Modelling the Quality of Attributes in Wikipedia Infoboxes

Krzysztof Węcel, Włodzimierz Lewoniewski

Poznań University of Economics,
Al. Niepodległości 10, 61-875 Poznań, Poland
{krzysztof.wecel,wlodzimierz.lewoniewski}@ue.poznan.pl

Abstract. Quality of data in DBpedia depends on underlying information provided in Wikipedia's infoboxes. Various language editions can provide different information about given subject with respect to set of attributes and values of these attributes. Our research question is which language editions provide correct values for each attribute so that data fusion can be carried out. Initial experiments proved that quality of attributes is correlated with the overall quality of the Wikipedia article providing them. Wikipedia offers functionality to assign a quality class to an article but unfortunately majority of articles have not been graded by community or grades are not reliable. In this paper we analyse the features and models that can be used to evaluate the quality of articles, providing foundation for the relative quality assessment of infobox's attributes, with the purpose to improve the quality of DBpedia.

Key words: data quality, information quality, DBpedia, Wikipedia, infobox, data mining, wikirank

1 Introduction

One of the important parts of a Wikipedia article is an infobox – the distinguished table, located in the top-right corner, which concisely presents the most important facts about a subject of the article, e.g. date of birth, length, area. An infobox allows to get acquainted with the most important facts about a subject of an article, without reading the article. The facts are expressed as attribute-value pairs and as such are more suitable for machine processing.

The same subject can be described in many languages in separate Wikipedia chapters. In general, articles in different languages are edited independently[1], therefore facts can differ. This also applies for infobox attributes – not only values can vary but also a set of attributes can be different for each language as language-specific structures have to be obeyed. This opens opportunities for learning new facts from cross-language comparison.

Our research hypothesis is that more accurate and complete facts will be observed in the articles which are distinguished for their quality by the Wikipedia community. The problem is that majority of Wikipedia articles have not been graded and comparing quality of articles in different languages is hampered. We have conducted an initial

[1] except those edited by multi-lingual editors and resulting from translation

analysis concerning the possibilities to build a model for automatic estimation of quality of a Wikipedia article based on measurable features of the article. We found that differences between languages seem to be significant, therefore separate models have to be built. For example, for Russian edition of Wikipedia length of an article is the most important feature determining the quality, while the number of references is leading in Polish and English Wikipedia.

The paper is organised as follows. In section 2 we start with the description of a general concept of data and information quality, followed by the approach to quality in Wikipedia. In section 3 various methods for quality evaluation are considered. Section 4 presents outcomes of the experiments and the discussion of the obtained results. Section 5 concludes the paper.

2 Data and Information Quality

2.1 General Concept of Quality

In the literature there is no agreed frontier between a quality of data and a quality of information. Very often they are defined by analogy to differences between data and information. Madnick et al have observed that there is a tendency to use notion "quality of data" in the technical context (e.g. data integration) and "quality of information" in personal contexts (e.g. relevance of information) [1]. Heinrich & Klier explain the quality of data as a multidimensional construct embracing multiple dimensions, e.g. precision, completeness, timeliness, consistency. Each dimension contributes its own view on a quality of attributes in an information system [2].

Only recently the Semantic Web community started serious research in the area of linked open data quality. Behkamal et al defined five features specifying the quality of such data: semantic accuracy, syntactic accuracy, uniqueness, consistency and completeness [3]. In DBpedia, for some of the attributes we should also add another important characteristics – timeliness.

The research in the area of quality of information is well established. There is a plethora of approaches to define attributes of information for the purpose of quality evaluation. A good summary is presented by Eppler who proposed 70 attributes of information, then narrowed the list to 16 most important [4].

Methods and criteria for quality evaluation differ also in various domains, e.g. business, medical or technical information. For example, the Commission of the European Communities have elaborated dedicated quality criteria for web pages related to healthcare. In this case the quality of web page (effectively, of information) is measured based on the following criteria: transparency, honesty, authority, privacy and data protection, updating of information, accountability, accessibility [5]. For information published in encyclopaedia the following quality features should be considered: scope, format, uniqueness, accuracy, currency, accessibility [6].

2.2 Quality in Wikipedia

The notion of the quality in Wikipedia is not language-independent, nor stable over time. Currently Wikipedia has 279 active language editions[2]. Each language has its own community, which can define own quality criteria within their edition, independently of other languages [7]. Quality evaluation systems are established in an evolutionary process through discussion about rules and principles for granting awards for quality.

Virtually in all editions of Wikipedia highest grades are awarded only by a community, after an extensive discussion. They are equivalent to English edition's "featured article (FA)" and "good article (GA)". Less respected grades can be attached by users themselves, according to the scale agreed by the community. In English edition there are additionally: A-class, B-class, C-class, start and stub. Altogether there are seven quality classes.[3]

In some cases different language editions have similar quality classes, e.g. Russian and Ukrainian. Nevertheless, rules for assigning articles to classes can still vary. For example, in English Wikipedia the featured article has to be: well-written, comprehensive, well-documented, neutral, in good style, with multimedia elements. The same quality class in German requires: references to reputable literature, balanced scope, high quality and up to date sources (evidence), proper linguistic style, good design including graphics and photos.

Not all language editions use the extensive grading scale. In some cases only the two highest grades can be awarded and German Wikipedia edition is an example. It distinguishes only two levels: *Exzellente Artikel* (featured article) and *Lesenswerte Artikel* (good article). As a consequence, it is not possible to track the evolution of an article.

Table 1 presents numbers of articles of various quality in different language editions. The biggest number of classified articles can be found in the biggest edition of Wikipedia – English. It is worth to note that even though it contains over 4.7 million articles, over 3.8 million are stub or start, i.e. over 80% of articles are problematic from the quality point of view. Lower classes of quality are frequently used also in Russian and Ukrainian editions. In Polish edition low grades are assigned infrequently and German edition does not use these classes at all. Number of featured articles is in most cases proportional to the overall number of articles – ca. 0.06%. Only German edition has much bigger share of 0.13%.

3 Quality Evaluation Methods

The quality of information should be understood as a degree to which user needs are satisfied [8]. Methods for quality evaluation can be divided into objective and subjective. In information manufacturing systems, quality of raw data is evaluated using an objective approach, whereas quality of information products is judged by people who are the customers receiving information, hence a subjective approach [9]. The objective approach to quality evaluation consists in defining rules and patterns that can be applied

[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias
[3] https://en.wikipedia.org/wiki/Template:Grading_scheme

Table 1. Number of articles of various quality in different language editions (as of June 2015). Similar classes are grouped

Name/Language	BE	DE	EN	PL	RU	UK
Number of articles	89,923	1,828,090	4,741,168	1,099,441	1,198,199	557,590
Featured Article	54	2,383	4,481	648	768	204
Good Article	97	3,785	21,697	1,993	2,041	531
Solid article					1,661	
A-class			821			
Four				157		
Full article					3,978	195
B-class			80,919			
Developed					15,492	707
C-class			194,063			
Developing					60,801	3,490
Start			1,177,495	928		
Stub	860		2,634,800	2,071	429,926	189,742
Unevaluated	88,912	1,821,922	626,892	1,093,644	683,532	362,721

to automatically determine the quality of data in a database. The subjective approach assumes involvement of users who are asked to grade the usefulness of delivered information. The fundamental difference in these approaches can lead to a problematic situation when objective method will assign high quality grade for data that are not useful for the user. For example, very detailed and complete information (objective) can be hard to understand (subjective). In the subjective quality evaluation, the statistical approach is prevailing for identification of the most important features influencing the perceived quality of information [10]. Both approaches are applied in this paper and in fact a relation between them is studied.

In Wikipedia, there are generally two groups of grades: those that can be awarded after a discussion by a community (higher quality classes) and those that can be assigned by users alone (lower quality classes, reflecting rather the stage of the development of an article). The highest quality classes, mostly "featured article" and "good article", can only be awarded as a result of a positive voting. The voting has to be initiated by an interested user, typically the main contributor, who very often also works on improving the article. As not all users are interested in starting the evaluation process, there is a problem of big number of unevaluated articles (see table 1).

Our concern is also the subjectiveness of the assignment of lower quality classes. A quality class can be assigned by a single user, without any prior discussion. A user interested in self-promotion might assign a higher class to the article than it would result

from the rules and principles. It would be desirable to devise a method for verification of such grades.

In a longer term, the evolution of quality criteria has also to be considered. The rules for grading can change and as a result some articles can bear the out-of-date class. For example, rules in English Wikipedia have been amended many times since the inception in 2002, making requirements more and more strict. This forced verification of the classification of articles and over 1000 articles, which had been awarded on less restrictive conditions, lost the grade "features article". In Polish edition, over 200 articles have lost the "medal".

Taking above into consideration, there is a strong need for automation of the grading of articles. This research area is already relatively well-developed. Many models have been proposed, which classify Wikipedia articles to appropriate classes based on metadata and different measures that can be extracted from the article. The relatively simple approaches base on measuring the "volume" of contents (number of letters, images, headers, references) [11, 12, 13, 14] – the more content, the higher the quality. Some external measures can be added for quality estimation, like number of links to the article, Gunning fog index (measures the readability of English writing) and others [15]. The behaviour of contributors has also been included in some models.

Unfortunately, majority of methods focused on English Wikipedia only and have not taken internationalisation dimension into account. Our contribution is in considering many languages in a linked way, focusing both on popular (English) and less developed language editions (Belarussian).

We consider the hypothesis that the overall quality of a Wikipedia article (as decided by users) allows to derive the quality measure of attributes in the infobox. The reasoning can be as follows: an editor wishing to submit an article to the evaluation process polishes the contents and also fixes all possible issues. However, the other hypothesis can be tested as well – attributes in an infobox are part of the article and as such are also subject to quality evaluation. Therefore, quality of attributes influences the overall quality of an article. Our initial experiments confirm this relation. In our approach we will assume a two-directional mutual influence. From statistics point of view, we can calculate correlation between those "qualities", i.e. determine that there is a relation but we cannot say anything about direction of the influence.

As already mentioned, quality of data and quality of information are characterised by different features. Attributes in an infobox are equivalent to data and an article is equivalent to information. Various sets of features are presented in figure 1. These features can be measured with various effectiveness and precision. Some of them characterise the same aspect in a different way, thus they can be related.

In our overall research plan, estimated quality of articles is input to our next method, which should estimate the quality of attributes contained in infoboxes. For this we need to study in greater details the quality of attributes contained in infoboxes, hence the relation to linked data and DBpedia. Reliable methods for quality evaluation are then crucial.

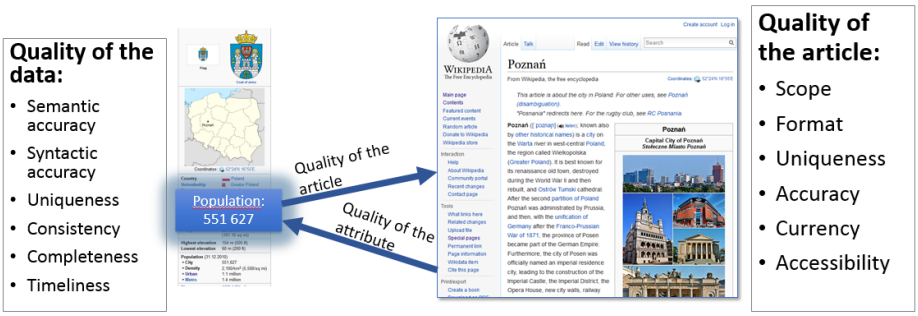

Fig. 1. Features related to quality of an article and an infobox

4 Modelling the Quality of Wikipedia Articles

In order to build a model, we need first to prepare a training dataset. Using Wikipedia API it is possible to obtain lists of articles belonging to certain quality class. We have first sampled articles of each quality class in each language edition separately: Belarussian (BE), German (DE), English (EN), Polish (PL), Russian (RU) and Ukrainian (UK). We initially planned to have 2000 articles in each class but it was only possible for English and German. Unfortunately, in the case of the less popular languages, some classes encompassed smaller number of articles and consequently number of sampled articles had to be reduced to keep the numbers equal in each class. Once the limits have been determined, we randomly selected articles for each class in each language edition. The statistics concerning sampling is presented in table 2.

Table 2. Size of sample in various language editions

Language	Number of articles in each class
English (EN)	2000
German (DE)	2000
Russian (RU)	744
Ukrainian (UK)	191
Polish (PL)	152
Belarusian (BE)	52

Using own software *WikiAnalyzer2*, we have extracted numerous parameters characterising articles, which are normally available via various APIs, e.g. length, number of letters, observers, editions, incoming links. We have also added derived parameters, e.g. references/lenght. Prepared datasets were then subject to exploration.

We have analysed the distribution of values of various attributes in relation to quality classes. It is basically not possible to unambiguously assign article to the class, based only on one attribute. Even though the median of the value of analysed attribute is increasing, the range of values is so wide that it covers several classes. This phenomenon

is presented in figure 2. Left chart presents the distribution of length of articles in bytes (horizontal axis) in relation to seven quality classes in English (vertical axis), where JA7 (top of the chart)is the worst class, and JA1 (bottom of the chart) is the best class. Right chart presents number of headers in English articles. Concluding, we need to increase the discriminative power of attributes in order to more precisely distinguish various classes.

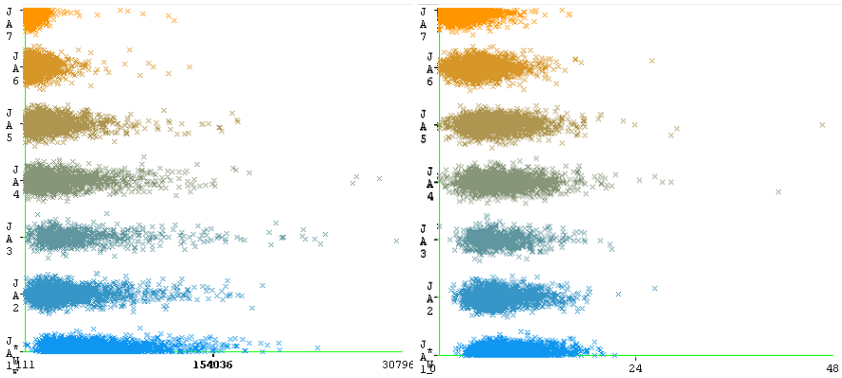

Fig. 2. Attribute values by quality classes: length of article (left), number of headers (right)
Source: Statistica

Our initial findings that can roughly be summarised as "the more content, the higher class" have been confirmed in several data mining models. Figure 3 presents scaled average deviations of attribute values as calculated by boosting tree model in SAS Enterprise Miner. It can be clearly observed that these values are organised in layers, i.e. attributes of better articles (towards JA1, top of the chart) have higher values than worse articles (towards JA7, bottom of the chart).

Based on the above observation we have built a baseline model for quality class prediction. It is basically a linear regression model, which takes into account variables that are positively correlated with quality. There is, however, small modification that allows to cope with the excessive values of some of the attributes. For any given variable (for given language) we calculate the median value in the highest quality class. This value is used as a threshold. If the value of the given attribute exceed the threshold, it is set to 100 points, otherwise its value is linearly scaled to reflect the relation of the value to the median value. Let us assume that the median for the number of images in the highest class is 20. Any article with higher number of articles will score 100 for this variable; article with 12 images will get proportionally 60 points ($\frac{12}{20} \times 100$). The score for quality is calculated as a weighted average of single transformed variables, where weights are derived from significance of these variables as estimated by linear regression model.

The above procedure for calculation of the quality score is underlying the web application[4] for automatic estimation of quality for easier comparison of articles about

[4] alfa version available at `http://wikirank.net`

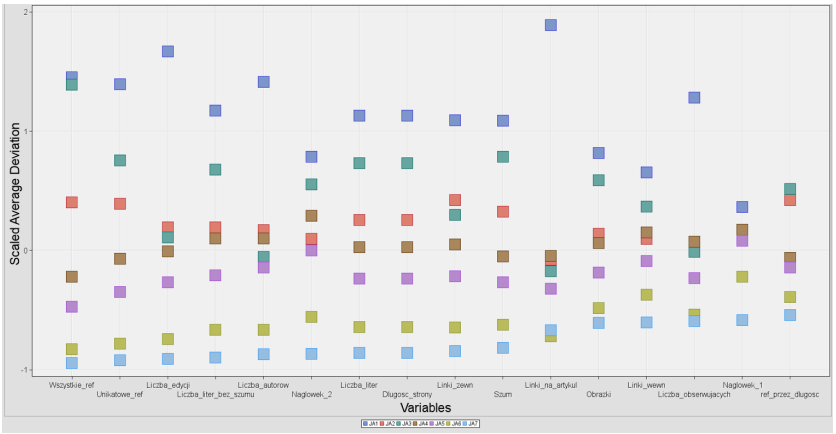

Fig. 3. Relation between volume of contents and quality
Source: boosting tree model in SAS Enterprise Miner

single subject in various languages. Sample results are presented in figure 4 – *Poznań* is best described in Polish language where it obtained 93.05 points. The worst description, according to the model, is in Belarussian – only 26.73 points. In Polish, attributes *length*, *number of references*, *images* and *headers* are no less than the median. Only attribute *references by length* had value of 65.27% of the median. This approach is not perfect but works correctly for language-sensitive subjects, e.g. Berlin is best described in German (100 pts for quality), Liverpool – in English (89.93 pts), Lviv – in Ukrainian (100 pts).

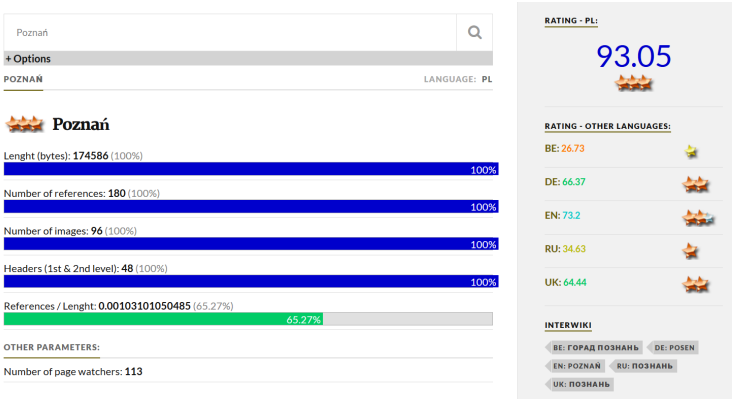

Fig. 4. Wikirank screenshot - Poznań

More rigorous approach to quality prediction is offered by data mining techniques. We have applied a range of them, in various applications with various settings. In data mining, it is important to set the scale on which the target variable is measured. It

determines which data mining techniques can be used for prediction. Thus, quality can be modelled in different ways. Initially we have modelled quality as a nominal variable, i.e. values were discrete, e.g. “good article”, “stub”. Binary variable is a special case of the above, where only two distinct classes are considered, e.g. “good” and ”useless”. Quality can also be modelled as interval variable. This approach requires to assign a real number value in range 0.0 to 1.0 (or 0 to 100) for each class, not necessarily according to linear scale. This process is called variable encoding and the way it is conducted is crucial for good quality models. The last case is not considered in this paper.

We have first built models for quality as a nominal variable, using various state of the art methods from software like SAS Enterprise Miner, Statistica and WEKA. Finally, for the prediction of a quality class we have used the Random Forests classifier, which provided results with high precision in similar tasks [14]. Results for this model are presented in table 3. Unfortunately, they are not satisfactory as only Polish and Belarussian have misclassification rate smaller than 40% for test dataset. It was mostly caused by bigger number of classes and small differences between them (compare figure 2). Model for German language was one of the worst but at the same time it was the most stable (similar results for training and test).

Table 3. Misclassification for models with the quality as the nominal variable

	Misclassification rate	
Language	Train	Test
Belarusian (BE)	.150	.235
English (EN)	.382	.436
German (DE)	.450	.465
Polish (PL)	.273	.365
Russian (RU)	.383	.436
Ukrainian (UK)	.349	.503

In the second approach – quality as binary variable – all articles have been split into two classes: 1 – complete (labelled GoodEnough), including FA and GA, which are the most reliable grades (voting); 0 – developing (labelled NeedsWork), including all other articles from lower quality classes. This ordering bases on an observation that practically in all language editions there are quality classes that are equivalent to English featured article (FA) and good article (GA). Table 4 presents result for this approach, which are significantly better than in the case of nominal target variable.

The binary approach may be preferred to nominal. First of all, results are much more precise in terms of misclassification rate.[5] Second, instead of having various number of quality classes in different languages we have just two in all cases. Thus, interpretation of results can be homogenised. Third, outcomes are more reliable as only the articles ranked in a voting by a community are considered for quality calculation.

[5] This is obvious as with reduced number of classes we avoid misclassification within combined classes

Table 4. Misclassification for models with the quality as the binary variable

	Misclassification rate	
Language	Train	Test
Belarusian (BE)	.014	.000
English (EN)	.115	.125
German (DE)	.033	.036
Polish (PL)	.087	.079
Russian (RU)	.067	.081
Ukrainian (UK)	.082	.092

For the models that we have built, significance of variables has also been calculated. Attributes along with their significance form a specific profile of a language, i.e. one attribute is important for one language, another better characterises quality of the other language. It is then possible to compare various languages, what is presented in figure 5. The profiles indicate that in each language various features have different significance.

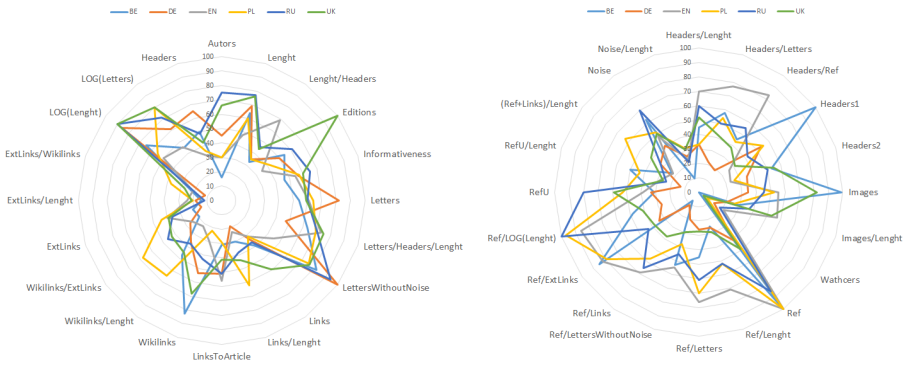

Fig. 5. Significance of variables, the quality as the binary variable
Source: *Statistica*, random forest of 100 trees

Verification of the proposed method has been done manually. We actually evaluated how good is the method in spotting the best articles among various language versions about the same topic. In the experiment, we have based on a simple assumption: if there are six articles, each in another language, and only one article is featured, then our method should identify this article as carrying the most up-to-date information in infoboxes. The process of preparing the dataset is presented in figure 6. Based on DBpedia, we first queried for all articles in English Wikipedia that belong to the class `PopulatedPlace`. There were over 500,000 such locations. In next step, we only considered locations that have description in all six analysed languages (BE, DE, EN, PL, RU, UK), and we obtained exactly 10,000 populated places. Finally, we have se-

lected only 6-tuples in which only one article, in any language edition, is graded as featured article. Last filtering left us with just 385 articles. For manual verification of the quality of attribute *Population*, we checked if the best article indeed provides the most accurate value for population. It was the case only in 60% of articles, so this is the baseline precision that we aim to improve.

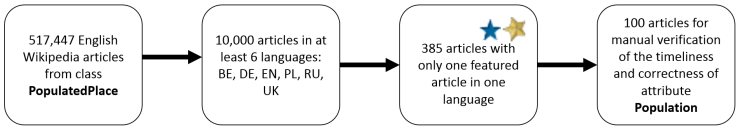

Fig. 6. Data flow in timeliness experiment

Concluding the section about modelling the quality of Wikipedia articles we need to state that there is no single, simple method to predict the quality of the article. Best results, in terms of misclassification rate, are obtained by binary model, but at the same time this model is the least useful for our future purposes – the relative quality of an attribute to point the language version with the most correct attribute value. The most useful results would be provided by models using the interval scale for measuring the quality but we need to work on value encoding to obtain better results.

5 Conclusions and Future Work

Our approach to quality evaluation of attributes in infoboxes can be classified as analytical. We have built numerous data mining models, covering various aspects: from discovering relations between attributes to prediction of a quality class. Our results are exploratory in nature and there is room for improvements. Below we put some of the ideas.

Neutrality of articles is one of the most important criteria for good articles. It would be then interesting to add features related to sentiment analysis. Articles with positive or negative polarity, i.e. not neutral, should be penalised and obtain lower quality grade. Sometimes flaws in articles are provided explicitly, in dedicated templates, such as *unreferenced*, *citation needed*, *orphan*, *dead link*, *notability*, *no footnotes* [6].

Another group of features can be labelled as external, as they are not directly bound to an article. Quality of references is one of the best examples – trust in referred data source is then transmitted to the attribute. One can assume that articles referring to sources maintained by governmental bodies tend to offer more accurate information and should be generally of higher quality. Open data comes into play here.

Our special interest is nevertheless the international dimension in quality evaluation. Analysis of multi-lingual Wikipedia users is a promising direction. These people generally introduce changes in several language editions of the same article, making this better comparable and perfect for matching. Information is transferred from one language to the other, and the question is what is the impact of such changes on quality.

The relation between infoboxes (data) and article's text (information) will be researched in next phase.

References

1. Madnick, S.E., Wang, R.Y., Lee, Y.W., Zhu, H.: Overview and Framework for Data and Information Quality Research. ACM Journal of Data and Information Quality **1**(1) (2009) 1–22
2. Heinrich, B., Klier, M.: Metric-based data quality assessment — Developing and evaluating a probability-based currency metric. Decision Support Systems **72** (2015) 82–96
3. Behkamal, B., Kahani, M., Bagheri, E., Jeremic, Z.: A metrics-driven approach for quality assessment of linked open data. Journal of Theoretical and Applied Electronic Commerce Research **9**(2) (2014) 64–79
4. Eppler, M.J.: Managing information quality : increasing the value of information in knowledge-intensive products and processes ; ... 25 tables. Springer (2003)
5. Commission of the European Communities: eEurope 2002: Quality criteria for health related websites (2002)
6. Anderka, M.: Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. Phd, Bauhaus-Universitaet Weimar Germany (2013)
7. Stvilia, B., Al-Faraj, A., Yi, Y.J.: Issues of cross-contextual information quality evaluation-The case of Arabic, English, and Korean Wikipedias. Library and Information Science Research **31**(4) (2009) 232–239
8. Abramowicz, W.: Filtrowanie informacji. Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań (2008)
9. Ge, M., Helfert, M.: Data and information quality assessment in information manufacturing systems. In: 11th International Conference, BIS 2008, Innsbruck, Austria, May 5-7, 2008. (2008) 380–389
10. Xu, H.: What are the most important factors for accounting information quality and their impact on ais data quality outcomes? J. Data and Information Quality **5**(4) (March 2015) 14:1–14:22
11. Hu, M., Lim, E.P., Sun, A., Lauw, H.W., Vuong, B.Q.: Measuring article quality in wikipedia. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management - CIKM '07. (2007) 243–252
12. Blumenstock, J.E.: Size matters: word count as a measure of quality on wikipedia. In: WWW. (2008) 1095–1096
13. Wöhner, T., Peters, R.: Assessing the quality of Wikipedia articles with lifecycle based metrics. Proceedings of the 5th International Symposium on Wikis and Open Collaboration WikiSym 09 (2009) 1
14. Warncke-wang, M., Cosley, D., Riedl, J.: Tell Me More : An Actionable Quality Model for Wikipedia. In: WikiSym 2013. (2013) 1–10
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. (2009) 295–304