. The resulting average quality of articles is presented in Figure 8.
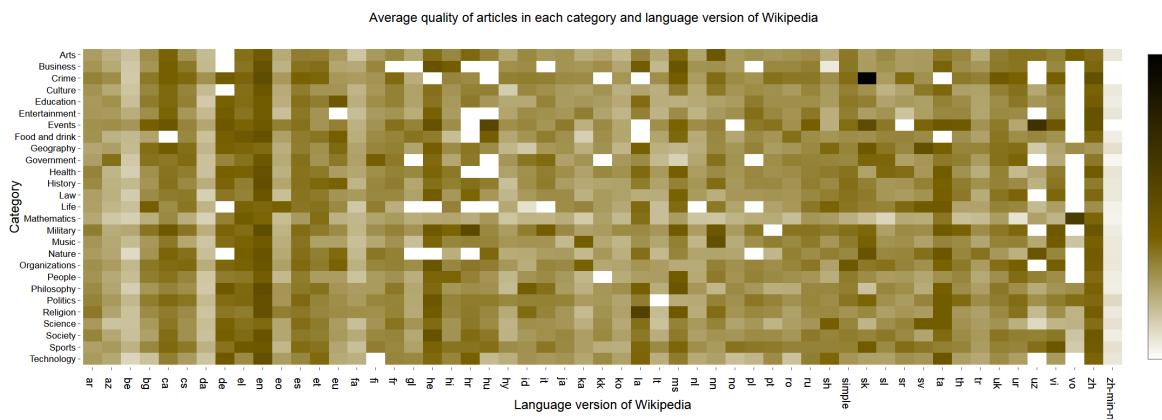
Figure 8. Average quality of articles in each category and language version of Wikipedia. Source: own calculation based on Wikipedia dumps in April, 2019. More detailed and interactive chart can be found on the Web page: http://data.lewoniewski.info/computers/heatmap-cat-quality.

The highest average quality have articles in category Crime in Slovak Wikipedia (sk) - 63.92 points. This is due to the fact that in this language chapter only a few articles fall into this category and they are generally well written according to studied features. Articles about crime also have relatively higher quality scores in English (en) and Chinese (zh) Wikipedia.

Second place in ranking of the highest average quality have articles about events in Uzbek Wikipedia (uz) - 43.96 points. Again, this main category does not contain much content – there are only 31 articles. If we take into account the development of the Uzbek Wikipedia (about 130 thousand of articles), we can conclude that this topic is rather important for local community of editors. Articles about events also have relatively higher quality scores in Hungarian (hu), Slovak (sk), Hebrew (he) and Chinese (zh) Wikipedia.

Third place regarding the quality is taken by articles about mathematics in Volapük Wikipedia - 39.63 points. However, in this language chapter the category consist only 2 articles. Latin Wikipedia (la) has the fourth place with average quality of the articles about religion - 37.77.

If we take into account the most developed English Wikipedia, the highest average quality of articles can be found in categories: Philosophy, Crime, Military and History. Generally, we can conclude that English Wikipedia articles usually have high value of average quality in different topics.

Next heatmap presented in Figure 9 shows average page views per article in year 2018 for each category and language version of Wikipedia.
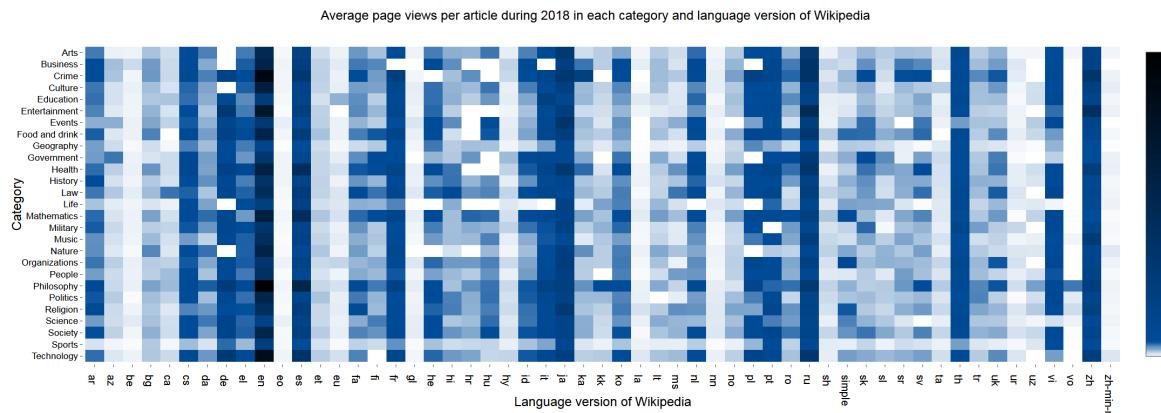
Figure 9. Average page views per article in year 2018 for each category and language version of Wikipedia. Source: own calculation based on Wikipedia dumps. More detailed and interactive chart can be found on the Web page: http://data.lewoniewski.info/computers/heatmap-cat-views

Generally, page views values are higher for the most popular languages. This led to the fact that the first 11 positions in the rank are occupied by English (en) Wikipedia. The most popular topic in this language is Philosophy. One of the highest average popularity in this language characterizes also articles about crime, technology, entertainment, mathematics, culture, and health. All these categories had at least 20 thousand page views in year 2018.

Second most popular language version is Spanish (es). Similarly to English, the most visited category is Philosophy. It is also worth to mention two other popular categories in this language: Mathematics and Health. Articles in three mentioned main categories of Spanish Wikipedia have at least 14 thousand page views per year.

Third place is taken by Russian (ru) Wikipedia and category Entertainment, with about 16 thousand page views per year. Entertainment is also the most popular topic in Chinese (zh) Wikipedia.

Finally, heatmap in Figure 10 shows average number of authors per article in 2018 in each category and language version of Wikipedia.
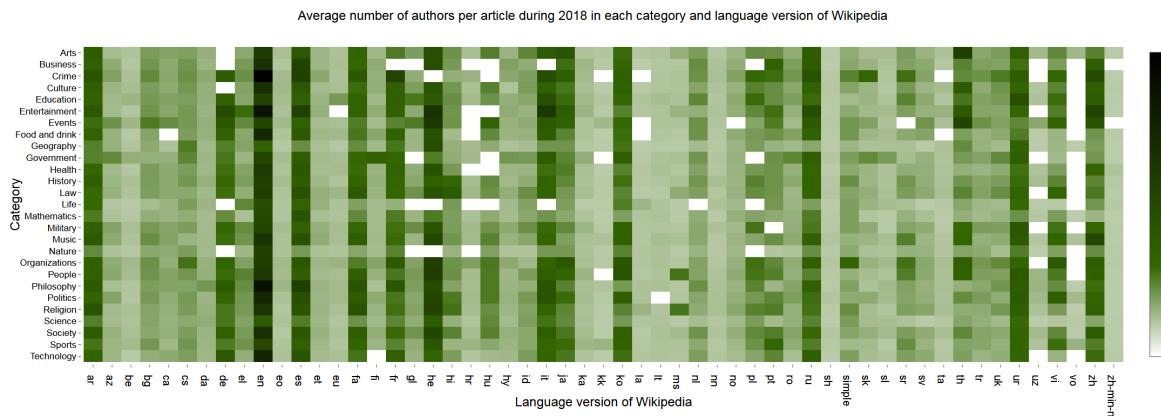
Figure 10. Average number of authors per article during 2018 in each category and language version of Wikipedia. Source: own calculation based on Wikipedia dumps. More detailed and interactive chart can be found on the Web page: http://data.lewoniewski.info/computers/heatmap-cat-authors

As in the case of the popularity of page views, in the ranking of authors' interests topics from English Wikipedia topped the ranking. Here we have such popular categories as Crime, Philosophy, Entertainment. Articles about topics were edited at least by 8 authors during the 2018 year.

Second language version that has most active authors is Hebrew (he) Wikipedia with articles about entertainment. During a year at least 6 authors have edited each article in this topic. Entertainment is

301   also popular among authors in Italian (it), Spanish (es) and Chinese (zh) Wikipedia. At the same time
302   Italian Wikipedia we can met as the third language in the authors' interest ranking.
303   Table 4 presents main categories which has the highest value of average quality, average popularity
304   and authors' interest in each language version of Wikipedia.

Table 4. Main category of articles with the highest value of average quality, average popularity and authors' interest in each language version of Wikipedia. Source: own calculations.

Language version	Quality	Popularity	Authors' interest
Arabic (ar)	Religion	Religion	Religion
Azerbaijani (az)	Government	Government	Government
Belarusian (be)	Government	Business	Events
Bulgarian (bg)	Events	Food and drink	Life
Catalan (ca)	Events	Law	Events
Czech (cs)	Organizations	Health	Crime
Danish (da)	Philosophy	Philosophy	Crime
German (de)	Entertainment	Entertainment	Events
Greek (el)	Entertainment	Health	Food and drink
English (en)	Crime	Philosophy	Philosophy
Esperanto (eo)	Philosophy	Events	Life
Spanish (es)	Philosophy	Philosophy	Crime
Estonian (et)	Crime	Food and drink	Crime
Basque (eu)	Education	Education	Education
Persian (fa)	Religion	Philosophy	Religion
Finnish (fi)	Government	Government	Government
French (fr)	Crime	Crime	Crime
Galician (gl)	Education	Events	Food and drink
Hebrew (he)	Entertainment	Events	Events
Hindi (hi)	Law	Law	Business
Croatian (hr)	Organizations	Mathematics	Military
Hungarian (hu)	Events	Events	Events
Armenian (hy)	Government	Government	Crime
Indonesian (id)	Arts	Business	Philosophy
Italian (it)	Entertainment	Education	Military
Japanese (ja)	Organizations	Events	Events
Georgian (ka)	Government	Crime	Music
Kazakh (kk)	Sports	Philosophy	Health
Korean (ko)	People	Business	Military
Latin (la)	Religion	Religion	Religion
Lithuanian (lt)	Education	Mathematics	Sports
Malay (ms)	People	Law	Business
Dutch (nl)	Education	Philosophy	Events
Norwegian (Nynorsk) (nn)	History	History	Music
Norwegian (Bokmål) (no)	Crime	Mathematics	Sports
Polish (pl)	Crime	Crime	Entertainment
Portuguese (pt)	Business	Health	Crime
Romanian (ro)	Government	Government	Food and drink
Russian (ru)	Entertainment	Entertainment	Events
Serbo-Croatian (sh)	Music	Mathematics	Science
Simple English (simple)	Organizations	Organizations	Organizations
Slovak (sk)	Crime	Crime	Crime
Slovenian (sl)	Government	Government	Government
Serbian (sr)	Crime	Crime	Life
Swedish (sv)	Events	Health	Geography
Tamil (ta)	Entertainment	Philosophy	Technology
Thai (th)	Arts	Military	Events
Turkish (tr)	Events	Politics	Nature
Ukrainian (uk)	Crime	Philosophy	Crime
Urdu (ur)	Education	Military	Organizations
Uzbek (uz)	Events	Philosophy	Events
Vietnamese (vi)	Organizations	Law	Sports
Volapük (vo)	Sports	Philosophy	Mathematics
Chinese (zh)	Entertainment	Entertainment	Crime
Min Nan (zh-min-nan)	Health	Technology	Politics

Finally, we do the similar calculations for articles in semantic classes: actor, automobile, business, city, film, footballplayer, human, programming, university, videogame, website. Figure 11 shows average quality, authors interest and pageviews in 2018 per article in each semantic class and language version of Wikipedia
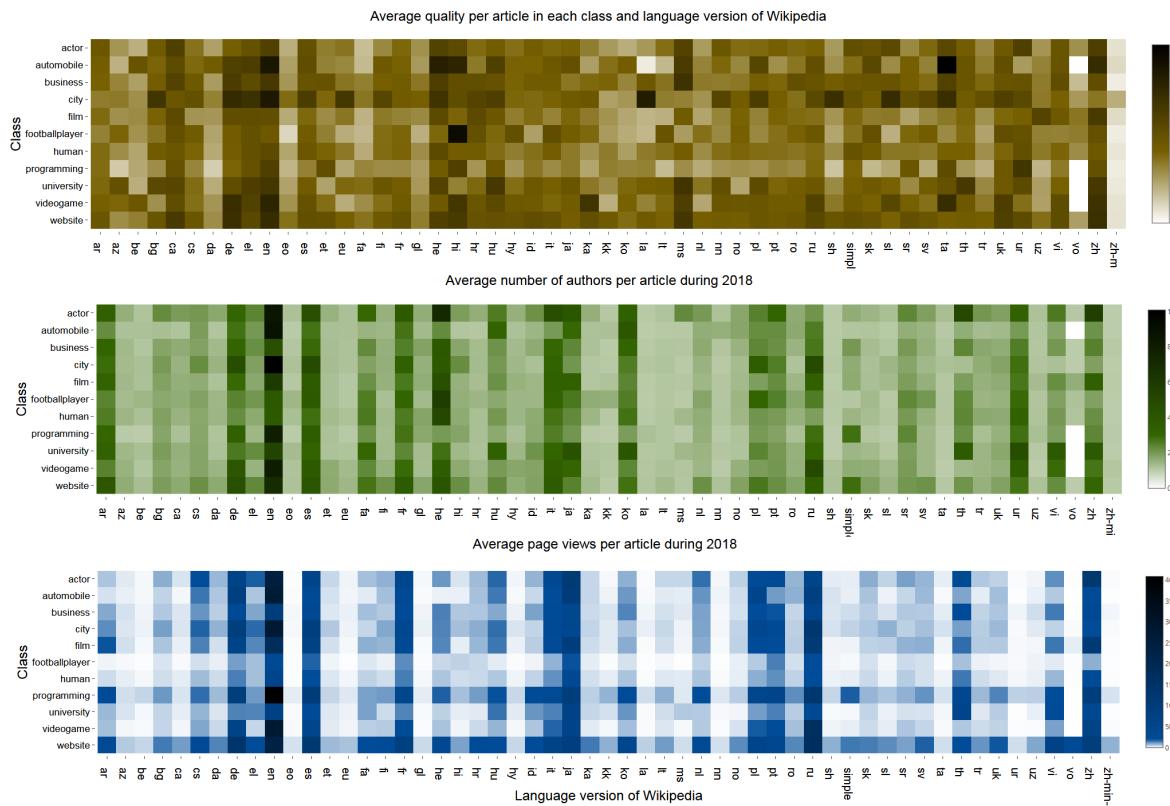
Figure 11. Average quality, authors interest and pageviews during 2018 per article in each class and language version of Wikipedia. More detailed and interactive chart can be found on the Web page: http://data.lewoniewski.info/computers/heatmap-classes

The leader in terms of the value of average quality is Tamil (ta) Wikipedia with articles that describes cars (automobiles) - 43.22 points. The second place in this ranking occupies articles about football players in Hindi (hi) Wikipedia - 40.35 points for quality per article. The third place in quality took English (en) Wikipedia with articles about cars - 37.39 points. Articles about cars have also relative high quality un Hebrew (he), Hindi (hi) and Chinese (zh) Wikipedia - over 31 points. In this quality ranking most often we can met articles about cities in English (en), Latin (la), German (de), Slovenian (sl), Serbo-Croatian (sh), Greek (el) Wikipedia - over 30 points per article.

As for page views, we have similar situation as it was in the case of main category classifications - English Wikipedia has here the highest values. The most popular class in this language versions is programming, which has over 40 thousand page visits per article during 2018. Next the most popular classes with over 23 thousand visits per articles during a year are related to video games, cities, cars, actors, web sites. Second language version that we can met in the top of the popularity ranking - Russian (ru) Wikipedia with articles about web sites and video games. Next is German (de) version with articles about web sites.

Authors' interest ranking of the classes shows also leadership of the English (en) Wikipedia. Here the highest number of authors per article in 2018 has articles about cities - over 10 authors edits each article during a year. Next popular articles. Popular among authors are also articles about cars, actors, video games and programming languages - over 8 authors per article during a year. Next we can met in this ranking articles from Hebrew (he) Wikipedia which describes actors - over 7 authors per

article during past year. Relatively high interest among authors we can observe Chinese (zh), Thai (th), Italian (it), Spanish and Japanese (ja) Wikipedia - over 4 authors per article during 2018. Articles about universities has similar values of average authors' interest in English (en), Urdu (ur), Japanese (ja) and Korean (ko) Wikipedia.

Table 4 presents classes which has the highest value of average quality, average popularity and authors' interest in each language version of Wikipedia.

Table 5. Classes of articles with the highest value of average quality, average popularity and authors' interest in each language version of Wikipedia. Source: own calculations.

Language version	Quality	Popularity	Authors' interest
Arabic (ar)	website	website	website
Azerbaijani (az)	website	website	university
Belarusian (be)	footballplayer	programming	automobile
Bulgarian (bg)	actor	website	city
Catalan (ca)	actor	website	website
Czech (cs)	city	website	city
Danish (da)	actor	website	automobile
German (de)	city	website	city
Greek (el)	actor	website	city
English (en)	city	programming	automobile
Esperanto (eo)	footballplayer	website	city
Spanish (es)	city	website	city
Estonian (et)	website	website	programming
Basque (eu)	website	website	city
Persian (fa)	university	website	university
Finnish (fi)	website	website	city
French (fr)	actor	website	website
Galician (gl)	business	website	city
Hebrew (he)	actor	website	automobile
Hindi (hi)	city	website	footballplayer
Croatian (hr)	actor	website	city
Hungarian (hu)	university	website	university
Armenian (hy)	videogame	website	footballplayer
Indonesian (id)	actor	programming	website
Italian (it)	actor	website	footballplayer
Japanese (ja)	university	actor	automobile
Georgian (ka)	footballplayer	website	videogame
Kazakh (kk)	footballplayer	website	website
Korean (ko)	university	website	automobile
Latin (la)	programming	website	city
Lithuanian (lt)	website	website	footballplayer
Malay (ms)	actor	university	business
Dutch (nl)	website	website	website
Norwegian (Nynorsk) (nn)	automobile	website	city
Norwegian (Bokmål) (no)	website	website	videogame
Polish (pl)	city	website	city
Portuguese (pt)	actor	programming	website
Romanian (ro)	website	website	business
Russian (ru)	videogame	website	videogame
Serbo-Croatian (sh)	website	website	city
Simple English (simple)	website	programming	actor
Slovak (sk)	website	website	automobile
Slovenian (sl)	website	website	city
Serbian (sr)	actor	actor	website
Swedish (sv)	website	website	city
Tamil (ta)	actor	website	automobile
Thai (th)	actor	university	university
Turkish (tr)	actor	website	city
Ukrainian (uk)	actor	website	videogame
Urdu (ur)	university	programming	programming
Uzbek (uz)	film	website	film
Vietnamese (vi)	university	website	videogame
Volapük (vo)	film	website	film
Chinese (zh)	actor	actor	automobile
Min Nan (zh-min-nan)	videogame	website	city

6. Local and global rankings of the Wikipedia articles

Based on assessment of over 39 million articles we built rankings of articles in each language version of Wikipedia separately and also leveraged knowledge about links between languages to build multilingual rankings. Page views and authors' interest can change in time, therefore we also conducted calculations for individual months – from January 2018 till March 2019. Thus, it is possible to observe changes of preferences of Wikipedia authors and readers.

Popularity measure for an article can be calculated as a median value of the page visits per day, as it was proposed in previous study [22]. For the purpose of ranking, if median is not sufficient to sort articles we use additional criterion – total number of visits in selected month is considered. Another measure, authors' interest, is calculated as a number of unique authors who introduced changes to an article during selected period (e.g. month). If the number of authors for selected articles is the same, we further sort based on total number of the page visits.

Tables 6, 7 and 8 show top 3 articles about cars, films and video games respectively with the highest values of page views and author interest in each period in all considered language version.

Table 6. Top 3 articles about cars with highest value of page views and authors' interest in multilingual ranking, monthly. Source: own calculations.

Month	Page views	Authors' interest
January 2018	Volkswagen Golf BMW 3 Series Audi A4	Lancia Delta Daihatsu Tanto Astra (1954 automobile)
February 2018	BMW 3 Series Volkswagen Golf Audi A4	Opel Corsa Chevrolet Tracker (Americas) Mercedes-Benz W128
March 2018	BMW 3 Series Ford Mustang Volkswagen Golf	Triumph Roadster UD Quon UD Condor
April 2018	Ford Mustang BMW 3 Series Volkswagen Golf	Chrysler Fifth Avenue Oreca 05 Škoda 130
May 2018	Ford Mustang BMW 3 Series Volkswagen Golf	Citroën C1 Cadillac Brougham Mitsubishi Colt
June 2018	Ford Mustang BMW 3 Series Volkswagen Golf	Oka (automobile) Nissan NV200 Alfa Romeo Giulia TZ
July 2018	BMW 3 Series Ford Mustang Volkswagen Golf	Dongfeng Fengshen L60 Nissan NV200 Brabham BT24
August 2018	BMW 3 Series Ford Mustang Volkswagen Golf	Audi Front Vauxhall B-Type Inspiration (car)
September 2018	BMW 3 Series Ford Mustang Volkswagen Golf	SEAT Leon X-PERIENCE Daihatsu Applause Daihatsu Pyzar
October 2018	BMW 3 Series BMW 3 Series (F30) Volkswagen Golf	Cadillac V-16 Citroën C3 Aircross Ferrari 812 Superfast
November 2018	BMW 3 Series Tesla Model S Volkswagen Golf	Citroën C3 Aircross Nissan Kicks Cadillac V-16
December 2018	BMW 3 Series Volkswagen Golf Tesla Model S	Volvo 200 Series Subaru BRAT Citroën C3 Aircross
January 2019	BMW 3 Series Toyota Supra Volkswagen Golf	Opel Rekord P1 SEAT Leon X-PERIENCE Volkswagen CrossBlue
February 2019	BMW 3 Series Volkswagen Golf Ford Mustang	Opel Rekord P1 Geely Binyue Ford Model TT
March 2019	BMW 3 Series Tesla Model S Ford Mustang	Alvis 10/30 MG ZS BMW Z1

Table 7. Top 3 articles about films with highest value of page views and authors' interest in multilingual ranking, monthly. Source: own calculations.

Month	Page views	Authors' interest
January 2018	Black Mirror The End of the F***ing World Star Wars: The Last Jedi	This Is Us (season 3) 9/11 (2017 film) This Is Us (season 1)
February 2018	Black Panther (film) Altered Carbon (TV series) Money Heist	Black Panther (film) Alfie (2004 film) Yo soy Betty, la fea
March 2018	Black Panther (film) The Shape of Water Avengers: Infinity War	The Shape of Water Avengers: Infinity War Black Panther (film)
April 2018	Avengers: Infinity War A Quiet Place (film) Money Heist	Avengers: Infinity War Ready Player One (film) A Quiet Place (film)
May 2018	Avengers: Infinity War Deadpool 2 Black Panther (film)	Avengers: Infinity War Deadpool 2 Solo: A Star Wars Story
June 2018	Jurassic World: Fallen Kingdom Avengers: Infinity War Westworld (TV series)	Jurassic World: Fallen Kingdom Avengers: Infinity War Starship Troopers: Traitor of Mars
July 2018	Ant-Man and the Wasp Avengers: Infinity War The Handmaid's Tale (TV series)	Ant-Man and the Wasp Idol School (2017 TV series) Mamma Mia! Here We Go Again
August 2018	Story of Yanxi Palace Avengers: Infinity War Crazy Rich Asians (film)	Champions (2018 film) Story of Yanxi Palace Charlie St. Cloud
September 2018	Story of Yanxi Palace The Nun (2018 film) The Matrix	The Twelve Chairs (1971 film) The Nun (2018 film) Say "I love you"
October 2018	Venom (2018 film) A Star Is Born (2018 film) The Haunting (TV series)	Venom (2018 film) Bad Boys for Life O.G. (film)
November 2018	Bohemian Rhapsody (film) Fantastic Beasts: The Crimes of Grindelwald Fantastic Beasts and Where to Find Them (film)	Bohemian Rhapsody (film) Ralph Breaks the Internet Fantastic Beasts: The Crimes of Grindelwald
December 2018	Aquaman (film) Spider-Man: Into the Spider-Verse Bohemian Rhapsody (film)	Spider-Man: Into the Spider-Verse Aquaman (film) Bumblebee (film)
January 2019	Glass (2019 film) You (TV series) Aquaman (film)	Viy (2014 film) Aquaman (film) Glass (2019 film)
February 2019	Alita: Battle Angel The Umbrella Academy (TV series) Green Book (film)	Viy (2014 film) Alita: Battle Angel Bohemian Rhapsody (film)
March 2019	Captain Marvel (film) Us (2019 film) Game of Thrones	Heavenly Sword and Dragon Slaying Saber Captain Marvel (film) Laplace's Witch (film)

Table 8. Top 3 articles about video games with highest value of page views and authors' interest in multilingual ranking, monthly. Source: own calculations.

Month	Page views	Authors' interest
January 2018	Assassin's Creed Devilman PlayerUnknown's Battlegrounds	The Surge (video game) Pillars of Eternity II: Deadfire Dirt 4
February 2018	Assassin's Creed Kingdom Come: Deliverance Fortnite	Mortal Kombat: Deception Wild Gunman Teenage Mutant Ninja Turtles: The Hyperstone Heist
March 2018	Fortnite Assassin's Creed Call of Duty	VBS2 Questprobe Star Trek: Judgment Rites
April 2018	God of War (2018 video game) Fortnite Far Cry 5	Fortnite Darksiders III God of War (2018 video game)
May 2018	Fortnite God of War (2018 video game) Assassin's Creed	Fortnite Mega Man (original series) The Lord of the Rings: The Third Age
June 2018	Detroit: Become Human Fortnite Assassin's Creed	Monument Valley 2 Clonk Oddworld
July 2018	Fortnite Detroit: Become Human Assassin's Creed	Verdun (video game) The Lord of the Rings: Aragorn's Quest Star Wars Battlefront: Elite Squadron
August 2018	Fortnite Assassin's Creed World of Warcraft	Blacklight: Tango Down Tokyo Xtreme Racer: Zero Ace Combat Zero: The Belkan War
September 2018	Borderlands: The Pre-Sequel Spider-Man (2018 video game) Fortnite	The Walking Dead (video game series) FIFA 19 Fortnite
October 2018	Borderlands: The Pre-Sequel Assassin's Creed Red Dead Redemption 2	Diner Dash Red Dead Redemption 2 Fortnite
November 2018	Borderlands: The Pre-Sequel Red Dead Redemption 2 Fallout 76	Dramatical Murder Red Dead Redemption 2 Fortnite
December 2018	Borderlands: The Pre-Sequel Fortnite Red Dead Redemption 2	Adventure Time: The Secret of the Nameless Kingdom Starsiege Ecco the Dolphin
January 2019	Borderlands: The Pre-Sequel Fortnite Minecraft	The Walking Dead (video game series) Fortnite Snake Pass (video game)
February 2019	Borderlands: The Pre-Sequel Apex Legends Fortnite	Darwin Project Delta Force (video game) Marvel vs. Capcom
March 2019	Borderlands: The Pre-Sequel Fortnite Sekiro: Shadows Die Twice	Yu Yu Hakusho Makyō Tōitsusen Everything (video game) Puzzle & Action: Tant-R

348   These ranking show that the most popular articles from readers' point of view usually do not
349 match with the priorities of the community of Wikipedia authors. For example, in each month of
350 the considered period we can see the same models of the cars from the readers ranking, such as
351 Volkswagen Golf, BMW 3 Series, Ford Mustang, Tesla Model S. However the authors' interest ranking
352 can show unique titles of articles about cars each months. This may be due to the fact that popular
353 articles are sufficiently developed and do not require significant revisions. When new car models are
354 introduced, there is of course a need to create new high-quality content. In the case of video games we

can observe bigger overlaps between popular articles from readers and writers point of views – for example Fortnite.

Such global quality rankings can show how specific product is popular worldwide. For example, one of the most popular films in 2018-2019 were: Black Mirror, Black Panther, Avengers: Infinity War, Aquaman, Alita: Battle Angel, Captain Marvel, Game of Thrones, Star Wars, and Fantastic Beasts. Among popular video games according to multilingual Wikipedia in 2018-2019 were: Assassin's Creed, PlayerUnknown's Battlegrounds, God of War, Fortnite, Fallout 76, Borderlands: The Pre-Sequel, Red Dead Redemption 2, Apex Legends, Spider-Man, Detroit: Become Human, and Call of Duty.

Various multilingual rankings are implemented in WikiRank service [78], where it is possible to analyze how the position of a particular article has changed in rankings in comparison with the previous period, what is the most popular language version, what is the quality of the popular language version article etc. Figure 12 presents example of the ranking of the articles about films with different parameters.
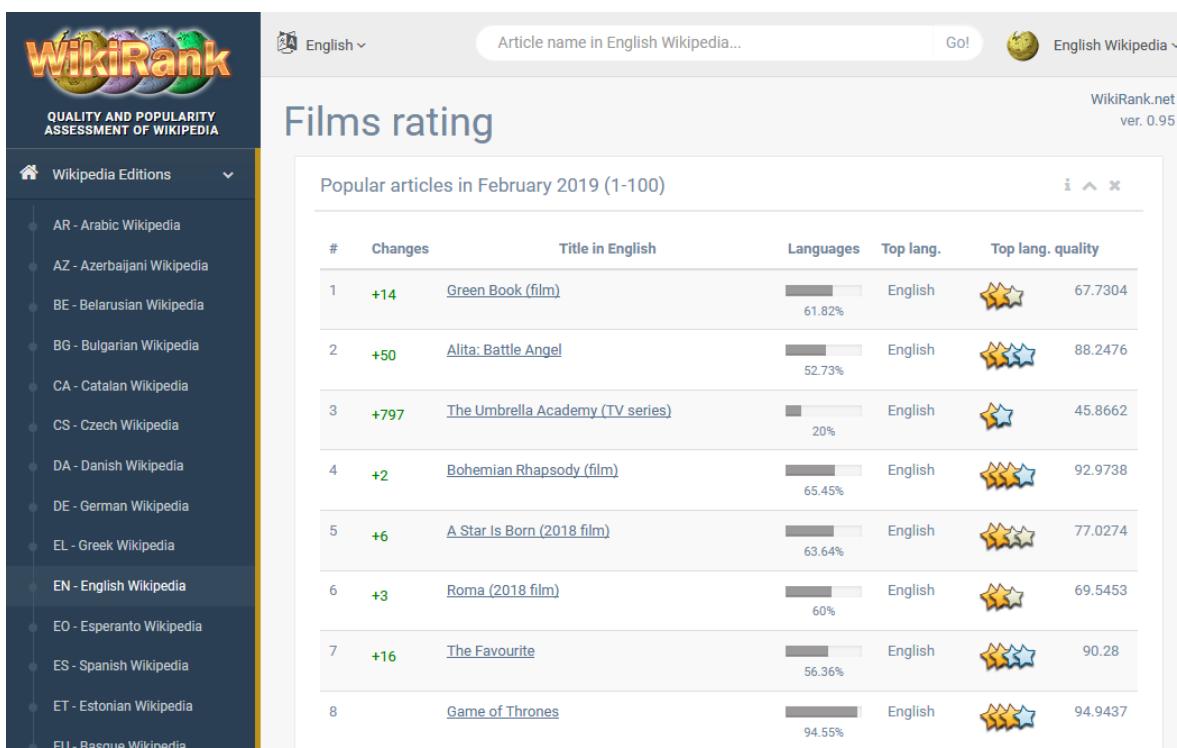
Figure 12. List of the most popular articles about films in multilingual Wikipedia in WikiRank service. Source: https://wikirank.net/top/film

Combination of measures from different languages makes it possible to create global rankings of all articles. Additionally, for each language version there is possible to generate local rankings – here measures from one language can be taken into account. Example of the local ranking with quality distribution of all articles in English Wikipedia is shown in Figure 13

Calculated measures can be gathered to create individual profile for each article in each language version. For example, figure 14 presents such a profile for article „Fortnite" in English Wikipedia on WikiRank with information about places in local and global rankings, quality and popularity scores, and also history of popularity rank.

If an article is written in more than one language, additional ranking of the most popular language versions as well as languages with the highest quality are displayed. Figure 15 shows example of such ranking of the best language versions about Minecraft.

Profiles of Wikipedia articles can also be used to compare the demand for a specific product between various language communities. For example video game Dota 2 is most popular in English,
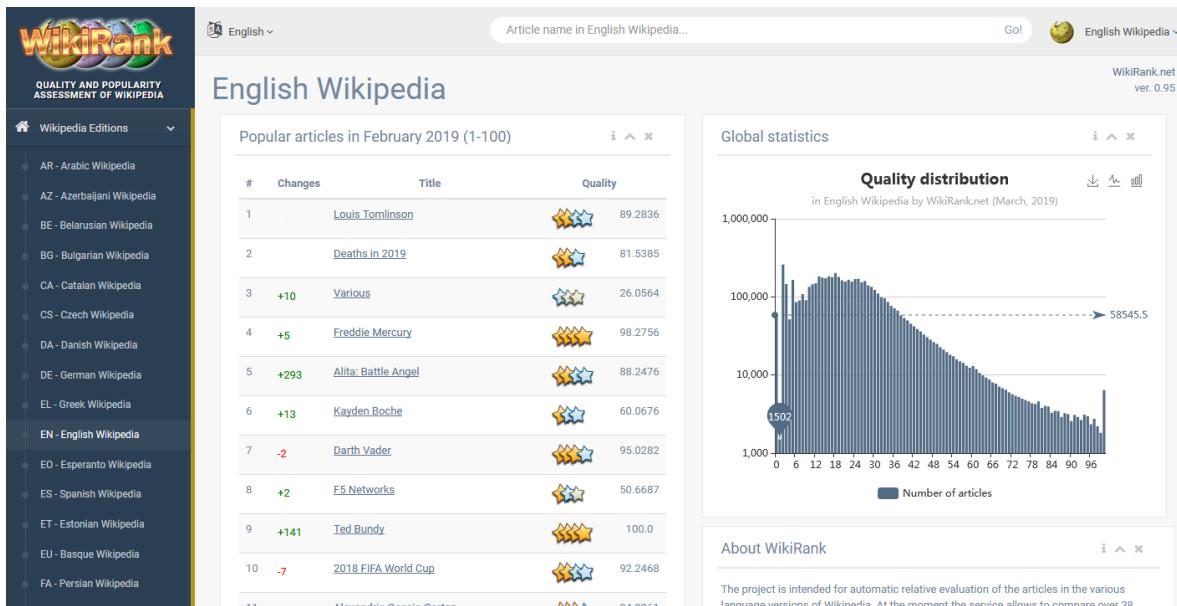
Figure 13. Local ranking with quality distribution of all articles in English Wikipedia in WikiRank service. Source: https://wikirank.net/en/
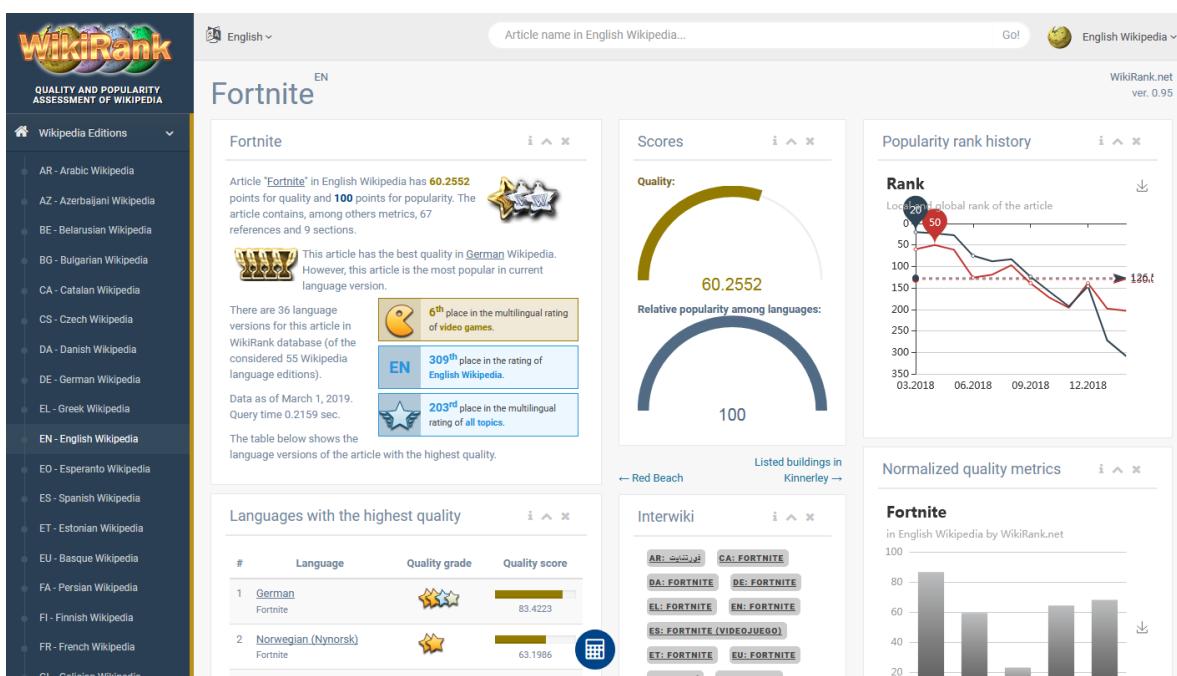
Figure 14. Profile on WikiRank of the article about Fortnite in English Wikipedia with information about places in local and global rankings, quality and popularity scores, history of popularity rank. Source: https://wikirank.net/en/Fortnite

Figure 15. The most popular language versions and languages with the highest quality for article about Minecraft on WikiRank. Source: https://wikirank.net/en/Minecraft

Russian, Chinese, German, and Spanish [79]. Another multilayer video game Counter-Strike: Global Offensive is most popular in English, Russian, German, Spanish and Polish Wikipedia [80].

7. Results and Discussion

During the research we encountered several restrictions, mainly related to the differences between language versions of Wikipedia. For example, as we show in table 3, some main categories do not have links to all considered language versions. This is also true for developed languages. For example, category "Art" in English Wikipedia does not have direct equivalent in German Wikipedia, which uses category "Kunst und Kultur" [81] ("Arts and Culture") to describe part of this topic.

Regarding categories, our experiments showed that each language version has specific ratio between number of articles and number of categories. Additionally, some language version can have a lot of undefined pages for the categories. There is also a difference between number of categories that are assigned to each article. Some languages can use an average of 30 categories to describe one article, while the others are limited to 2-3 categories per article.

Results for authors popularity can be sometimes biased due to temporal or permanent restrictions. According to one of the main principles of Wikipedia anyone can edit content. However, in some particular situations this right can be revoked to protect content from unwanted changes (vandalism) [82]. Each language version can define own levels of page protection. For example, in English Wikipedia there is a full protection, where only administrators can edit an article, and semi-protection, which prevents editing by unregistered users or users that are not confirmed. Each article can be protected for a specified period. As a result, some articles can have less authors' interest than it would in the situation without protection.

In our work we provide classification to main topics according to structure of main categories in English Wikipedia. However, each language can have own definition of whether it is main category. In future, we plan to develop more sophisticated methods to take into account refined category structures.

Some research results are available online at WikiRank service [78]. In research we used some tools that are available on GitHub [83].

8. Conclusions and future work

In this paper we presented results of quality and popularity assessment of articles in multilingual Wikipedia. For this purpose we calculated over 200 million values characterizing quality and popularity of articles in 55 language versions of Wikipedia. Additionally, we analyzed over 10 million categories, over 26 million links between them, and about 400 million links from articles to categories in order

to determine assignment of articles to one of the topics in main classification. In order to assign articles from different languages to various topics we also used semantic databases – Wikidata and DBpedia. Similarly to other studies, we combined data from these sources to have more comprehensive classifications of articles.

Results of the research showed not only how quality and popularity differ for articles from various topics and languages but also how the same topic is developed in different languages of Wikipedia in terms of quality and popularity of content. With regard to popularity, we proposed to pay attention not only at how often users visits certain articles but also what is authors' interest in them. The author's interest measure can be calculated for a language version or can be combined within selected languages. This measure can show what priorities and preferences have Wikipedians in various periods of time. We can also measure correlation between this measure and other measures, such as quality or popularity from readers point of views.

Presented results can be used to build more complex models for quality assessment of information in Wikipedia in different languages and topics. In the future, they can help not only to automatically enrich each less-developed language version of Wikipedia but also can be used to build massive semantic databases with powerful inference system, which can create new knowledge for humanity in a relatively short time.

The work towards more precise assessment of Wikipedia quality will be continued, especially different measures and approaches for quality assessment in Wikipedia and other collaborative knowledge bases can be analyzed. As of April 2019, based on our calculations, there were over 70 thousand wiki services in the Internet, which potentially can be used to enrich Wikipedia and other knowledge bases. Additionally, there are over 1300 linked databases [84], which use data from open sources. We can also take into account the web-services that allow companies and individuals to share their databases for research, such as Kaggle [85].

Author Contributions: K.W. and W.L. conceived the research problem; W.L. conducted state of the art analysis; K.W. proposed research methodology and designed the experiments, starting from hypotheses to be verified statistically; W.L. collected data and performed the analysis; W.L. and K.W. interpreted the results; W.A. provided an overall guidance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Price, R.; Shanks, G. A semiotic information quality framework: development and comparative analysis. In *Enacting Research Methods in Information Systems*; Springer, 2016; pp. 219–250.

2. Xu, H.; Koronios, A. Understanding information quality in e-business. *Journal of Computer Information Systems* **2005**, *45*, 73–82.

3. Wikipedia Meta-Wiki. List of Wikipedias. https://meta.wikimedia.org/wiki/List_of_Wikipedias, accessed on 2019-05-05.

4. Alexa. wikipedia.org Traffic Statistics. https://www.alexa.com/siteinfo/wikipedia.org, accessed on 2018-10-08.

5. Thompson, N.; Hanley, D. Science is shaped by wikipedia: Evidence from a randomized control trial. *Social Science Research Network* **2018**.

6. Osman, K. The role of conflict in determining consensus on quality in Wikipedia articles. Proceedings of the 9th International Symposium on Open Collaboration. ACM, 2013, p. 12.

7. Callahan, E.S.; Herring, S.C. Cultural bias in Wikipedia content on famous persons. *Journal of the American society for information science and technology* **2011**, *62*, 1899–1915.

8. Laufer, P.; Wagner, C.; Flöck, F.; Strohmaier, M. Mining cross-cultural relations from Wikipedia: a study of 31 European food cultures. Proceedings of the ACM Web Science Conference. ACM, 2015, p. 3.

9. Gieck, R.; Kinnunen, H.M.; Li, Y.; Moghaddam, M.; Pradel, F.; Gloor, P.A.; Paasivaara, M.; Zylka, M.P. Cultural differences in the understanding of history on Wikipedia. In *Designing Networks for Innovation and Improvisation*; Springer, 2016; pp. 3–12.

461   10.   Samoilenko, A.; Karimi, F.; Edler, D.; Kunegis, J.; Strohmaier, M. Linguistic neighbourhoods: explaining
462   cultural borders on Wikipedia through multilingual co-editing activity. *EPJ data science* **2016**, *5*, 9.
463   11.   Kim, S.; Park, S.; Hale, S.A.; Kim, S.; Byun, J.; Oh, A.H. Understanding editing behaviors in multilingual
464   Wikipedia. *PloS one* **2016**, *11*, e0155305.
465   12.   Węcel, K.; Lewoniewski, W. Modelling the Quality of Attributes in Wikipedia Infoboxes. In *Business*
466   *Information Systems Workshops*; Abramowicz, W., Ed.; Springer International Publishing, 2015; Vol. 228,
467   *Lecture Notes in Business Information Processing*, pp. 308–320. doi:10.1007/978-3-319-26762-3_27.
468   13.   Lewoniewski, W. The method of comparing and enriching information in multlingual wikis based on the
469   analysis of their quality. Phd, Poznań University of Economics and Business, 2018.
470   14.   Xu, Y.; Luo, T. Measuring article quality in Wikipedia: Lexical clue model. Web Society (SWS), 2011 3rd
471   Symposium on. IEEE, 2011, pp. 141–146. doi:10.1109/SWS.2011.6101286.
472   15.   Anderka, M.; Stein, B.; Lipka, N. Predicting quality flaws in user-generated content: the case of wikipedia.
473   Proceedings of the 35th international ACM SIGIR conference on Research and development in information
474   retrieval. ACM, 2012, pp. 981–990.
475   16.   Warncke-wang, M.; Cosley, D.; Riedl, J. Tell Me More : An Actionable Quality Model for Wikipedia.
476   WikiSym 2013, 2013, pp. 1–10. doi:10.1145/2491055.2491063.
477   17.   Su, Q.; Liu, P. A Psycho-Lexical Approach to the Assessment of Information Quality on Wikipedia. 2015
478   IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT),
479   2015, Vol. 3, pp. 184–187. doi:10.1109/WI-IAT.2015.23.
480   18.   Lewoniewski, W.; Węcel, K.; Abramowicz, W. Quality and Importance of Wikipedia Articles in Different
481   Languages. In *Information and Software Technologies: 22nd International Conference, ICIST 2016, Druskininkai,*
482   *Lithuania, October 13-15, 2016, Proceedings*; Springer International Publishing: Cham, 2016; pp. 613–624.
483   doi:10.1007/978-3-319-46254-7_50.
484   19.   Dang, Q.V.; Ignat, C.L. Quality assessment of Wikipedia articles without feature engineering. 2016
485   IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2016, pp. 27–30.
486   20.   Halfaker, A.; Taraborelli, D. Artificial intelligence service 'ORES' gives Wikipedians X-ray specs to see
487   through bad edits. https://blog.wikimedia.org/2015/11/30/artificial-intelligence-x-ray-specs/, accessed
488   on 2017-12-31.
489   21.   Wikimedia Foundation. ORES. https://ores.wikimedia.org/, accessed on 2019-05-05.
490   22.   Lewoniewski, W.; Węcel, K.; Abramowicz, W. Relative Quality and Popularity Evaluation of Multilingual
491   Wikipedia Articles. *Informatics* **2017**, *4*. doi:10.3390/informatics4040043.
492   23.   Lewoniewski, W.; Härting, R.C.; Węcel, K.; Reichstein, C.; Abramowicz, W. Application of SEO Metrics to
493   Determine the Quality of Wikipedia Articles and Their Sources. Information and Software Technologies;
494   Damaševičius, R.; Vasiljevienė, G., Eds.; Springer International Publishing: Cham, 2018; pp. 139–152.
495   doi:10.1007/978-3-319-99972-2_11.
496   24.   Wikimedia Meta-Wiki. Wikipedia article depth. https://meta.wikimedia.org/wiki/Wikipedia_article_
497   depth, accessed on 2019-04-26.
498   25.   Bao, P.; Hecht, B.; Carton, S.; Quaderi, M.; Horn, M.; Gergle, D. Omnipedia: bridging the wikipedia
499   language gap. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM,
500   2012, pp. 1075–1084.
501   26.   Kahn, B.K.; Strong, D.M.; Wang, R.Y. Information quality benchmarks: product and service performance.
502   *Communications of the ACM* **2002**, *45*, 184–192.
503   27.   Tayi, G.K.; Ballou, D.P. Examining data quality. *Communications of the ACM* **1998**, *41*, 54–57.
504   28.   Giles, J. Internet encyclopaedias go head to head, 2005.
505   29.   Holman Rector, L. Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in
506   historical articles. *Reference services review* **2008**, *36*, 7–22.
507   30.   Crawford, H. Encyclopedias. *Reference and information services: An introduction* **2001**, pp. 433–459.
508   31.   Lewoniewski, W. Measures for Quality Assessment of Articles and Infoboxes in Multilingual Wikipedia.
509   International Conference on Business Information Systems; Abramowicz, W.; Paschke, A., Eds. Springer,
510   Springer International Publishing, 2019, pp. 619–633. doi:doi.org/10.1007/978-3-030-04849-5_53.
511   32.   Dalip, D.H.; Gonçalves, M.A.; Cristo, M.; Calado, P. A general multiview framework for assessing the
512   quality of collaboratively created content on web 2.0. *Journal of the Association for Information Science and*
513   *Technology* **2017**, *68*, 286–308.

514   33.   Yaari, E.; Baruchson-Arbib, S.; Bar-Ilan, J. Information quality assessment of community generated content:
515   A user study of Wikipedia. *Journal of Information Science* **2011**, *37*, 487–498.
516   34.   Dang, Q.V.; Ignat, C.L. Measuring Quality of Collaboratively Edited Documents: The Case of Wikipedia.
517   Collaboration and Internet Computing (CIC), 2016 IEEE 2nd International Conference on. IEEE, 2016, pp.
518   266–275.
519   35.   Shen, A.; Qi, J.; Baldwin, T. A Hybrid Model for Quality Assessment of Wikipedia Articles. Proceedings of
520   the Australasian Language Technology Association Workshop 2017, 2017, pp. 43–52.
521   36.   Zhang, S.; Hu, Z.; Zhang, C.; Yu, K. History-Based Article Quality Assessment on Wikipedia. Big Data and
522   Smart Computing (BigComp), 2018 IEEE International Conference on. IEEE, 2018, pp. 1–8.
523   37.   Warncke-Wang, M.; Ranjan, V.; Terveen, L.G.; Hecht, B.J. Misalignment Between Supply and Demand of
524   Quality Content in Peer Production Communities. ICWSM, 2015, pp. 493–502.
525   38.   Lerner, J.; Lomi, A. Knowledge categorization affects popularity and quality of Wikipedia articles. *PloS one*
526   **2018**, *13*, e0190674.
527   39.   Blumenstock, J.E. Automatically Assessing the Quality of Wikipedia Articles. Technical report, 2008.
528   doi:10.1080/17439880802324251.
529   40.   Dalip, D.H.; Gonçalves, M.A.; Cristo, M.; Calado, P. Automatic Assessment of Document Quality
530   in Web Collaborative Digital Libraries.   *Journal of Data and Information Quality* **2011**, *2*, 1–30.
531   doi:10.1145/2063504.2063507.
532   41.   Stvilia, B.; Twidale, M.B.; Smith, L.C.; Gasser, L. Assessing information quality of a community-based
533   encyclopedia. *Proc. ICIQ* **2005**, pp. 442–454.
534   42.   Ransbotham, S.; Kane, G.C. Membership turnover and collaboration success in online communities:
535   Explaining rises and falls from grace in Wikipedia. *Mis Quarterly* **2011**, pp. 613–627.
536   43.   Flekova, L.; Ferschke, O.; Gurevych, I. What makes a good biography?: multidimensional quality analysis
537   based on wikipedia article feedback data. Proceedings of the 23rd international conference on World wide
538   web. ACM, 2014, pp. 855–866.
539   44.   Wu, K.; Zhu, Q.; Zhao, Y.; Zheng, H. Mining the factors affecting the quality of Wikipedia articles.
540   Information Science and Management Engineering (ISME), 2010 International Conference of. IEEE, 2010,
541   Vol. 1, pp. 343–346.
542   45.   Stvilia, B.; Twidale, M.B.; Smith, L.C.; Gasser, L. Assessing information quality of a community-based
543   encyclopedia. *Proc. ICIQ* **2005**, pp. 442–454.
544   46.   Conti, R.; Marzini, E.; Spognardi, A.; Matteucci, I.; Mori, P.; Petrocchi, M. Maturity assessment of Wikipedia
545   medical articles. Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on.
546   IEEE, 2014, pp. 281–286.
547   47.   Wikipedia. Featured article criteria. https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria,
548   accessed on 2019-05-05.
549   48.   Wikipedia. Verifiability. https://en.wikipedia.org/wiki/Wikipedia:Verifiability, accessed on 2019-05-05.
550   49.   Blumenstock, J.E. Size matters: word count as a measure of quality on wikipedia. WWW, 2008, pp.
551   1095–1096. doi:10.1145/1367497.1367673.
552   50.   Dalip, D.H.; Gonçalves, M.A.; Cristo, M.; Calado, P. Automatic quality assessment of content created
553   collaboratively by web communities: a case study of wikipedia. Proceedings of the 9th ACM/IEEE-CS
554   Joint Conference on Digital Libraries, 2009, pp. 295–304. doi:10.1145/1555400.1555449.
555   51.   Ferschke, O.; Gurevych, I.; Rittberger, M. FlawFinder: A Modular System for Predicting Quality Flaws in
556   Wikipedia. CLEF (Online Working Notes/Labs/Workshop), 2012, pp. 1–10.
557   52.   di Sciascio, C.; Strohmaier, D.; Errecalde, M.; Veas, E. WikiLyzer: interactive information quality assessment
558   in Wikipedia. Proceedings of the 22nd International Conference on Intelligent User Interfaces. ACM, 2017,
559   pp. 377–388.
560   53.   Soonthornphisaj, N.; Paengporn, P. Thai Wikipedia Article Quality Filtering Algorithm. Proceedings of the
561   International MultiConference of Engineers and Computer Scientists, 2017, Vol. 1.
562   54.   Liu, J.; Ram, S. Using big data and network analysis to understand Wikipedia article quality. *Data &*
563   *Knowledge Engineering* **2018**.
564   55.   Kane, G.C. A multimethod study of information quality in wiki collaboration. *ACM Transactions on*
565   *Management Information Systems (TMIS)* **2011**, *2*, 4.

566 56. Shang, W. A Comparison of the Historical Entries in Wikipedia and Baidu Baike. International Conference
567   on Information. Springer, 2018, pp. 74–80.

568 57. Roll, U.; Mittermeier, J.C.; Diaz, G.I.; Novosolov, M.; Feldman, A.; Itescu, Y.; Meiri, S.; Grenyer, R. Using
569   Wikipedia page views to explore the cultural importance of global reptiles. *Biological conservation* **2016**,
570   *204*, 42–50.

571 58. Wikimedia Toolforge. Pageviews Analysis. https://tools.wmflabs.org/pageviews/, accessed on
572   2019-05-05.

573 59. WMF Analytics. Wikistats pageview files. https://dumps.wikimedia.org/other/pagecounts-ez/, accessed
574   on 2019-05-05.

575 60. Lih, A. Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative
576   media as a news resource. *5th International Symposium on Online Journalism* **2004**, p. 31.

577 61. Wilkinson, D.M.; Huberman, B.a. Cooperation and quality in wikipedia. *Proceedings of the 2007 international*
578   *symposium on Wikis WikiSym 07* **2007**, pp. 157–164. doi:10.1145/1296951.1296968.

579 62. Kittur, A.; Kraut, R.E. Harnessing the wisdom of crowds in wikipedia. *Proceedings of the ACM 2008*
580   *conference on Computer supported cooperative work - CSCW '08* **2008**, p. 37. doi:10.1145/1460563.1460572.

581 63. Wilkinson, D.M.; Huberman, B.a. Cooperation and quality in wikipedia. *Proceedings of the 2007 international*
582   *symposium on Wikis WikiSym 07* **2007**, pp. 157–164. doi:10.1145/1296951.1296968.

583 64. Kittur, A.; Chi, E.H.; Suh, B. What's in Wikipedia?: mapping topics and conflict using socially annotated
584   category structure. Proceedings of the SIGCHI conference on human factors in computing systems. ACM,
585   2009, pp. 1509–1512.

586 65. Boldi, P.; Monti, C. Cleansing wikipedia categories using centrality. Proceedings of the 25th International
587   Conference Companion on World Wide Web. International World Wide Web Conferences Steering
588   Committee, 2016, pp. 969–974.

589 66. English Wikipedia. Category:Main topic classifications. https://en.wikipedia.org/wiki/Category:
590   Main_topic_classifications, accessed on 2019-04-27.

591 67. Vrandečić, D. Wikidata: A new platform for collaborative data collection. Proceedings of the 21st
592   international conference on world wide web. ACM, 2012, pp. 1063–1064.

593 68. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A nucleus for a web of open
594   data. In *The semantic web*; Springer, 2007; pp. 722–735.

595 69. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey,
596   M.; Van Kleef, P.; Auer, S.; others. DBpedia–a large-scale, multilingual knowledge base extracted from
597   Wikipedia. *Semantic Web* **2015**, *6*, 167–195.

598 70. Abramowicz, W.; Auer, S.; Heath, T. Linked Data in Business. *Business & Information Systems Engineering*
599   **2016**, *58*, 323–326. doi:10.1007/s12599-016-0446-0.

600 71. Lewańska, E. Towards Automatic Business Networks Identification. International Conference on Business
601   Information Systems. Springer, 2017, pp. 389–398. doi:doi.org/10.1007/978-3-319-52464-1_36.

602 72. Filipiak, D.; Filipowska, A. Improving the Quality of Art Market Data Using Linked Open Data and
603   Machine Learning. Business Information Systems Workshops; Abramowicz, W.; Alt, R.; Franczyk, B., Eds.;
604   Springer International Publishing: Cham, 2017; pp. 418–428.

605 73. Stróżyna, M.; Eiden, G.; Abramowicz, W.; Filipiak, D.; Małyszko, J.; Węcel, K. A framework for the
606   quality-based selection and retrieval of open data - a use case from the maritime domain. *Electronic Markets*
607   **2018**, *28*, 219–233. doi:10.1007/s12525-017-0277-y.

608 74. Färber, M.; Bartscherer, F.; Menne, C.; Rettinger, A. Linked data quality of dbpedia, freebase, opencyc,
609   wikidata, and yago. *Semantic Web* **2018**, *9*, 77–129.

610 75. DBpedia. Ontology Classes. http://mappings.dbpedia.org/server/ontology/classes/, accessed on
611   2019-05-05.

612 76. Ringler, D.; Paulheim, H. One knowledge graph to rule them all? Analyzing the differences between
613   DBpedia, YAGO, Wikidata & co. Joint German/Austrian Conference on Artificial Intelligence (Künstliche
614   Intelligenz). Springer, 2017, pp. 366–372.

615 77. Ismayilov, A.; Kontokostas, D.; Auer, S.; Lehmann, J.; Hellmann, S.; others. Wikidata through the Eyes of
616   DBpedia. *Semantic Web* **2018**, *9*, 493–503.

617 78. WikiRank. Quality and popularity assessment of Wikipedia. https://wikirank.net/, accessed on
618   2019-04-27.

79.   WikiRank. Dota 2. https://wikirank.net/en/Dota_2, accessed on 2019-05-05.

80.   WikiRank. Counter-Strike: Global Offensive. https://wikirank.net/en/Counter-Strike:_Global_Offensive, accessed on 2019-05-05.

81.   Deutschsprachige Wikipedia. Kategorie: Kunst und Kultur. https://de.wikipedia.org/wiki/Kategorie: Kunst_und_Kultur, accessed on 2019-05-05.

82.   English Wikipedia. Wikipedia:Protection policy. https://en.wikipedia.org/wiki/Wikipedia:Protection_ policy, accessed on 2019-05-05.

83.   GitHub. Lewoniewski - user profile. https://github.com/lewoniewski, accessed on 2019-05-05.

84.   The Linked Open Data Cloud. Datasets. https://lod-cloud.net/datasets, accessed on 2019-05-05.

85.   Kaggle. Datasets. https://www.kaggle.com/datasets, accessed on 2019-05-05.