Reliability of User-Generated Data: the Case of Biographical Data in Wikipedia

Robert VISEUR
CETIC
Rue des Frères Wright, 29/3
B-6041 Charleroi
robert.viseur@cetic.be

UMONS Faculty of Engineering
Rue de Houdain, 9
B-7000 Mons
robert.viseur@umons.ac.be

ABSTRACT
Wikipedia is a collaborative multilingual encyclopedia launched in 2001. We already conducted a first research on the extraction of biographical data about personalities from Belgium in order to build a large database with biographical data. However, the question of the reliability of the data arises. In particular, in the case of Wikipedia, the data are generated by users and could be subject to errors. In consequence, we wanted to answer to the following question: are the data introduced in Wikipedia articles reliable? Our research is organized in three sections. The first section provides a brief state of the art about the reliability of the user-generated data. A second section presents the methodology of our research. A third section will present the results. The error rates that were measured for the birthdate is low (0.75%), although it is higher than the 0.21% score that we observed for the baseline (reference sources). In a fourth section, the results are discussed.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries - *collection*, *dissemination*, *standards*, *systems issues*, *user issues*.

General Terms
Reliability.

Keywords
wikipedia, open data, reliability, data quality, data extraction, information retrieval, biography.

1. INTRODUCTION
Wikipedia (wikipedia.org) is a collaborative multilingual encyclopedia launched in 2001. The project is financially supported since 2003 by the Wikimedia Foundation (wikimediafoundation.org). The volume of the encyclopedia has grown steadily since its inception. In January 2013, the largest editions of Wikipedia were English edition (more than four million articles), German edition (more than one and a half million articles), French edition (more than one million three hundred thousand articles) and Dutch edition (over one million one hundred thousand articles).

We already conducted a first research on the extraction of biographical data about personalities from Belgium [14, 15, 16]. Indeed, using Wikipedia for supplying a biographical database seems appropriate, due to the breakdown by type of content within the encyclopedia. The articles related to biographies represent 15% of the total content in January 2008, behind the articles about culture and the arts [11].

The data are extracted in order to build a large database with biographical data. The extraction and integration steps in a common database were successfully completed [16]. However, the question of the reliability of the data that are entered into this unified database arises. In particular, in the case of Wikipedia, the data are generated by users -the term “crowdsourced” was not used here given the discussions on the Wikipedia character to be crowdsourced or not crowdsourced [7]- and could be subject to errors.

On the basis of those data, we want to answer to the following question: are the data that are introduced in Wikipedia articles reliable? Our results are based on the comparison between data from Wikipedia and data from reference databases. The data that we used for the comparison are the birthdates extracted from biographical articles.

Our research is organized in three sections. The first section will provide a state of the brief art about the reliability of the user-generated data. A second section will present the methodology of our research. A third section will present the results. In a fourth section, the results will be discussed.

2. STATE OF THE ART
The quality of data that are extracted from websites whose the content is generated by users is the subject of discussions and researches in the recent years. This is especially true for websites like OpenStreetMap (www.openstreetmap.org), an alternative to Google Maps (maps.google.fr), which is free and supplied by its users, or the free encyclopedia Wikipedia (www.wikipedia.org). Everyone can contribute to the content of these websites: how to ensure a quality that is comparable to commercial services that require a quality process that is considered as more strict and systematic?

In practice, the quality of Wikipedia is comparable to its commercial competitors [3, 5, 10]. OpenStreetMap quickly improves its coverage of territory [9]. The researchers particularly point out the practice of peer review (if the community is large enough, the errors will be detected and can be corrected) and the fact that the users who contribute to the content may have knowledge that companies do not have [2, 8, 9].

The methods for measuring the quality generally revolve around two strategies.

The first method consists in using control data. The latter have been audited by experts. This technique is used for geographical

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
Copyright is held by the owner/author(s).
OpenSym '14 , Aug 27-29 2014, Berlin, Germany
ACM 978-1-4503-3016-9/14/08.
http://dx.doi.org/10.1145/2641580.2641581

information provided by users [6]. The main limitation is the errors that may exist in the control data.

The second method consists in taking a community assessment criterion as a criterion of quality. In the case of Wikipedia, it is usually the value of "featured article" associated with remarkable articles. On this basis, the researchers try to develop models and to determine the characteristics of a quality article. It is clear from these studies that criteria such as the length of the article, the number of external references, the number of contributors, the number of editions or the length of the discussion page for an article would have a positive impact on quality of articles [2, 5, 12, 13, 17]. Thus, these criteria could be used to build a combination of metrics for estimating the quality of a Wikipedia article.

3. METHODOLOGY

To assess the reliability of data included in the Wikipedia encyclopedia, we used the principle of comparing the data extracted from Wikipedia with data from 9 reference databases that were supplied by experts and provided by the sponsor (anonym) of this study. We were aware of the risk of errors in the control data. In consequence, when the values were different, we searched additional sources of information in order to determine if the error was in the Wikipedia encyclopedia or in the reference sources.

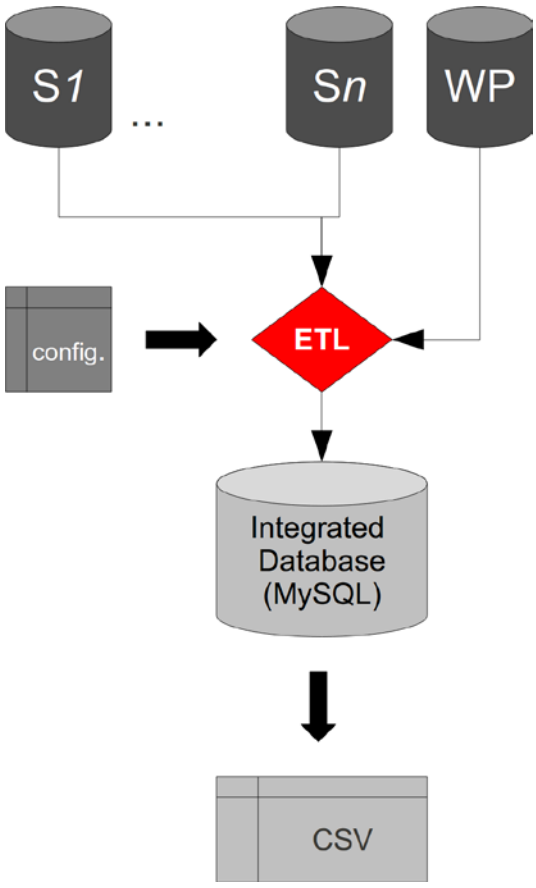

Figure 1. Fusion of the 10 data sources.

Our evaluation of the reliability of data is based on the personalities' birth year. We used the people with an entry in the Wikipedia online encyclopedia and at least one entry in one of the 9 reference databases.

In practice, the 10 data sources were integrated into a single database through an ETL (Extract- Transform- Load) tool that was developed for the needs of the study (see Figure 1). This tool uses a configuration file to specify the rules to homogenize the names and the dates. It also generates a set of keys that helps to associate multiple entries for the same person. Once they are homogenized and complemented by their keys, the records are loaded into a MySQL database that allows their manipulation (via SQL queries) and the export of work files.

In practice, a file with 938 lines was created from the integrated database (see Figure 2). Each line represents a people that is characterized by the birthdate that is mentioned in the 10 databases that we used (including Wikipedia). The null value is assigned when the corresponding entry does not exist in a database or when the birthdate is unknown.

Albert Bruylants	0	1915	0	0	1915	0	0	0	0	0
Thomas Buffel	0	1981	0	0	0	0	0	1981	0	0
Auguste Buisseret	0	1888	0	0	1888	0	0	0	0	0
Charles Buls	0	1837	0	0	1837	0	0	0	0	0
Ernest Burnelle	0	1908	0	0	1908	0	0	0	0	0
Jan Burssens	0	1925	0	0	1925	0	0	0	0	0
Max Buset	0	1896	0	0	1896	0	0	0	0	0
Yoni Buyens	0	1988	0	0	0	0	0	1988	0	0

Figure 2. Example of entries in CSV export file.

4. RESULTS

We first considered that the entries were correct when the years (birthdates) were the same in the different databases. We kept the entries with different years in the different sources. It represents 14.4% of the entries. These entries may contain an error or simply represent homonyms.

We conducted a manual review of these entries in order to distinguish errors due to the extraction process, errors in Wikipedia and errors in data from reference sources. The errors in the reference sources were determined by comparison with reference websites such as foundations websites, museums websites, etc. that are directly or indirectly dedicated to individuals involved.

Table 1. Error rates (birthdate)

Errors	%
Errors in Wikipedia	0.75%
Extraction errors	1.71%
Errors in reference sources	0.21%
Undetermined	0.75%

The error rate in Wikipedia is 0.75%, against 0.21% for the reference sources. In less than one percent of cases, due to the lack of information in reference sources, it was not possible to determine if it was a homonym or an error. In 1.71% of cases, the error in the file was due to data extraction error in the text of Wikipedia. That is a result very close to the 1.9% that we obtained in our previous study by using the dates that are in the Infobox as data control to evaluate the quality of data extraction [14, 15, 16].

5. DISCUSSION

The reliability of data was unknown at the start of the project. The quality of Wikipedia content has been studied for several years. The results tend to reassure, at least for items that receive a major peer review activity. The error rate that was measured for the birthdates is 0.75 %. This error rate is low, although it is higher than the 0.21% score that we observed for the reference sources

Note that the evaluation method that we developed for this study suffers from a limitation. Indeed, the comparisons are made for personalities that are encoded in multiple databases. We can therefore assume that they are famous people, whose articles are good candidates for a more intensive peer review activity.

We did not address the evolution in time of the reliability of articles in this study. However, the Alfonseca's study [1] provides an encouraging information because he found that the dates of death that were in the Infobox were updated within 2 days after the death of the personality (for 50% of the deceased people). In addition, we found, in the few errors reported in Wikipedia for the birthdates, that errors had been corrected (with the addition of a source) between the time when the collection of data was done and the time when the study about the reliability was conducted. More generally, the Javanmardi and Lopes' study [10] shows that the quality of articles tends to increase over time and that the variation in quality tends to decrease.

It is interesting to note that the personalities included in our database are essentially personalities still alive or recently deceased. The average year of birth is 1880. However, this value is quite close to that one of the reference sources (1878). The over-representation of recent personalities is common with other language versions (see the page "*Kategorie Diskussion : Person nach Geschlecht*" in Wikipedia.de). The variation around this average year is greater in the case of Wikipedia (156 vs 66). The biographical articles in Wikipedia are therefore not limited, as the user-generated aspect could let fear, to contemporary figures close to the general public. The recent controversy about the refusal to add a page about the Nabilla Benattia French starlet also illustrates the care provided by Wikipedia to keep the encyclopedic interest of the biographical articles [4].

Automating the detection of records containing false data would be possible by using measures of the quality of the articles that were identified in the literature. Such a method would be useful to improve the quality of data that were extracted from Wikipedia and were integrated in a multi-source database.

6. REFERENCES

[1] Alfonseca, E. 2013. Distributing the Edit History of Wikipedia Infoboxes, Google Research, May 30, 2013. URL: http://googleresearch.blogspot.fr/2013/05/distributing-edit-history-of-wikipedia.html (read: April 25, 2014).

[2] Blumenstock, J. E. 2008. Size matters: word count as a measure of quality on wikipedia, Proceedings of the 17th international conference on World Wide Web, ACM, pp. 1095-1096.

[3] Brown, A. R. 2011. Wikipedia as a data source for political scientists: accuracy and completeness of coverage, PS Political Science and Politics, 44(2), pp. 339-343, 2011.

[4] Cesbron, M. 2013. Nabilla n'est pas «un objet de savoir» pour Wikipédia, LeFigaro.fr, April 29, 2013. URL: http://www.lefigaro.fr/culture/2013/04/29/03004-20130429ARTFIG00429-nabilla-n-est-pas-un-objet-de-savoir-pour-wikipedia.php (read: April 25, 2014).

[5] Chevalier, F., Huot, S., Fekete, J. D. 2010. Visualisation de mesures agrégées pour l'estimation de la qualité des articles Wikipedia, EGC 2010: Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances.

[6] Comber A., See L., Fritz S., Van Der Velde M., Perger C., Foody G. 2013. Using control data to determine the reliability of volunteered geographic information about land cover, International Journal of Applied Earth Observation and Geoinformation, Vol. 23, pp. 37-48.

[7] Estelles-Arolas, E., Gonzalez-Ladron-de-Guevara, F. 2012. Towards an integrated crowdsourcing definition, Journal of Information science, 38(2), pp. 189-200.

[8] Goodchild, M. F., Li, L. 2012. Assuring the quality of volunteered geographic information, Spatial statistics, Vol. 1, pp. 110-120.

[9] Heipke, C. 2010. Crowdsourcing geospatial data, ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 65(6), pp. 550-557.

[10] Javanmardi, S., Lopes, C. 2010. Statistical measure of quality in Wikipedia, Proceedings of the First Workshop on Social Media Analytics, ACM, pp. 132-138.

[11] Kittur, A., Chi, E.H., Suh, B. 2009. What's in Wikipedia?: Mapping Topics and Conflict using Socially Annotated Category Structure, Proceedings of the 27th international Conference on Human Factors in Computing Systems, April 04-09, 2009.

[12] Stvilia, B., Twidale, M. B., Gasser, L., Smith, L. 2005a. Information quality discussions in Wikipedia. Proceedings of 2005 ICKM conference, pp. 101-113.

[13] Stvilia, B., Twidale, M.B., Smith L.C., Gasser, L. 2005b. Assessing information quality of a community-based encyclopedia, Proceedings of the International Conference on Information Quality, pp. 442-454.

[14] Viseur, R. 2013a. Extraction de données biographiques depuis Wikipedia, Actes du 31ème colloque Inforsid, Paris, 30 mai 2013.

[15] Viseur, R. 2013b. Extraction of Biographical Data from Wikipedia, Proceedings of International Conference on Data Technologies and Applications 2013, Iceland, July 29-31, 2013.

[16] Viseur, R. 2013c. Collecter des données sur Wikipédia : application à la création d'une base de données biographiques, Conférence Wikipédia, objet scientifique non identifié, ISCC (CNRS), 5 juin 2013.

[17] Wilkinson, D.M., Huberman, B.A. 2007, Cooperation and quality in wikipedia, Proceedings of the 2007 international symposium on Wikis, ACM, pp. 157-164.