Measures for Quality Assessment of Articles and Infoboxes in Multilingual Wikipedia

Włodzimierz Lewoniewski[1][0000−0002−0163−5492]

Department of Information Systems, Poznań University of Economics and Business,
al. Niepodległości 10, 61-875 Poznań, Poland
{wlodzimierz.lewoniewski}@ue.poznan.pl
http://kie.ue.poznan.pl

Abstract. * One of the most popular collaborative knowledge bases on the Internet is Wikipedia. Articles of this free encyclopaedia are created and edited by users from different countries in about 300 languages. Depending on topic and language version, quality of information there may vary. This study presents and classifies measures that can be extracted from Wikipedia articles for the purpose of automatic quality assessment in different languages. Based on a state of the art analysis and own experiments, specific measures for various aspects of quality have been defined. Additional, in this work they were also defined measures for quality assessment of data contained in the structural parts of Wikipedia articles - infoboxes. This study describes also an extraction methods for various sources of measures, that can be used in quality assessment.

Keywords: Wikipedia · Data Quality · Quality Measures · DBpedia · Wikidata · Quality Dimensions · Web 2.0 · Encyclopedia

1 Introduction

Nowadays, often decision making in different areas depends on information that is found in the various open sources. On the one hand, peoples care about having access to as wide range of related data as possible. On the other hand, the quality of the data is also important. Therefore, searching for relevant information, Internet users need to understand how choose data and information with high quality from the Web.

Technologies Web 2.0 for more than 10 years allow everyone to contribute to common human knowledge on the Internet. One of the best examples of such online repositories is Wikipedia with over 48 million articles [76]. Information in this free encyclopedia can be edited even by anonymous users independently in about 300 various language versions. The most developed is English version with over 5.7 million articles. However, this does not mean that this language version contains data and information of the best quality. Despite its popularity

* The final publication is available at https://link.springer.com/chapter/10.1007/978-3-030-04849-5_53

(the 5th most visited website in the world [2]) Wikipedia often criticized for the poor quality of content [21]. That quality depends on topic and language version of the articles [48].

Community of Wikipedia users separately in each language version defined rules and criteria to be followed by contributor when creating and editing the content of the articles. When all (or almost all) criteria are met, the article can get special award for quality. For example, in English Wikipedia the best articles have name „Featured" [23] (when all criteria are met) and „Good" [24] (when almost all criteria are met). In other language versions can be found equivalents for these awards with different spelling. However, very small number of articles in each language version of Wikipedia can boast such high quality content - they have a share of less than 1 percent [48].

In some language versions of Wikipedia articles can get other (lower) grades for quality. Articles assessment requires initiative and time from users, which should check whether the content meets the accepted quality criteria. Additionally, the content of the previously evaluated article can be corrected and updated at any time several times, which does not mean that the quality grade will also be corrected. Therefore, a large number of articles in different language versions do not have an assessment or have an irrelevant grade.

Quality in Wikipedia is broad topic in scientific works [77] and there are different researches in the field of automatic predicting of quality grade of the Wikipedia articles. Each study usually used own set of measures and specific algorithm to build a model to solve this task. This work presents known and new measures which can be related for different quality dimensions of the Wikipedia articles.

Articles in Wikipedia often includes dedicated table with main facts about the subject infobox. Depending on topic, the presence of an infobox can affect the quality of whole article. Infobox usually placed on a visible part of the page. That one of the most important elements. In wiki markup infobox contains list of items „parameter = value". However, sometimes it data can be inserted automatically from other sources: from Tabular Data [26] or WikiData [74]. Example of such infobox with its data sources is shown in figure 1.

These infoboxes are also used to enrich other others public knowledge bases such as DBpedia [18]. Data from such bases have been successfully applied in a number of domains: Life Sciences, Web Search, Digital Libraries, Maritime Domain, Art Market and others [1,66,29]. So this article presents also dimensions and measures for quality assessment of the infoboxes.

2 Quality Dimensions of the Wikipedia Articles

Quality can be defined as a degree to which information has content, form, and time characteristics, which give it value to specific end users [57]. If we take into the account user needs, quality will be the degree to which information is meeting this needs according to external, subjective user perceptions.[69]. In other words, quality of information is fitness for use [36].

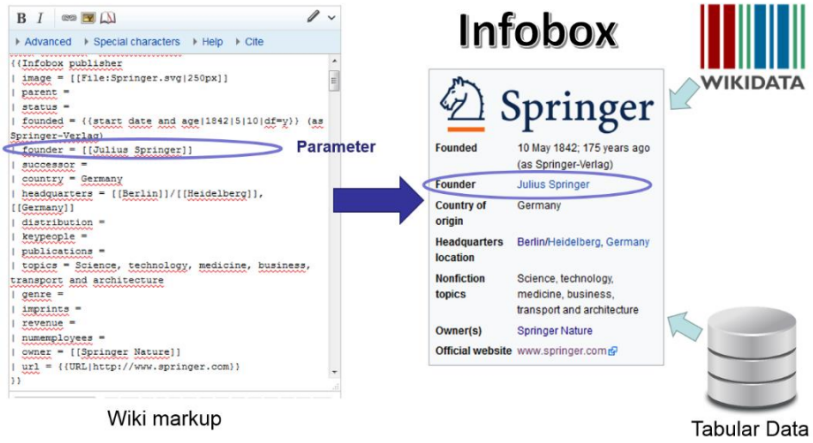

Fig. 1. Infobox with its data sources in English Wikipedia about publisher in article "Springer Science+Business Media"

According to ISO 8402, quality is „the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs" [58]. For the needs of this work, the concept of quality from ISO 8402 will be used.

There are different approaches that defined measures and dimensions of information quality in the literature. For example, Eppler proposed 70 characteristics (or dimensions) of information that narrows down to 16 most important [27]. Depending on a source of information, complementary definitions of quality, outlining the various important dimensions of quality (e.g. accuracy, timeliness, etc) can be defined.

Due the fact that on the one hand Wikipedia is an encyclopedia, and on the other hand - a representative of Web 2.0 services, based on the literature below are presented the most important quality dimensions for three sources of information:

- **Traditional encyclopedias**: Authority, Completeness, Format, Objectivity, Style, Timeliness, Uniqueness
- **Web 2.0 services**: Accessibility, Completeness, Credibility, Involvement, Objectivity, Readability, Relevance, Reputation, Style, Timeliness, Uniqueness, Usefulness
- **Wikipedia**: Completeness, Credibility, Objectivity, Readability, Relevance, Style, Timeliness

Figure 2 shows coverage of the quality dimensions of three mention sources of information.

Short description of each quality dimension are presented below:

- **Credibility**: whether the information provided can be checked with reliable sources

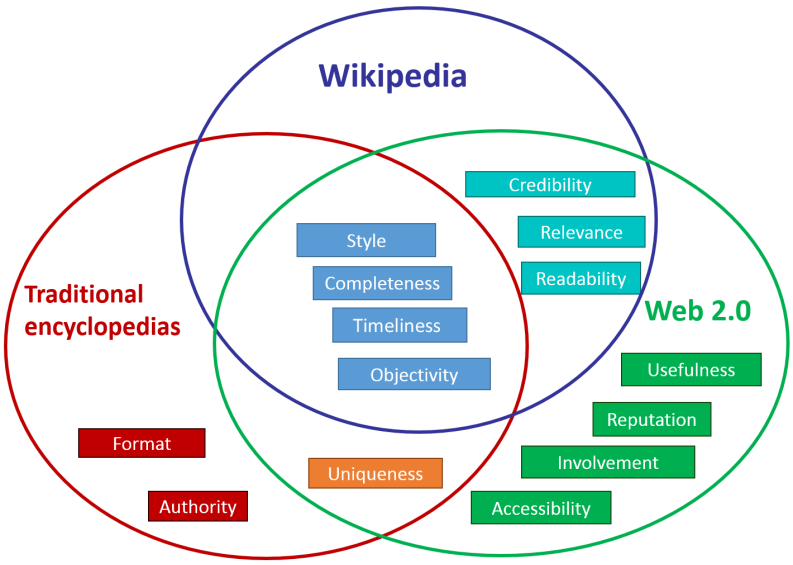

Fig. 2. Coverage of the quality dimensions of three sources of information: Traditional encyclopedias, Wikipedia, Web 2.0 services. Source: own study.

- **Completeness**: how comprehensive the description of the topic is in article
- **Objectivity**: to what extent the content of the article meets the criterion of a neutral point of view, does it contain pictures and other multimedia materials related to this article
- **Readability**: to what extent the text is understandable and free from unnecessary complexity
- **Relevance**: to what extent the article is relevant (important) for readers/users
- **Style**: How the content of the article is organized.
- **Timeliness**: to what extent the article describes the current state of a certain reality (degree to which information is up-to-date).

3 Quality Measures of the Wikipedia articles

Each of 7 quality dimension of the Wikipedia cave has own set of measures. Each measure can be represented as statistical value. In section describes the most popular quality measures of the Wikipedia Articles related to particular quality dimension.

3.1 Credibility

Using reliable sources in Wikipedia is one of the important criteria for writing articles with high quality [22]. Readers of the encyclopedia must be able

to check where the information come from [25]. Therefore, one of the most commonly used measure related to credibility is number of the references in Wikipedia articles [6,15,17,28,81,13,61,65,64,84,72,47,46,48,43] or external link count [6,15,68,82,30,28,13,65,72,47]. On of the related research has shown that depending on the references users can assess the trustworthiness of Wikipedia articles[52].

Here it can be also taken info the account not only quantity, but also quality of the sources. On of the possibilities is to estimate popularity of the reference and its domain based on visiting count or number of incoming links from other websites. For this, data from search engines can be useful. For this we can use data from search engines such as Google, Baidu, Yahoo, Bing, Yandex and also specific tools such as Alexa. Another possibility is to evaluate scientific references using Altmetric [3] and other tools.

3.2 Completeness

Wikipedia articles with high quality must neglects no major facts or details and places the subject in context. One of the most popular measure for this dimension is content volume measured by articles length [82,17,64,84,72,47,46,48,43,71,41,5,16,68,59,30,81,67,13]. Length can be measured in different ways: bytes, characters, words and others.

3.3 Objectivity

Wikipedia article must presents views fairly and without bias. Objectivity can expected from article, which was jointly created by a large number of different users. So, the most popular measure is the number of unique authors [50,68,80,39,51,79,37,82,30,28,81,67,13,61,72,47]. Here it can be also used image count measure [6,15,68,82,81,67,13,84,51,37,63,72,47,46,48,43]

3.4 Readability

Measures related to this quality dimension must show to what extent the text is understandable and free from unnecessary complexity. Therefore, first of all, here it is necessary to take into account special readability formulas such as Automated Readability Index [62,5,16,51,60,59,30,17,64], Bormuth Index [7,4], Coleman-Liau Index [12,5,16,59,30,17,64], FORCAST Readability [10,5], Flesch Reading Score [31,5,16,68,30,17,81,67,13,61,64], Flesch-Kincaid grade level [38,5,16,68,70,30,17,81,67,13,61,64], Gunning Fog Index [33,5,16,30,17,64], LIX [16,30], Miyazaki EFL Readability Index [32,4], Dale-Chall [14,17,64], SMOG Grading [53,5,30,17,64], Linsear write formula [11,17,64] and others. These formulas often based on pre-calculated words of different types. So, this dimension can also consist various linguistic features. Depending on language version, it is possible to defined up to over 100-150 such measures [45,49]

3.5 Relevance

This dimension shows how popular or important for readers is selected Wikipedia Article. For this reason, it can be used such measures as articles age [15,68,70,39,51,60,59,37,41,30,28,81,67,13,61 number of page watchers [72,47], number of page visits [47,48], incoming internal link count (number of times that the article is cited by other Wikipedia articles) [15,30,28,65,72,47] and others, including more complex (e.q. PageRank [8]). Also it can be taken into the account measures that shows number of the links from external sources, such as Reddit [56], Facebook, Youtube, Twitter, Linkedin, VKontakte and other social services [44].

3.6 Style

Wikipedia articles with high quality must follows the style guidelines, including appropriate structure. So, one of the most simplest and popular measure for this dimension is number of the sections in the article [6,15,70,41,28,81,13,84,72,47]. Here also can be used such measures, as number of tables [4,6], number of templates [4,70,43,41].

3.7 Timeliness

Information on certain topics may change with time (living people, populated places etc.), therefore it is important that the article has actual data. Some measures can help to assess this dimension: number of unique editors and number of contributions for the last selected time. Measures of this quality dimension can be related to currency and volatility of the information [34].

3.8 Extraction Methods for Articles Measures

There are different possibilities and techniques to get measures values of the Wikipedia articles. The vast majority of the measures can be extracted from Wikipedia database dumps. Below is a list of some files for the latest dump of English Wikipedia [75] and a brief description of what can be extracted:

- **enwiki-latest-pages-meta-current.xml.bz2**: recombine all pages (including articles), current versions only. This file is used for obtaining a majority of the articles measures.
- **enwiki-latest-pages-articles.xml.bz2**: consist articles, templates, media/file descriptions, and primary meta-pages. Can be used also for obtaining a majority of the articles measures (excluding statistics from discussion pages).
- **enwiki-latest-pagelinks.sql.gz** : wiki page-to-page link records. Used for network measures - for example incoming links from other articles.
- **enwiki-latest-categorylinks.sql.gz**: wiki category membership link records. Can be used for category count measure.
- **enwiki-latest-externallinks.sql.gz**: wiki external URL link records. can be used for external link count measure.
- **enwiki-latest-stub-meta-history.xml.gz**: contain only historical revision metadata. Can be used to extract number of the editors from different groups (bots, anonymous users, administartors etc.) and alsa number of the edits of various types (e.g. minor edits, edits comments).
- **enwiki-latest-iwlinks.sql.gz**: Interwiki link tracking records. Can be used to extract number of the unique internal links (links to other Wikipedia articles).
- **enwiki-latest-templatelinks.sql.gz**: Wiki template inclusion link records. Used for templates count measure, also it is possible to check if article has infobox
- **enwiki-latest-page.sql.gz**: base per-page data (id, title, old restrictions, etc). Can be used to extract last edit time, page length in bytes.

– **enwiki-latest-imagelinks.sql.gz**: wiki media/files usage records. Can be used to image count measure.

Mention files can give different opportunities for extracting values of measures. For example, some of the studies count number of images by taken into the account tag [[image:...]] in the wiki markup (source code of the article) [6,15,68,82,81,67,13,84]. However, other images that are inserted (for example using special templates) will not be considered. Therefore, it can be used other approach, which extracted number of the images from wiki media usage records file [63,72,47,46,48,43].

Another example - number of internal links and number of incoming internal link (from other articles). It is possible to study the code of each article to find links, but links which was inserted by special templates will be not considered. Therefore, it can be used file with wiki page-to-page link records.

Some of the measures can not be extracted from dumps files. For example, to obtain number of page watcher for each article it is necessary to send request to Wikipedia API [20]. Measures from external resources (such as Facebook, Twitter, Reddit etc.) are also must be obtained from other sources.

3.9 Derivative Measures

Most of the related works took into the account combination of two or more measures. For instance, one of the most popular derivative measure is number of the references per article length [15,59,28,70,9,72,47,46,48,43]. Here length can be defined as volume in bytes [70,9,72,47,46,48,43], number of the words [15,4], number of the characters [59,28]

Some approaches based on normalised measures. For example, to build synthetic measure for Wikipedia articles quality online service WikiRank [78] use normalised values of 5 measures based on the threshold from Featured articles [72,46]. It is also possible to measure relative popularity using normalised value of some measures related to relevance quality dimension [48]. Some studies used log-transformed measures [9,41,71,72,47]

3.10 Multidimensional Quality Measures

Some measures can be related to two or more quality dimension. For example, editors count can show objectivity of the article (different point of view), but additionally can help to measure relevance of the content (more users are interested in this topic).

Another example - images count. On the one hand, pictures can help assess objectivity of the presented in the article material, but on the other hand we can measure completeness (because articles on a particular topic should contain pictures) and style (for example, to avoid writing a lot of text, the authors of the article decided to add more images).

Number of the citation templates [17,71,64,84] can help to measure quantity of the references (credibility) as well as in what degree information about the source is available for the reader (completeness).

4 Quality Measures of the Infoboxes

Interpretation of the quality of data depends on who will use this information [54]. Based on the literature [54,40,83,42]and own observations four important dimensions was selected to assess quality of the infoboxes: completeness, credibility, relevance, timeliness. Subsections below briefly describes quality measures of the infoboxes related to these dimensions.

4.1 Completeness

Completeness of the infobox can be measured as the ratio of the number of parameter values to the number of all defined parameters in the infobox of a given type. Other related to this quality dimension measure can also consider weights for each filled parameter, where weight is based on the frequency of filling this parameter [42]. Here we can also take into the account length of the infobox, number of templates and other elements that the infobox contains.

In some topics, infoboxes can consist similar parameters, which can be omitted when calculating completeness. For example, to describe cities of Poland in some language versions of Wikipedia there is a special infobox, so the parameter about country is absent. At the same time other languages to describe the same city can use common infobox for different cities in the world, so parameter about country is important there.

4.2 Credibility

As in the case of Wikipedia articles credibility is related to analysis of the references. Depending on the topic and the language version, within each infobox you can find parameters with similar references. To assess credibility it can be used such measures as number of references, number of unique references, references to filled parameters ratio.

4.3 Relevance

Data in infoboxes can be provided by different users. Relevance can be measured as number of unique authors of the infoboxes. Authors can be divided to different categories: bots, anonymous users, administrators etc.

4.4 Timeliness

For this dimension we must take into the account measure related to number of recent changes of whole infobox and its individual parameters. As in the case of Wikipedia articles, timeliness measures can be related to currency and volatility of the infoboxes.

5 Quality of the Infobox Parameter

Quality of each parameter of the infoboxes can also be evaluated. One of the important dimensions of quality is timeliness. It can be measured based on the values of specific parameters.

For example, often in infoboxes that describe cities, there is a parameter that indicates the date (year) when the population size was evaluated. However, most of the parameters do not have this additional information.

For example, in the same infoboxes about cities, there is no explicit information when the value of the parameter about the city mayor has been entered. This may be particularly important in the periods of local government elections, when the election results were announced, but officially the new mayor has still can not perform this function.

The chart 3 shows the history of changes in the "leader name" parameter of the Poznań infobox in the Wikipedia language versions in question from the moment of announcing the results of the exit pool on WTK television until the oath made by the new mayor of Poznań. On the basis of this chart, we can see that in Polish version changed parameter about mayor quickly after posting news on media portals. In addition, it can be seen that in the Polish Wikipedia there was no consensus on the value of the "leader name" parameter of the infobox in the article about Poznań in the presented period, because the new city mayor was formally elected, but can gain authority after taking the oath. However, in English Wikipedia there was no controversy on the subject, and the name of the new mayor was entered after the election results were announced, but a bit later than the Polish language version did. The Russian Wikipedia has twice changed the value of the parameter about the city mayor in the audited period. The first arose from the announcement of the results of the vote, and the second change arose from a minor correction of the name, according to the rules of transliteration to the Cyrillic alphabet. As for the Belarusian and Ukrainian Wikipedia - there were no changes there and the new value appeared much later. This is related to the fact that entering the value of this parameter in the Belarusian and Ukrainian version of the infobox is not mandatory, because the value can be automatically inserted from the Wikidata, where the value was updated almost 3 years after the announcement of the election and the oath.

6 Discussion and Future Work

In this paper quality measures and dimensions for quality assessment of Wikipedia articles and infoboxes were described.

Most of the previous studies works with the most developed language version of Wikipedia - English. Therefore, it is necessary to filter some of the measures (especially related to readability dimension) to be able to assess quality of articles between different languages.

Using machine learning and artificial intelligence algorithms proposed measures can help to build more accurate models for quality assessment of articles

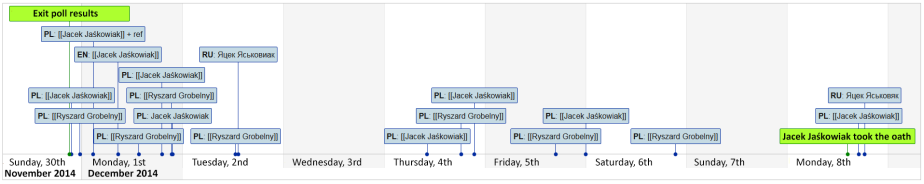

Fig. 3. History of changes of the "leader name" parameter of the infobox about Poznań in selected language versions of Wikipedia since the publication of the exit pool results on WTK television until the taking the oath by the new mayor of Poznań. Source: own study based on historical Wikipedia data.

and infoboxes in different language versions of Wikipedia. To build such models it is planned to use also cloud computing platforms such as Microsoft Azure [55]. Comparing the quality of information between different language versions of Wikipedia can also be done without taking into account other external sources of the related data (that can be closer to real-world description), by analogy with the theory of relativity [19].

Additional to improve the quality models, it is planned to use data from projects, which collect data from Internet users that compare quality of the multilingual information in Wikipedia. For example, project WikiBest [73] allows to choose best language version of infoboxes of particular topic in four nominations: the best quality, the best completeness, the best credibility, the best timeliness. User ratings can help to improve projects related to infoboxes evaluation [35].

Future works will be concentrated in defining new measures and in researches that will help to find dimensions and the most important measures for quality assessment of articles and infoboxes in multilingual Wikipedia.

References

1. Abramowicz, W., Auer, S., Heath, T.: Linked data in business. Business & Information Systems Engineering **58**(5), 323–326 (Oct 2016). https://doi.org/10.1007/s12599-016-0446-0
2. Alexa: Wikipedia.org traffic, demographics and competitors. `https://www.alexa.com/siteinfo/wikipedia.org`
3. Altmetric: Free tools. `https://www.altmetric.com/products/free-tools/`
4. Anderka, M.: Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. Phd, Bauhaus-Universitaet Weimar Germany (2013)
5. Blumenstock, J.E.: Automatically Assessing the Quality of Wikipedia Articles. Tech. rep. (2008). https://doi.org/10.1080/17439880802324251
6. Blumenstock, J.E.: Size matters: word count as a measure of quality on wikipedia. In: WWW. pp. 1095–1096 (2008). https://doi.org/10.1145/1367497.1367673
7. Bormuth, J.R.: Readability: A new approach. Reading research quarterly pp. 79–132 (1966)
8. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems **30**(1-7), 107–117 (1998)

9. De la Calzada, G., Dekhtyar, A.: On measuring the quality of wikipedia articles. In: Proceedings of the 4th workshop on Information credibility. pp. 11–18. ACM (2010)
10. Caylor, J.S., Sticht, T.G.: Development of a simple readability index for job reading material. (1973)
11. Chen, H.H.: How to use readability formulas to access and select english reading materials. Journal of Educational Media & Library Sciences **50**(2) (2012)
12. Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology **60**(2), 283 (1975)
13. Conti, R., Marzini, E., Spognardi, A., Matteucci, I., Mori, P., Petrocchi, M.: Maturity assessment of wikipedia medical articles. In: Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on. pp. 281–286. IEEE (2014)
14. Dale, E., Chall, J.S.: A formula for predicting readability: Instructions. Educational research bulletin pp. 37–54 (1948)
15. Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P.: Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 295–304 (2009). https://doi.org/10.1145/1555400.1555449
16. Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P.: Automatic Assessment of Document Quality in Web Collaborative Digital Libraries. Journal of Data and Information Quality **2**(3), 1–30 (2011). https://doi.org/10.1145/2063504.2063507
17. Dang, Q.V., Ignat, C.L.: Measuring quality of collaboratively edited documents: The case of wikipedia. In: Collaboration and Internet Computing (CIC), 2016 IEEE 2nd International Conference on. pp. 266–275. IEEE (2016)
18. DBpedia: Main page. `https://wiki.dbpedia.org`
19. Einstein, A.: The meaning of relativity. Routledge (2003)
20. English Wikipedia: Api sandbox. `https://en.wikipedia.org/wiki/Special:ApiSandbox`
21. English Wikipedia: Criticism of wikipedia. `https://en.wikipedia.org/wiki/Criticism_of_Wikipedia`
22. English Wikipedia: Featured article criteria. `https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria`
23. English Wikipedia: Featured articles. `https://en.wikipedia.org/wiki/Wikipedia:Featured_articles`
24. English Wikipedia: Good articles. `https://en.wikipedia.org/wiki/Wikipedia:Good_articles`
25. English Wikipedia: Verifiability. `https://en.wikipedia.org/wiki/Wikipedia:Verifiability`
26. English Wikipedia: Wikiproject tabular data. `https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Tabular_Data`
27. Eppler, M.J.: Managing information quality : increasing the value of information in knowledge-intensive products and processes ; ... 25 tables. Springer (2003)
28. Ferschke, O., Gurevych, I., Rittberger, M.: Flawfinder: A modular system for predicting quality flaws in wikipedia. In: CLEF (Online Working Notes/Labs/Workshop). pp. 1–10 (2012)
29. Filipiak, D., Filipowska, A.: Improving the quality of art market data using linked open data and machine learning. In: Abramowicz, W., Alt, R., Franczyk, B. (eds.) Business Information Systems Workshops. pp. 418–428. Springer International Publishing, Cham (2017)

30. Flekova, L., Ferschke, O., Gurevych, I.: What makes a good biography?: multidimensional quality analysis based on wikipedia article feedback data. In: Proceedings of the 23rd international conference on World wide web. pp. 855–866. ACM (2014)
31. Flesch, R.: A new readability yardstick. Journal of applied psychology **32**(3), 221 (1948)
32. Greenfield, G.R.: Classic Readability Formulas in an EFL Context: Are They Valid for Japanese Speakers? Ph.D. thesis, Temple University (1999)
33. Gunning, R.: The technique of clear writing (1952)
34. Hazen, B.T., Boone, C.A., Ezell, J.D., Jones-Farmer, L.A.: Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics **154**, 72 – 80 (2014). https://doi.org/https://doi.org/10.1016/j.ijpe.2014.04.018
35. Infoboxes.net: Quality comparison of infoboxes in miltilingual wikipedia. `http://infoboxes.net`
36. Juran, J., Godfrey, A.B.: Quality handbook. Republished McGraw-Hill pp. 173–178 (1999)
37. Kane, G.C.: A multimethod study of information quality in wiki collaboration. ACM Transactions on Management Information Systems (TMIS) **2**(1), 4 (2011)
38. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., Naval Technical Training Command Millington TN Research Branch (1975)
39. Kittur, A., Kraut, R.E.: Harnessing the wisdom of crowds in Wikipedia: quality through coordination. Proceedings of the ACM 2008 conference on Computer supported cooperative work - CSCW '08 p. 37 (2008). https://doi.org/10.1145/1460563.1460572
40. Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A.: Test-driven evaluation of linked data quality. In: Proceedings of the 23rd international conference on World Wide Web. pp. 747–758. ACM (2014)
41. Lerner, J., Lomi, A.: Knowledge categorization affects popularity and quality of wikipedia articles. PloS one **13**(1), e0190674 (2018)
42. Lewoniewski, W.: Completeness and reliability of wikipedia infoboxes in various languages. In: International Conference on Business Information Systems. pp. 295–305. Springer (2017)
43. Lewoniewski, W.: Enrichment of information in multilingual wikipedia based on quality analysis. In: International Conference on Business Information Systems. pp. 216–227. Springer (2017)
44. Lewoniewski, W., Härting, R.C., Węcel, K., Reichstein, C., Abramowicz, W.: Application of seo metrics to determine the quality of wikipedia articles and their sources. In: Damaševičius, R., Vasiljevienė, G. (eds.) Information and Software Technologies. pp. 139–152. Springer International Publishing, Cham (2018)
45. Lewoniewski, W., Khairova, N., Węcel, K., Stratiienko, N., Abramowicz, W.: Using Morphological and Semantic Features for the Quality Assessment of Russian Wikipedia, pp. 550–560. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-67642-5_46
46. Lewoniewski, W., Węcel, K.: Relative Quality Assessment of Wikipedia Articles in Different Languages Using Synthetic Measure, pp. 282–292. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-69023-0_24

47. Lewoniewski, W., Węcel, K., Abramowicz, W.: 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, pp. 613–624. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-46254-7_50
48. Lewoniewski, W., Węcel, K., Abramowicz, W.: Relative quality and popularity evaluation of multilingual wikipedia articles. Informatics **4** (2017). https://doi.org/10.3390/informatics4040043
49. Lewoniewski, W., Węcel, K., Abramowicz, W.: Determining quality of articles in polish wikipedia based on linguistic features. In: Damaševičius, R., Vasiljevienė, G. (eds.) Information and Software Technologies. pp. 546–558. Springer International Publishing, Cham (2018)
50. Lih, A.: Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. 5th International Symposium on Online Journalism p. 31 (2004)
51. Liu, J., Ram, S.: Using big data and network analysis to understand wikipedia article quality. Data & Knowledge Engineering (2018)
52. Lucassen, T., Schraagen, J.M.: Trust in wikipedia: how users trust information from an unknown source. In: Proceedings of the 4th workshop on Information credibility. pp. 19–26. ACM (2010)
53. Mc Laughlin, G.H.: Smog grading-a new readability formula. Journal of reading **12**(8), 639–646 (1969)
54. Mendes, P.N., Mühleisen, H., Bizer, C.: Sieve: linked data quality assessment and fusion. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops. pp. 116–123. ACM (2012)
55. Microsoft Azure: Cloud computing platform & services. `https://azure.microsoft.com/en-us/`
56. Moyer, D., Carson, S.L., Dye, T.K., Carson, R.T., Goldbaum, D.: Determining the influence of reddit posts on wikipedia pageviews. In: Ninth International AAAI Conference on Web and Social Media. pp. 75–82. AAAI Press Oxford, UK (2015)
57. O'Brien, J.A., Marakas, G.M.: Introduction to information systems, vol. 13. McGraw-Hill/Irwin New York City, USA (2005)
58. OECD Glossary of Statistical Terms: Iso 8402 - quality. `http://stats.oecd.org/glossary/detail.asp?ID=5150`
59. Ransbotham, S., Kane, G.C.: Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in wikipedia. Mis Quarterly pp. 613–627 (2011)
60. Ransbotham, S., Kane, G.C., Lurie, N.H.: Network characteristics and the value of collaborative user-generated content. Marketing Science **31**(3), 387–405 (2012)
61. di Sciascio, C., Strohmaier, D., Errecalde, M., Veas, E.: Wikilyzer: interactive information quality assessment in wikipedia. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces. pp. 377–388. ACM (2017)
62. Senter, R., Smith, E.A.: Automated readability index. Tech. rep., CINCINNATI UNIV OH (1967)
63. Shang, W.: A comparison of the historical entries in wikipedia and baidu baike. In: International Conference on Information. pp. 74–80. Springer (2018)
64. Shen, A., Qi, J., Baldwin, T.: A hybrid model for quality assessment of wikipedia articles. In: Proceedings of the Australasian Language Technology Association Workshop 2017. pp. 43–52 (2017)

65. Soonthornphisaj, N., Paengporn, P.: Thai wikipedia article quality filtering algorithm. In: Proceedings of the International MultiConference of Engineers and Computer Scientists. vol. 1 (2017)
66. Stróżyna, M., Eiden, G., Abramowicz, W., Filipiak, D., Małyszko, J., Węcel, K.: A framework for the quality-based selection and retrieval of open data - a use case from the maritime domain. Electronic Markets **28**(2), 219–233 (May 2018). https://doi.org/10.1007/s12525-017-0277-y
67. Stvilia, B., Twidale, M.B., Gasser, L., Smith, L.C.: Information quality discussions in wikipedia. In: Proceedings of the 2005 international conference on knowledge management. pp. 101–113. Citeseer (2005)
68. Stvilia, B., Twidale, M.B., Smith, L.C., Gasser, L.: Assessing information quality of a community-based encyclopedia. Proc. ICIQ pp. 442–454 (2005)
69. Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. Journal of management information systems **12**(4), 5–33 (1996)
70. Warncke-wang, M., Cosley, D., Riedl, J.: Tell Me More : An Actionable Quality Model for Wikipedia. In: WikiSym 2013. pp. 1–10 (2013). https://doi.org/10.1145/2491055.2491063
71. Warncke-Wang, M., Ranjan, V., Terveen, L.G., Hecht, B.J.: Misalignment between supply and demand of quality content in peer production communities. In: ICWSM. pp. 493–502 (2015)
72. Węcel, K., Lewoniewski, W.: Modelling the quality of attributes in wikipedia infoboxes. In: Abramowicz, W. (ed.) Business Information Systems Workshops, Lecture Notes in Business Information Processing, vol. 228, pp. 308–320. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-26762-3_27
73. WikiBest: Online game about comparing data quality between various languages of the wikipedia. `https://wikibest.net`
74. Wikidata: Main page. `https://www.wikidata.org/wiki/Wikidata:Main_Page`
75. Wikimedia Downloads: English wikipedia latest database backup dumps. `https://dumps.wikimedia.org/enwiki/latest/`
76. Wikipedia Meta-Wiki: List of wikipedias. `https://meta.wikimedia.org/wiki/List_of_Wikipedias`
77. Wikipedia Quality: Scientific works. `https://wikipediaquality.com/wiki/Category:Scientific_works`
78. WikiRank: Quality and popularity assessment of wikipedia. `https://wikirank.net`
79. Wilkinson, D.M., Huberman, B.A.: Assessing the value of coooperation in wikipedia. arXiv preprint cs/0702140 (2007)
80. Wilkinson, D.M., Huberman, B.a.: Cooperation and quality in wikipedia. Proceedings of the 2007 international symposium on Wikis WikiSym 07 pp. 157–164 (2007). https://doi.org/10.1145/1296951.1296968
81. Wu, K., Zhu, Q., Zhao, Y., Zheng, H.: Mining the factors affecting the quality of wikipedia articles. In: Information Science and Management Engineering (ISME), 2010 International Conference of. vol. 1, pp. 343–346. IEEE (2010)
82. Yaari, E., Baruchson-Arbib, S., Bar-Ilan, J.: Information quality assessment of community generated content: A user study of wikipedia. Journal of Information Science **37**(5), 487–498 (2011)
83. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web **7**(1), 63–93 (2016)
84. Zhang, S., Hu, Z., Zhang, C., Yu, K.: History-based article quality assessment on wikipedia. In: Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on. pp. 1–8. IEEE (2018)