Completeness and Reliability of Wikipedia Infoboxes in Various Languages *

Włodzimierz Lewoniewski

Poznań University of Economics and Business, Poland
wlodzimierz.lewoniewski@ue.poznan.pl

Abstract. Despite its popularity, Wikipedia is often criticized for poor information quality. Currently this online knowledge base consist over 45 million articles in almost 300 various languages. Articles in Wikipedia often includes special tables which present shortly important information about persons, places, products, organizations and other subjects. This table is usually placed in a visible part of the article and Wikipedia community called it „infobox". These infoboxes contains information in a structured form that allows automatically enrich popular public databases such as DBpedia. Wikipedia users can edit infoboxes in different languages independently. So, quality of information about the same thing may differ between various language versions. This article will examine the completeness and reliability of infoboxes about different topics in seven language versions of Wikipedia: English, German, French, Polish, Russian, Ukrainian and Belarussian. The results of the study can be used for automatic assessing and improving the quality of information in Wikipedia as well as in other public knowledge bases.

Keywords: Wikipedia, infobox quality, reliability, completeness, DBpedia.

1 Introduction

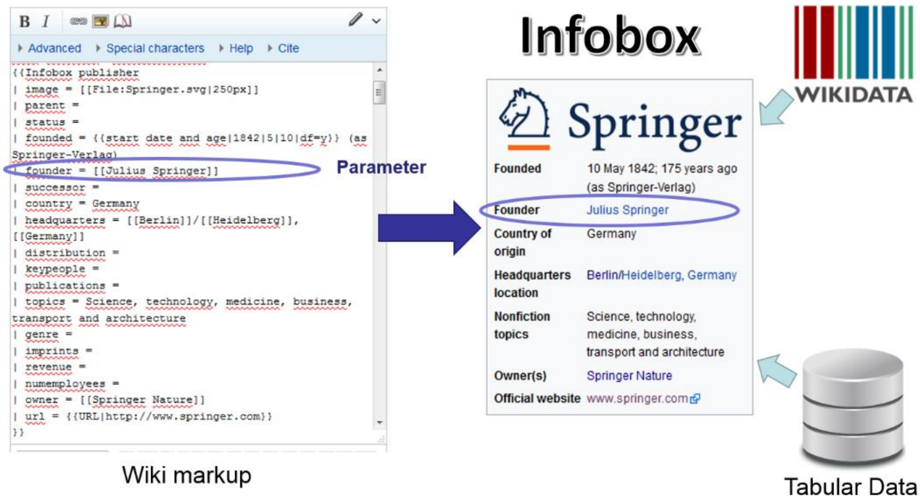
Wikipedia is on 5th place in the ranking of the most popular websites in the world[1]. Nowadays it is one of the most popular sources of knowledge and it allows everyone to participate in the content contribution in over 280 languages[2]. The largest English language version of Wikipedia over 5,4 million articles. Among language versions, which have more than 1 million articles are German, French, Russian and Polish.

Articles related to various topics in different languages can be created and edited even by anonymous users Wikipedia. The contributors of this encyclopedia do not have to formally demonstrate their competences or skills in a specific area. Often changes in articles are immediately available online to wide audience. These and other reasons allows criticizing Wikipedia for poor quality of information. However, some articles in can provide valuable information.

[1] http://www.alexa.com/siteinfo/wikipedia.org
[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias

* This is a preprint version. The final publication is available at Springer via https://doi.org/10.1007/978-3-319-69023-0_25

Wikipedia articles can includes dedicated table with main facts about the subject – so called "infobox". As one of the most important elements of the article, infobox usually placed on a visible part - top right-hand corner of article. That one of the most important elements. Infobox is in fact a Wikipedia template that contains list of items "parameter = value" in a wiki markup. Additionally, values of parameters can also be inserted from Tabular Data[3] or WikiData[4]. Example of such infobox with its sources is shown in figure 1.



```
{{Infobox publisher
| image = [[File:Springer.svg|250px]]
| parent =
| status =
| founded = {{start date and age|1842|5|10|df=y}} (as
Springer-Verlag)
| founder = [[Julius Springer]]
| successor =
| country = Germany
| headquarters = [[Berlin]]/[[Heidelberg]],
[[Germany]]
| distribution =
| keypeople =
| publications =
| topics = Science, technology, medicine, business,
transport and architecture
| genre =
| imprints =
| revenue =
| numemployees =
| owner = [[Springer Nature]]
| url = {{URL|http://www.springer.com}}
}}
```

Founded	10 May 1842; 175 years ago (as Springer-Verlag)
Founder	Julius Springer
Country of origin	Germany
Headquarters location	Berlin/Heidelberg, Germany
Nonfiction topics	Science, technology, medicine, business, transport and architecture
Owner(s)	Springer Nature
Official website	www.springer.com

Figure 1. Infobox with its data sources in English Wikipedia about publisher in article "Springer Science+Business Media".

Depending on the topic infoboxes have a different name, appearance and a strictly defined set of parameters. The structure of the infobox allows others public knowledge bases to extract the data. Among such projects, one of the most popular is DBpedia[5]. This crowd-sourced community effort uses a special framework[6] to extract information from these infoboxes and makes its available on the Web.

Infoboxes describing the same topics exist in different Wikipedia languages. Each language version of infobox can have its own set of parameters and parameters describing the same facts can be written differently. Therefore, DBpedia Extraction Framework uses the mappings defined by the community[7] to homogenize information extracted from Wikipedia in various languages. This makes possible to compare the filled parameters having different spellings. Later we can use this to determine which parameters are missing in certain language versions and automatically transfer them.

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

[4] https://www.wikidata.org

[5] http://wiki.dbpedia.org/

[6] https://github.com/dbpedia/extraction-framework

[7] http://mappings.dbpedia.org

One of the features of Wikipedia is that information about the same subject created in different languages often independently of each other. Therefore, an important issue is to identify the language version (or versions) with infoboxes that contains more complete and reliable data. In this paper presents an analysis of these two quality dimensions of infoboxes describing different topics in seven language versions of Wikipedia: English (EN), German (GE), French (FR), Russian (RU), Polish (PL), Ukrainian (UK), Belarussian (BE).

2 Related work

In Wikipedia, there is a system for assessing the quality of articles by community in particular language versions. However, a large number of articles have not yet been evaluated [1]. Therefore, automatic assessment of Wikipedia articles is a well-known and developed topic in scientific works. Such articles features as text length, number of references, images, sections can help in assessing Wikipedia articles [1,2,3]. Additionally for this purpose it is possible to analyze authors' reputation and articles edit history [4,5], in some cases natural language processing techniques can be useful [6]. At the same time, each language version of Wikipedia can have its own quality model [1,2]. Some of the proposed features used in online service WikiRank[8] to compare quality and popularity of articles in different languages. Those researches mainly focused on the analysis of the quality of articles as whole, not its individual elements such are infoboxes.

Preliminary experiments have shown that the articles evaluated with the highest grade by Wikipedia community in one language do not always contains infoboxes with the highest quality in comparison to other language versions, where articles received lower grades. Therefore, it is also necessary to be able to evaluate the quality of data in infoboxes taken into the account other measures.

As mentioned earlier, DBpedia extracts the information of the Wikipedia infoboxes. There are different approaches and tools to assess quality in this semantic database. RDFUnit uses pre-defined quality test patterns based on a SPARQL query template to analyze integrity constraints of dataset [7]. In contrast to the RDFUnit, ontology driven framework Luzzu allows implementation of different metrics without SPARQL querying [8]. Since DBpedia is the representative of Linked Open Data (LOD), in the area of fusion of data from different languages, Sieve framework can be used [9]. There are also algorithms that can identify missing type statements, and identifies faulty statements in LOD [10]. However, these approaches require domain experts to identify quality assessment metrics in a schema layer and for a more in-depth analysis of quality of the LOD it is necessary to take into account additional quality dimensions [11].

This paper proposed method to evaluate completeness and reliability of infoboxes that’s describes companies, universities, films, albums and video games in seven language versions.

[8] http://wikirank.net

3. Dataset

The results presented in the paper were carried out on Wikipedia dumps on May, 2017. First, articles with infoboxes on each topic in different languages were found. The number of such articles shown in the table 1.

Topic	BE	DE	EN	FR	PL	RU	UK
Album	130	8 348	137 972	36 379	22 026	14 144	6 522
Companies	371	21 703	56 678	15 427	4 660	9 449	3 628
Films	212	32 327	114 727	35 063	18 654	25 615	12 879
Universities	244	3 406	20 421	4 109	2 175	2 320	1 082
Video games	51	2 839	20 685	11 096	2 924	5 492	1 341

Table 1. Number of articles with infoboxes in particular topic in different Wikipedia languages. Source: own calculation in May, 2017.

It is easy to see that the largest English Wikipedia has the largest number of articles in each considered topic. The analysis of articles in this dataset also showed that only a small part of the articles are presented in all language versions Figure 2 presents the coverage of articles in three topics in some language versions of Wikipedia.

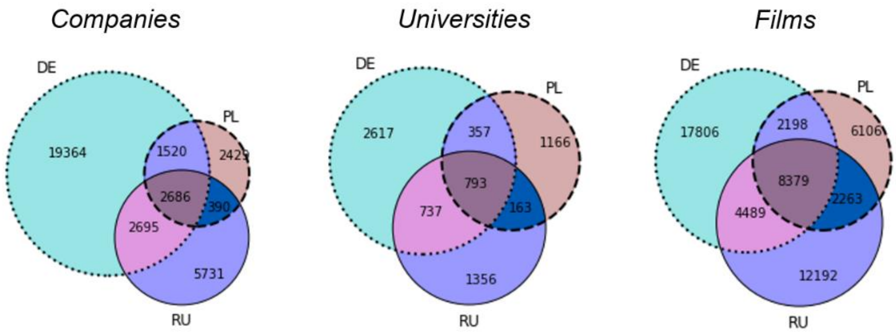
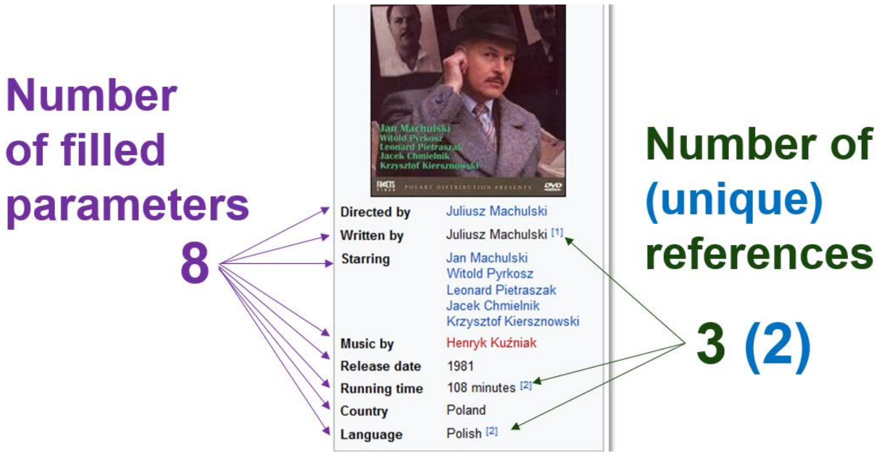


Figure 2. Coverage of Wikipedia articles about companies, universities and films in German (DE), Polish (PL) and Russian (RU) language. Source: own calculations in May, 2017.

To calculate quality metrics presented in this work, parameters of infoboxes were extracted using own parser. The article will use such concepts as number of filled parameters, number of references, number of unique references. For a more visual understanding, figure 3 shows an example of an infobox with this metrics.



Directed by Juliusz Machulski
Written by Juliusz Machulski [1]
Starring Jan Machulski
Witold Pyrkosz
Leonard Pietraszak
Jacek Chmielnik
Krzysztof Kierasznowski
Music by Henryk Kuźniak
Release date 1981
Running time 108 minutes [2]
Country Poland
Language Polish [2]

Figure 3. Example of the infobox about film and some of its metrics.

4 Completeness of infoboxes

Often users of Wikipedia do not fill in all the parameters of infoboxes. When calculating all presented in this paper metrics, the infoboxes parameters that Wikipedia users entered by mistake were ignored. Such incorrect parameters can easily be identified - as mentioned earlier, each infobox contains a certain set of predefined parameters names.

There are parameters that are filled more often than others. Figure 4 shows the top 20 most frequently filled parameters in the two selected infoboxes in English Wikipedia.

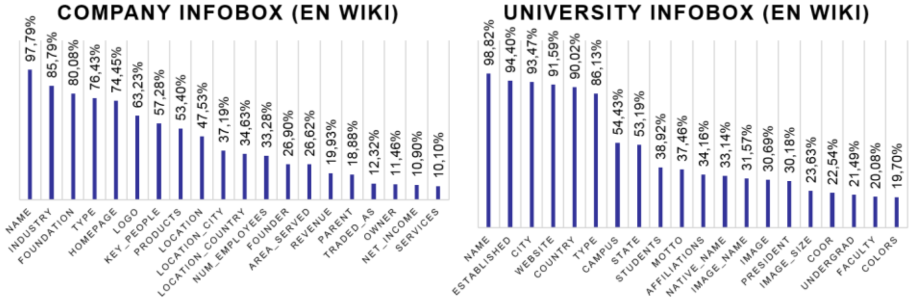


Figure 4. The top 20 most frequently filled parameters in the company and university infoboxes in English Wikipedia. Source: own calculations.

It should be noted that the frequency of filling the same parameters in different language versions is vary. Table 2 presents filling frequency of some parameters if infoboxes that describes companies in particular Wikipedia language edition. It is also important that in some language versions certain infoboxes do not use parameters that are used by other Wikipedia languages.

Parameter	BE	DE	EN	FR	PL	RU	UK
Name	75,76%	91,57%	97,79%	88,17%	99,87%	93,23%	78,00%
Industry	61,74%	91,22%	85,79%	78,81%	82,14%	79,04%	67,47%
Foundation	70,45%	93,14%	80,08%	87,46%	89,27%	87,29%	52,04%
Type	64,02%	85,01%	76,43%	47,79%	65,95%	70,97%	53,99%
Homepage	70,08%	74,86%	74,45%	79,96%	81,34%	75,70%	62,28%

Table 2. Filling frequency of some parameters of company infoboxes in different languages of Wikipedia. Source: own calculation in May, 2017.

In this work completeness of infoboxes measured by two metrics. The completeness C_1 of infobox calculated as the ratio of the number of parameter values to the number of all defined parameters in the infobox of a given type:

$$C_1 = \frac{FP}{AP},$$

where *FP* – number of filled parameters, *AP* – number of all defined parameters in considered infobox.

Topic	BE	DE	EN	FR	PL	RU	UK
Album	.339	.423	.533	.336	.432	.611	.415
Companies	.162	.589	.2	.295	.288	.221	.105
Films	.647	.399	.479	.509	.525	.617	.313
Universities	.284	.46	.186	.304	.347	.329	.213
Video games	.14	.418	.381	.434	.371	.153	.149

Table 3. Average completeness C_1 of Wikipedia infoboxes describing different topics in various languages. Source: own calculation in May, 2017.

Second method of measuring completeness based on the previous one with considering weights for each filled parameter:

$$C_2 = \frac{\sum_{i=1}^{FP} WP_i}{AP},$$

where *FP* – number of filled parameters, WP_i – weight of the parameter $P_{i,}$ *AP* – number of all defined parameters in considered infobox.

Weight is based on the frequency of filling this parameter. For example for university infobox in English Wikipedia weight of parameter “city” is 0,9347 (see figure 3).

Topic	BE	DE	EN	FR	PL	RU	UK
Album	.217	.343	.452	.223	.341	.474	.329
Companies	.071	.446	.107	.167	.153	.131	.04
Films	.527	.268	.36	.414	.402	.518	.199
Universities	.158	.335	.099	.19	.204	.19	.114
Video games	.044	.296	.287	.341	.266	.043	.048

Table 4. Average completeness C_2 of Wikipedia infoboxes describing different topics in various languages. Source: own calculation in May, 2017.

Figure 5 shows distribution of completeness C_1 and C_2 of infoboxes that describes different topics in particular language versions of Wikipedia. In presented boxplots the central box represents the middle 50% of the considered infoboxes in particular Wikipedia language, the central bar is the median and the bars at the end of the dotted lines (circles) close the most of the observations. Circles that lie beyond the end of the whiskers are data points that may be outliers.

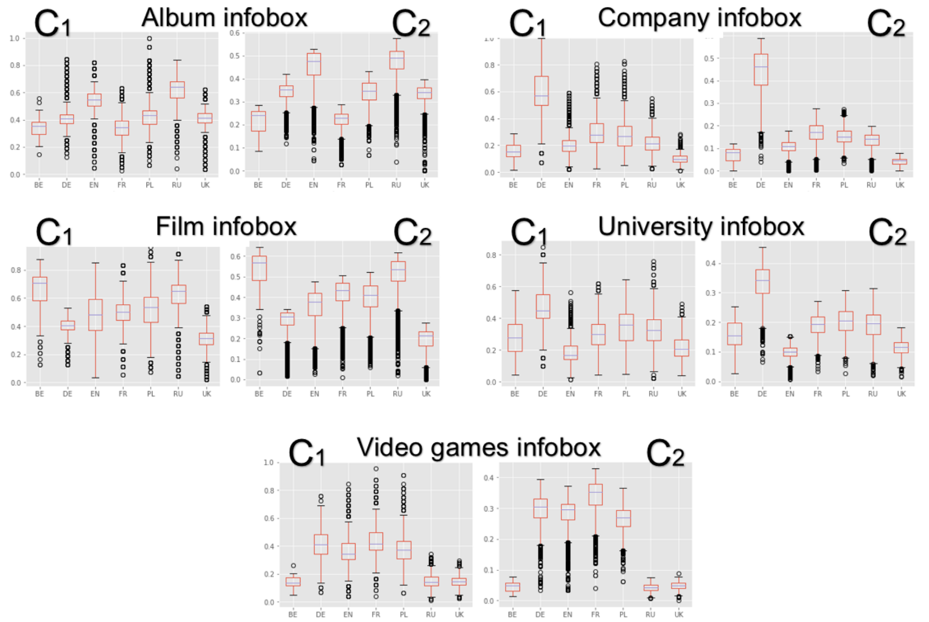


Figure 5. Distribution of completeness C_1 and C_2 of infoboxes that describes different topics in particular language version of Wikipedia.
Source: own calculations using pandas library[9].

Differences between completeness of the same infoboxes in various languages occurs due different sets of predefined parameters. For example in German Wikipedia infobox about company can have only 14 parameters while such infobox in Ukrainian edition can have over 40.

5 Reliability of infoboxes

One of the convenient ways to verify the reliability of information in Wikipedia is to check the sources (if they exist). So, to measure reliability of infoboxes the following

[9] http://pandas.pydata.org

metrics are used in this paper: number of references (R_1), number of unique references (R_2), references to filled parameters ratio R_3 calculated by the formula:

$$R_3 = \frac{R_1}{FP},$$

where R_1 – number of references in the infobox, FP – number of filled parameters.

Table 5 presents the results of average number of references R_1 in Wikipedia infoboxes describing different topics in various languages.

Topic	BE	DE	EN	FR	PL	RU	UK
Album	.22	.1	.153	.195	1.002	.641	.187
Companies	.386	.803	.649	.457	.352	.56	.459
Films	.553	.059	.441	.01	.403	.177	.316
Universities	.236	.774	.762	.47	.363	.329	.337
Video games	1.8	.467	.807	.278	1.944	.874	.641

Table 5. Average number of references R_1 of Wikipedia infoboxes describing different topics in various languages. Source: own calculation in May, 2017.

Depending on topic and language versions of Wikipedia number of references are vary. In particular topics some of the language version practically do not use references in infoboxes. For example in French Wikipedia only 277 of 35013 infoboxes that describes films have at least 1 reference. As a result, average number of references in these infoboxes in French at least 5 times less than in other considered Wikipedia language editions. Another interesting example is Belarussian and Polish Wikipedia with infoboxes that describes video games. Judging by the average value of R_1 almost all of those infoboxes must have at least 2 references. However, relatively high average R_1 associated with some part of infoboxes that have a large number of references. In Polish version about 10% of infoboxes about video games have over 6 references. There is even an infobox in that language version of Wikipedia with almost 40 references[10]. In the Belarusian Wikipedia 3 of 50 video game infoboxes have over 10 references.

Now let's look at the results of the analysis of unique references in the same dataset. Table 6 presents average number of unique references R_2.

Topic	BE	DE	EN	FR	PL	RU	UK
Album	.22	.086	.119	.169	.952	.567	.176
Companies	.273	.52	.335	.258	.247	.304	.248
Films	.398	.052	.329	.008	.131	.135	.284
Universities	.194	.554	.526	.33	.29	.23	.255
Video games	1.66	.367	.54	.2	.876	.656	.526

Table 6. Average number of unique references R_2 of Wikipedia infoboxes describing different topics in various languages. Source: own calculation in May, 2017.

[10] https://pl.wikipedia.org/wiki/StarCraft_II:_Wings_of_Liberty

Comparing with the results of the calculation of average R_1, table 5 shows lower values. This difference is due to the fact, that sometimes two or more parameters of particular infobox can have common source as a reference. Difference between Table 4 and 5 also shows how often do Wikipedia community use common source to describe different parameters of particular infobox in each language. For example in Polish Wikipedia infoboxes about video games in average one source can occur as 2 references in particular infobox. However, there are also such cases, where every or almost all references within particular infobox are unique. This concerns album and university infobox in Belarusian language, film infobox in French, German, and Russian Wikipedia.

In the previous section, the results of the measurement of completeness were considered. Table 7 shows how the infobox parameters supported by references through counting average references to filled parameters ratio R_3.

Topic	BE	DE	EN	FR	PL	RU	UK
Album	.039	.009	.015	.023	.098	.054	.019
Companies	.076	.115	.106	.065	.051	.121	.249
Films	.04	.013	.041	.001	.035	.014	.034
Universities	.03	.11	.095	.048	.04	.036	.054
Video games	.402	.052	.103	.032	.214	.218	.159

Table 7. Average references to filled parameters ratio R_3 of Wikipedia infoboxes describing different topics in various languages. Source: own calculation in May, 2017.

The results show, that relatively more often Wikipedia users inserts references to parameters of infoboxes about video games (especially in Belarussian, Polish and Russian) and companies (especially in Ukrainian).

6 Conclusions and Future Work

In this paper were introduced quality metrics of infoboxes related to completeness and reliability. Result of the analysis shows, that depending on the Wikipedia language version and described topic there different completeness of infoboxes. According to research can be observed a different culture of filling the parameters in infoboxes – in specified languages there are parameters usually filled by users more often than their counterparts in other language versions of Wikipedia. Additionally, some infoboxes in certain languages may not use parameters that are commonly used in other language versions. Therefore, some facts presented in infoboxes describing certain topics may be particularly important (or not important) for separate Wikipedia language community.

The methods proposed in the article for evaluating completeness and reliability can be used in other models to determine infobox with the best quality of data across language versions of Wikipedia. Using these quality models together with techniques of parameters unification it is possible to improve the quality of data in multilingual Wikipedia and other knowledge bases. Figure 6 presents example of extracting the parameters from company infobox in different languages and unification to common property names in DBpedia.

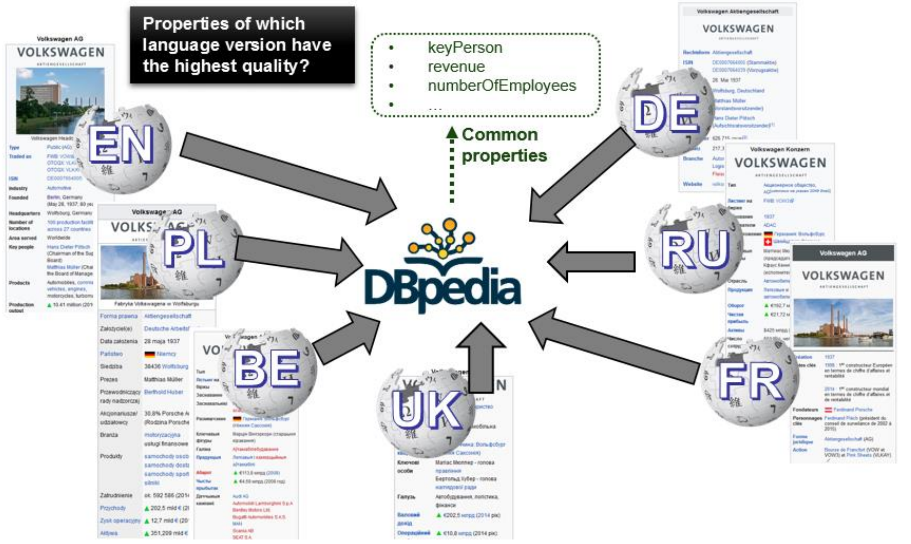


Figure 6. Extracting the parameters from company infobox in different languages of Wikipedia and unification to common property names in DBpedia.

Each Wikipedia article in certain language version without an infobox can be enriched potentially from other language versions. Presented in paper metrics with other quality models can help determine such language version (versions). Despite the fact that Wikipedia is the largest, it can be enriched by other language versions. Table 8 presents potential number of articles in each language and each topic that can be created or enriched using infoboxes from other language version of Wikipedia.

Topic	BE	DE	EN	FR	PL	RU	UK
Album	170 793	157 925	22 538	130 712	143 118	151 548	162 086
Companies	83 829	58 169	22 783	65 154	77 409	72 932	79 470
Films	146 876	114 263	28 355	100 265	128 189	119 812	133 739
Universities	24 325	20 273	3 420	19 804	22 298	21 728	23 072
Video games	24 325	21 184	2 924	12 953	21 245	18 559	22 917

Table 8. Number of Wikipedia articles that can be created or enriched using infoboxes from other languages. Source: own calculation in May, 2017.

Future works will continue researches in the field of quality measurement of Wikipedia infoboxes. Through research new metrics will be developed. For example, for research on reliability of infoboxes in various languages of Wikipedia it is planned to take into account similarities of the references [12].

References

1. Lewoniewski,W., Węcel, K., Abramowicz,W., (2016), Quality and importance of Wikipedia articles in different languages. In: Information and Software Technologies: 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016, Proceedings. Springer International Publishing, Cham (2016) 613–624.
2. Warncke-Wang, M., Cosley, D., Riedl, J., (2013), Tell me more: an actionable quality model for Wikipedia. In Proceedings of the 9th International Symposium on Open Collaboration (p. 8). ACM.
3. Węcel, K., Lewoniewski, W., (2015), Modelling the Quality of Attributes in Wikipedia Infoboxes. In Business Information Systems Workshops. Volume 228 of Lecture Notes in Business Information Processing. Springer International Publishing, pp.308–320
4. Suzuki, Y., Nakamura, S., (2016), Assessing the Quality of Wikipedia Editors through Crowdsourcing. In Proceedings of the 25th International Conference Companion on World Wide Web (pp. 1001-1006). International World Wide Web Conferences Steering Committee.
5. Ingawale, M., Dutta, A., Roy, R., Seetharaman, P., (2013), Network analysis of user generated content quality in Wikipedia. Online Information Review, 37(4), 602-619.
6. Khairova N., Lewoniewski W., Węcel K., (2017), Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction. In: Abramowicz W. (eds) Business Information Systems. BIS 2017. Lecture Notes in Business Information Processing, vol 288. Springer, Cham
7. Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A., (2014), Test-driven evaluation of linked data quality. In Proceedings of the 23rd international conference on World Wide Web (pp. 747-758). ACM.
8. Debattista, J., Auer, S., Lange, C., (2016), Luzzu - A Framework for Linked Data Quality Assessment. In Semantic Computing (ICSC), 2016 IEEE Tenth International Conference on (pp. 124-131). IEEE.
9. Mendes, P.N., Mühleisen, H., Bizer, C., (2012), Sieve: Linked data quality assessment and fusion. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops. EDBT-ICDT '12, New York, NY, USA, ACM, p. 116–123
10. Paulheim, H., Bizer, C., (2014), Improving the quality of linked data using statistical distributions. International Journal on Semantic Web and Information Systems (IJSWIS), 10(2), 63-86.
11. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S., (2016), Quality assessment for linked data: A survey. Semantic Web 7(1), p. 63–93
12. Lewoniewski, W., Węcel, K., Abramowicz W., (2017), Analysis of References across Wikipedia Languages, The 23rd International Conference on Information and Software Technologies (in press).