QUALITY ASSESSMENT OF WIKIPEDIA EXTERNAL LINKS

Paraskevi Tzekou[*], Sofia Stamou[*+], Nikos Kirtsis[*] and Nikos Zotos[*]
[*]*Computer Engineering and Informatics Department, Patras University 26500 Patras, Greece*
[+]*Department of Archives and Library Science, Ionian University, 49100 Ionian, Greece*

Keywords: Wikipedia, External links decay, Measurements, Quality assessment.

Abstract: Wikipedia is a unique source of information that has been collectively supplied by thousands of people. Since its nascence in 2001, Wikipedia is continuously evolving and like most websites it is interconnected via hyperlinks to other web information sources. Wikipedia articles contain two types of links: internal and external. Internal links point to other Wikipedia articles, while external links point outside Wikipedia and normally they are not used in the body of the article. Although there exist specific guidelines about both the style and the purpose of the article external links, no approach has been recorded that tries to capture in a systematic manner the quality of Wikipedia external links. In this paper, we study the quality of Wikipedia external links by assessing the degree to which these conform to their intended purpose; that is to formulate a comprehensive list of accurate information sources about the article contents. For our study, we estimate the decay of Wikipedia external links and we investigate their distribution in the Wikipedia articles. Our measurements give perceptible evidence for the value of external links and may imply their corresponding articles' quality in a holistic Wikipedia evaluation.

1 INTRODUCTION

Wikipedia is a dominant source of online information for millions of people. To date, the English version of Wikipedia hosts over 3 million articles[1] and it is constantly evolving as new material is been provided by numerous editors who work on a volunteer, collaborative basis. The most fascinating thing about Wikipedia is its open nature, which enables people to create, modify and/or extend an article and which has turned Wikipedia into an unprecedented source of collectively-supplied information.

However, Wikipedia's open nature has raised scepticism about the quality of its hosting articles and has paved the ground for studies that investigate the automatic assessment of Wikipedia's quality. In this respect, numerous quality assessment methods have been proposed, most of which rely on the cross-examination of Wikipedia internal characteristics, e.g. the articles' length (Blumenstock, 2008a), their contextual elements (Stvilia et al., 2005a), the number of edits (Buriol et al., 2006), their linkage within the Wikipedia graph (Kamps and Koolen, 2009), the formality of their language (Emigh and Herring, 2005), their factual accuracy (Giles, 2005), etc. But Wikipedia, like almost any other website, provides links from its articles to external (i.e. outside Wikipedia) web pages. The scope of external links is to provide supplementary information about the article's content that is accurate and which could or should not be added to the article for reasons such as copyright or amount of detail[2]. Although external links are intended to provide complementary information about the articles' contents, they have not been systematically accounted in the evaluation of Wikipedia's quality.

In this paper, we investigate the role of Wikipedia external links in conveying accurate information about the article contents. Our investigation is not confined to external links appearing under a homonymous section at the very end of articles, but rather it addresses the role of all external links within Wikipedia articles. The motive for our study is to quantify the external links' usefulness and merit towards improving the articles' comprehensiveness and explore them as implicit signals of the articles' quality. Our investigation relies on the intuition that the quality of an article is determined based not only on the accuracy of the information that the article per se brings, but also on the usefulness and the accuracy of its pointing external resources.

[1]http://en.wikipedia.org/wiki/Special:Statistics

[2]http://en.wikipedia.org/wiki/Wikipedia:External_links

Tzekou P., Stamou S., Kirtsis N. and Zotos N..
QUALITY ASSESSMENT OF WIKIPEDIA EXTERNAL LINKS.
DOI: 10.5220/0003299502480254
In *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (WEBIST-2011), pages 248-254
ISBN: 978-989-8425-51-5

In the course of our study, we built upon earlier work (Kirtsis et al., 2010) in which we assessed the amount to which Wikipedia external links add new content to their corresponding articles' body and we make the below-listed contributions:

- We examine the statistical ***distribution*** of external links in Wikipedia articles, we identify popular domain names in the articles' external resources and we capture the correlation between the articles' length and the fraction of their external links. Our examination helps assess the appropriateness of the articles' external links, since according to Wikipedia editing guidelines, external links should point to information that is not yet part of the article.
- We quantify ***decay*** in Wikipedia external links in an attempt to assist editors and administrators determine the amount of repair that the articles' linked sources should undergo. To measure decay of an article's external pages, we use the proportion of dead links to which the article points. A dead link points to a page that either is no longer available or redirects to a page that has nothing to do with the original page. Measuring decay of Wikipedia linked-to resources indicates the appropriateness of the articles' linking pages in providing readers with reliable yet valid information about the subject of the article.

The remainder of the paper is organized as follows. We start our discussion with an overview on related work. In Section 3, we describe how external links can serve as implicit indicators of Wikipedia's quality. Specifically, we discuss how to capture and interpret the distribution of external links in Wikipedia articles and we address the problem of identifying external links' decay. In Section 4, we present an experiment we carried out in which we assessed the usefulness of external links for communicating valid and complete information about the Wikipedia article contents. The discussion of experimental results sheds light on the article external features to be considered in future Wikipedia quality assessment efforts. We conclude the paper in Section 5 where we outline our plans for future work.

2 RELATED WORK

Related work falls in two main categories, which we discuss in turn. First, there is related work in studying the nature and the quality of Wikipedia. In this respect, researchers have suggested a number of article features that signify quality, e.g. their articles' survival period (Cross, 2006), the number and frequency of their edits (Wilkinson and Huberman, 2007), their revision history (Adler and de Alfaro, 2007), the amount of outbound citations to scientific publications (Nielsen, 2007), the dedication and expertise of their editors (Riehle, 2005), etc. Our work extends prior studies that infer the Wikipedia articles' quality based on their contents investigation in that we also examine the impact of external links to their corresponding articles' quality. In our work, we significantly enrich our earlier study (Kirtsis et al., 2010) by examining additional features in assessing the article links' and by running large-scale evaluation experiments on the Wikipedia outlinks' distribution and properties.

In a similar direction, (Buriol et al., 2006) study the evolution of the Wikipedia link graph over time and observe an increasing link density in Wikipedia as time goes by. This is also attested in the work of (Kamps and Koolen, 2009), who contrast the link structure of the Web and Wikipedia in an attempt to assess the impact of global and local link topology in web retrieval effectiveness. Although our work relates to the above studies that examine the Wikipedia link structure, it is different in that we investigate the contribution of external resources in complementing the article contents. Most importantly, we estimate not only external links' distribution but also their decay. Although, the encyclopaedic organization of Wikipedia is different from the Web data organization in that it is steered by specific guidelines about what and how to link[3] still Wikipedia like all large websites suffers from the link rot phenomenon. As of November 2006 dump, nearly 10% of external links were broken and although repairing and maintenance actions are been taken ever since, the problem of rotten external links persists.

Thus, our study also relates to existing works that focus on the identification of dead links. The majority of existing works concentrate on the web information sources' decay and examine millions of web pages for deriving statistics about the fraction of dead web links (Fetterly, et al., 2003; Morishima et al., 2008). Another signal of web decay, apart from dead links, is soft-404 server errors, which imply that a URL redirects to a page that returns an OK HTTP code, but contains a totally different content from the one requested (Bar-Yossef et al., 2004; Lee et al., 2009). Moreover, researchers attested that decayed pages are characterized by outdated or else abandoned content. A number of studies have been reported that try to capture abandoned pages based on the estimation of their links and content age (Jatowt et al., 2007; Popitsch and Haslhofer, 2010). Our

[3]http://en.wikipedia.org/wiki/Wikipedia:Linking.

study builds upon existing works on links' decay but instead of investigating the entire web's link rot, we concentrate on the links pointed-by the Wikipedia articles. This is because we seek to understand the value of Wikipedia external links for their corresponding articles and assess the effectiveness of Wikipedia maintenance policies in delivering valid, reliable and longevous content.

3 WIKIPEDIA LINKS' QUALITY

One thing about Wikipedia that has not been thoroughly examined in the course of its quality assessment is the impact of Wikipedia external links in determining their corresponding articles' value. The lack of such examination may be because Wikipedia has set specific linking guidelines, which, if respected, might assure the additive value of external links to the article contents. In brief, Wikipedia linking instructions request that external links are placed in identifiable sections under a primary heading (i.e. external links, citations, etc.) at the end of the article's body to identify sources outside Wikipedia that are directly relevant to the article that points them. Moreover, Wikipedia requests that external links to an article are helpful to the reader, are (and likely to remain) functional, provide consistent information, express a neutral stance for their contents and are kept minimal. The quest in establishing linking standards is to ensure that Wikipedia refers the reader to credible and accurate supplemental material, which would distinguish Wikipedia as an encyclopaedia from any other online resource.

However, the phenomenal growth of Wikipedia and the absence of a formal peer review process for its contents (Denning et al., 2005) make it extremely difficult for both editors and administrators to validate external links on a regular and coordinated basis. This, coupled with the insufficiency of existing link templates and policies to ensure proper external linking in all situations, impose the imperative need to automatically validate the contribution of external links to their respective articles' comprehensiveness and merit. In the following paragraphs, we examine a number of features in Wikipedia external links in order to capture how these adhere to the foreseen linking guidelines. In Section 3.1 we investigate the statistical distribution of external links in Wikipedia collection; we examine how link distribution correlates to the articles' length and we identify popular domains in Wikipedia' external resources. In Section 3.2, we address the problem of quantifying decay in Wikipedia external links, in an attempt to assist administrators determine effective maintenance and update Wikipedia policies.

3.1 External Links' Distribution

Identifying Wikipedia external links is fairly easy as these are placed in distinct sections at the end of the articles' body, released in the Wikipedia dump. Therefore, we explore the Wikipedia dump to detect and quantify the links of every article to non-wiki sources. Nevertheless, assessing the value of external links to their article contents is less straightforward as we would need to capture their appropriateness and their correlation to their linked-to articles.

To address that, we firstly need to identify the non-wiki sources to which every article points. Such identification is somewhat complex because there are different ways of linking to external resources. To tackle difficulties, we firstly merge all identified external links into a common file, we remove the space and/or bullet characters that appear at the beginning of every non-wiki link and then we organize them in three main clusters according to their syntax. The emerging clusters group together links provided in the following forms: (i) URLs placed in square brackets followed by a hyperlink with a serial number as its label, (ii) hyperlinked URL names and (iii) URLs placed in square brackets followed by a space and some text used to label the hyperlink.

Having extracted and clustered external links, we process the contents of every cluster in order to get the URL of each non-wiki linked source. We then parse the URLs of the external resources in order to identify their domain names and estimate their distribution in the Wikipedia collection. For computing the distribution of URL domain names in Wikipedia, we simply count the number of articles containing an external link to some URL listed under each of the identified domains. The last thing we examine with respect to Wikipedia external resources is the correlation between the articles' length and the amount of their external links. For estimating each article's length, we parse the XML file of the Wikipedia dump in order to extract for every file entry (i.e. article) its pure text. Then, based on the pure text of every article, we quantify the article's length by counting the number of characters it contains. Thereafter, we represent every article as a vector of paired numerical values identifying the article's length and the amount of its external resources. The computed vectors, when plotted, help us visualize the correlation between the article's textual size and its non-wiki linked resources.

In section 4.1, we report the statistics of our es-

timations, which serve as preliminary evidence about the degree to which Wikipedia external links comply with their editing guidelines and help us infer the quality of Wikipedia external links.

3.2 External Links' Decay

To estimate decay in Wikipedia external links, we measure the fraction of dead pages among the non-wiki linked sources. A dead page is a page that is not publicly available over the web. Determining if a page is dead is easy to detect as it will fail either the URL parsing or the resolution of its host address and the attempt to fetch it will return an error HTTP code response[4]. On the other hand, a page is alive if it can be successfully fetched and returns an HTTP code in the 2xx series or 4xx series, except for 403, 404 and 410. But, there exist cases where discriminating existent (alive) from non-existent (dead) pages is tedious. Such cases concern redirect pages that cannot be identified as such since their HTTP return code is not in the 3xx series. Instead their landing pages return either an error 404 HTTP code or an OK code. According to (Bar-Yossef et al., 2004) in the former case we can safely consider that the page is dead, but in the latter case there are two possibilities. The page is dead if one of the following applies: (i) the page redirects to its host's home page and a non-existing randomly generated URL in the same host directory also redirects to the host's home page, (ii) the page redirects to a URL, which in turn redirects and its fetching enters in a loop of redirects, or (iii) the page and a non-existing randomly generated URL redirect to pages of nearly-identical content, determined via shingling (Broder et al., 1997). On the other hand, the page is alive if: (i) it is a root of a website, or (ii) it redirects to a URL and a non-existing randomly generated URL in the same host directory returns a failure HTTP code. The above situations and their interpretation have been encapsulated into an elegant algorithm for identifying dead pages, introduced in (Bar-Yossef et al., 2004). In our study, we implemented the above algorithm and run it against the extracted Wikipedia external links. The results delivered by the algorithm, presented and discussed in Section 4.2, might assist Wikipedia editors take remedial actions with respect to out-link updates.

[4]For a list of error codes see: http://en.wikipedia.org/wiki/Wikipedia:Dead_ external_links

4 EXPERIMENTS AND RESULTS

To capture the contribution of Wikipedia external links to their respective articles' contents, we relied on the October 2009 dump of the English-language Wikipedia, a 24.5 GB XML document, which we processed in order to extract the pure text of every article and the URLs of its external resources. Table 1 summarizes the statistics of our dataset. As the Table shows, out of the 3,290,179 articles in the English Wikipedia 786,857 (i.e. 23.91%) provide no links to external resources, whereas the remaining 2,503,322 articles contain 13,355,687 external links in total; with the ratio of article/ outbound link mean distribution reaching to 4.06 (median 2). Based on the articles' identified external links, we downloaded their pages via a simple crawler module we built and we processed them as previously described. Then, we applied our measurements to the article external links in order to quantify their distribution and decay in the Wikipedia collection.

Table 1: Statistics of Wikipedia collection.

Collection ID	October 2009 Wikipedia dump
Number of articles	3,290,179
Number of external links	13,355,687
External links/ article	4.06
Articles without external links	786,857

4.1 Distribution of External Links

Figure1 plots the distribution of external links to Wikipedia articles.

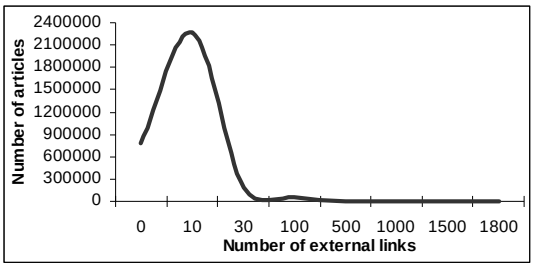

Figure 1: External links' distribution in Wikipedia articles.

As the figure shows this is a normal distribution, indicating that there are some Wikipedia articles with very many external links and some articles with very few links but the amount of external links that the majority of articles contain is close to the mean.

In particular, results indicate that 7.30% of the articles point to more than 100 non-wiki sources, 23.91% contain no out-links at all and 68.79% of the articles contain between 1 and 10 external links.

Based on the Wikipedia linking guidelines, which advise editors to give links to non-wiki sources when these contain useful material for the article subjects that is not yet added to the articles' body, we may interpret our finding as follows.

A small fraction of Wikipedia articles (i.e. those with many external links) provide incomplete information about their subjects, thus they point the reader to many external resources for additional information. Moreover, a small fraction of articles (i.e. those without or with very few external links) provide complete information about their subjects, therefore they either need no additional information or the non-wiki information that complements them is minimal. Finally, most of the articles contain adequate information about their subjects, which can be enriched by the contents of some external resources. But to check the validity of this interpretation, we need to investigate the correlation between the articles' length and the amount of their external links. This in order to attest whether a large amount of external links signifies incomplete article contents or not. Figure 2 plots the correlation between the articles' length and the average number of their external linked-to sources.

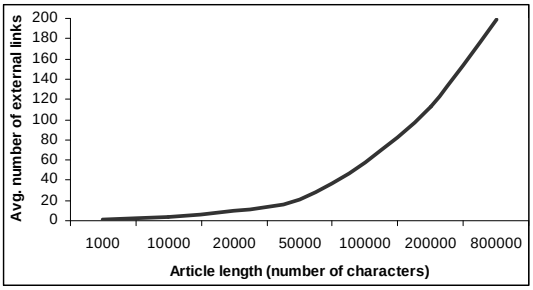

Figure 2: Correlation between article length and amount of external links.

According to the figure, there is an upward trend of external links over long articles, which peaks at 200 links on average for length articles, i.e. those containing nearly 800K characters and generally takes values between 3 and 8 links for articles containing on average 10K characters (note that 70.6% of the articles are 10K characters long). The peak in the quantity of external links is implicitly imposed by Wikipedia, which upon the detection of long link lists invokes a warning to editors and suggests alternative ways of linking. But still, the fact that long articles contain more links than short ones connotes a contradiction between the instructions given in Wikipedia linking guidelines and the way in which these are adopted in practice. In particular, Wikipedia editors are instructed to point to external resources if their content is proper in the article's context and is not yet part of the article. Thus, one would expect lengthy articles to contain fewer links than short ones in the sense that the more text the article contains the decreased the need to link it with supplemental non-wiki material. But, this is not observed in our results. Next, we report the distribution of link domain names across the Wikipedia collection. Such distribution will help us infer the appropriateness of Wikipedia external links as additional information sources, since according to linking guidelines editors should avoid separate links to multiple pages in the same website and link to the site directly.

Figure 3 depicts how domain names span to Wikipedia articles. Specifically, the *x*-axis represents the 50 most popular domain names within the Wikipedia external resources, where popularity is determined based on the amount of Wikipedia pointed URLs listed under each domain. The *y*-axis illustrates the amount of Wikipedia pointed URLs listed each of the examined domain names.

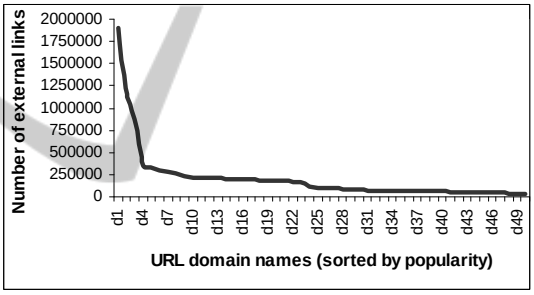

Figure 3: Correlation between article length and external links.

As the figure shows, this is a power-law distribution, indicating that only a few domains host a large fraction of the Wikipedia external links, while the majority of domains host only a small amount of the Wikipedia external links. In particular, we observe that the most popular domain alone hosts 14.29% of the Wikipedia external links, while 44.87% of all out-links are listed in only 10 different domain names. Considering that in-linked Wikipedia articles exhibit strong topical correlation (Koolen and Kamps, 2009) our finding make sense since URLs in a domain pointed by many articles of related contents give rise to that domain's popularity inside Wikipedia.

4.2 Decay of External Links

To compute the fraction of dead links in Wikipedia external resources, we run the dead link detection

algorithm against the identified non-wiki links of nearly 2,000,000 articles and estimated the amount of links that point to dead pages. For running the algorithm, we specified the following parameters: (i) we set a page fetching timeout to 10 seconds; so that if there is no server response within that time frame the page is declared dead, and (ii) we set the maximum number of redirects to 5; so that if more than 5 redirects are encountered the page is declared dead. Moreover, for generating random URLs in the host directory of the page the algorithm is trying to fetch, we adopted the method proposed in (Bar-Yossef et al., 2004), which appends to the host's URL a sequence of 25 randomly combined lower case Latin letters. Based on the above parameters, we run the algorithm against 4,575,154 linked external URLs that span to 1,939,445 Wikipedia articles and computed the fraction of external links that point to either dead (non-existing) or decayed (redirect to irrelevant content) pages. In total, we found 839,245 (roughly 18.34%) of the examined links be dead, i.e. returned hard or soft-404 codes. Figure 4 plots the distribution of decayed links in the examined Wikipedia articles. The *x*-axis reports the amount of articles that contain as many external links as their corresponding *y*-axis values.

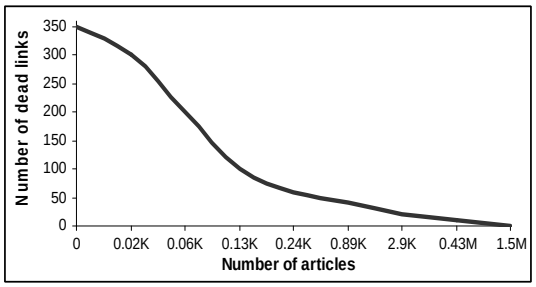

Figure 4: Distribution of dead links in Wikipedia articles.

According to the figure, the distribution of dead links follows a power-law with very few articles containing a considerable amount of dead links and the striking majority of the articles (i.e., 77.31%) containing no dead links at all. Results suggest that the majority of the web pages pointed by Wikipedia are reachable, thus implying that Wikipedia administrators have quite effective mechanisms for filtering rotten links and that Wikipedia editors carefully pick resources to link their articles.

5 CONCLUDING REMARKS

In this paper, we have examined the quality of Wikipedia external links in an attempt to infer how these contribute to the improvement of their linked article contents. For our examination, we quantified the distribution of external links in Wikipedia collection and we estimated the amount of external links' decay. Our estimations suggest that if Wikipedia is equipped with improved link control and filtering mechanisms, it will combat link rot and improve the overall quality of its external links and implicitly of its article contents.

REFERENCES

Adler, N. T., de Alfaro, L. 2007. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th International World Wide Web Conference*, pp. 261-270.

Bar-Yossef, Z., Broder, A. Z., Kumar, R., Tomkins, A. 2004. Sic transit gloria telae: towards an understanding of the web's decay. In *Proceedings of the 13th International World Wide Web Conference*, pp. 328-337.

Blumenstock, J. E. 2008(a). Automatically Assessing the Quality of Wikipedia Articles. *UCBiSchool Report* 021.

Blumenstock, J. E. 2008(b). Size matters: Word count as a measure of quality on Wikipedia. In *Proceedings of the 17th International World Wide Web Conference*, pp. 1095-1096.

Broder, A. Z., Glassman, S. C., Manasse, M. S., Zweig, G. 1997. Syntactic clustering of the web. In *Proceedings of the 6th International World Wide Web Conference*, pp. 391-404.

Buriol, J., Castillo, C., Donato, D., Leonardi, S., Millozzi, S. 2006. Temporal evolution of the wikigraph. In *Proceedings of the IEEE Web Intelligence Conference*, pp. 45-51.

Cross, T. 2006. Puppy smoothies: improving the reliability of open, collaborative wikis. First Monday 11(9).

Denning, P., Horning, J., Parnas, D., Weinstein, L. 2005. Wikipedia risks. *Communications of the ACM*, vol.48, no.12.

Emigh, W., Herring, S. 2005. Collaborative authoring on the Web. In *Proceedings of the 38th Hawaii Intl. Conference on System Sciences*.

Fetterly, D., Manasee, M., Najork, M., Wiener, J.L. 2003. A large-scale study of the evolution of web pages. In *Proceedings of the 12th International World Wide Web Conference*, pp. 669-678.

Giles, J. 2005. Internet encyclopedias go head to head. Nature, 438, pp. 900-901.

Jatowt, A., Kawai, Y., Tanaka, K. 2007. Detecting age of page content. In *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management*, pp. 137-144.

Kamps, J., Koolen M. 2009. Is Wikipedia link structure different? In *Proceedings of the 2nd International Conference on Web Search and Data Mining*, pp. 232-241.

Kirtsis N., Stamou S., Tzekou P., Zotos N. 2010. Information Uniqueness in Wikipedia. In *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WebIST)*, Valencia, Spain.

Koolen, M., Kamps, J. 2009. What's in a link? Form document importance to topical relevance. In *Proceedings of the 2nd International Conference on the Theory of Information Retrieval*, pp. 313-321.

Lee, T., Kim, J., Kim, J. W., Kim, R. S., Park, K. 2009. Detecting soft errors by redirection classification. In *Proceedings of the 18th International Web Conference*, pp. 1119-1120.

Mrishima, A., Nakamizo, A., Iida, T., Sugimoto, S., Kitagawa, H. 2008. PageCasher: A tool for the automatic correction of broken web links. In *Proceedings of the 24th International IEEE Conference on Data Engineering*, pp. 1486-1488.

Nielsen, F. A. 2007. Scientific citations in Wikipedia. Computing Research Repository.

Popitsch, N. P., Haslhofer, B. 2010. DSNotify: Handing Broken Links in the Web of Data. In *Proceedings of the World Wide Web Conference*.

Riehle, D. 2005. How and why Wikipedia works: an interview. In *Proceedings of the Intl. Symposium on Wikis*, pp. 3-8.

Stvilia, B., Twidale, M. B., Smith, L. C., Gasser, L. 2005(a). Assessing information quality of a community-based encyclopedia. In *Proceedings of the International Conference on Information Quality*, pp. 442-454.

Stvilia, B., Twidale, M. B., Gasser, L., Smith, L. C. 2005.(b) Information quality discussions in Wikipedia. In *Proceedings of the International Conference on Knowledge Management*.

Wilkinson, D. M., Huberman, B. A. 2007. Assessing the value of cooperation in Wikipedia. First Monday 12(4).