RENDER
FP7-ICT-2009-5
Contract no.: 257790
www.render-project.eu

RENDER

Deliverable D5.1.1

Definition and Evaluation of Metrics in Wikipedia

Editor:	Angelika Adam, Wikimedia
Author(s):	Angelika Adam, Wikimedia; Fabian Flöck, Karlsruher Institut für Technologie; Gerrit Holz, Wikimedia
Deliverable Nature:	Report (R)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	September 2011
Actual Delivery Date:	September 2011
Suggested Readers:	Researchers and professionals
Version:	1.2
Keywords:	Wikipedia, Diversity, metrics, editor behaviour, timeliness, completeness, objectivity, editing patterns

Disclaimer

Full Project Title:	RENDER – Reflecting Knowledge Diversity
Short Project Title:	RENDER
Number and Title of Work package:	WP5 Diversity case studies
Document Title:	D5.1.1 - Definition and Evaluation of Metrics in Wikipedia
Editor (Name, Affiliation)	Angelika Adam, Wikimedia
Work package Leader (Name, affiliation)	David Cadenas Sánchez, Telefónica Investigación y Desarrollo

Executive Summary

In this deliverable we defined the metrics of the Wikipedia case study. These are mainly motivated by the criteria for high-quality articles, which the Wikipedia community defined and imposed them-self. The main goal of the Wikipedia case study - to increase the quality of Wikipedia by supporting its users - fits very well with the aims of Wikimedia Deutschland e.V.

We defined three use case scenarios where the RENDER project could help Wikipedia's editors and readership.

The main questions in this context are:

- Does an article cover all important facts of its topic compared to related articles in other language versions or compared to external sources like news stream extractions?
- Does an article represent the most recent information about that topic, or are there new events in the world, which are not included in the article content yet?
- Could a reader feel motivated to contribute in Wikipedia, if he knows which facts are missing and where he could find them?

We provide metrics concerning completeness, currency, objectivity, and editor behaviour as the most important and appropriate ones to find, to understand, and to improve the lack of knowledge diversity in Wikipedia articles and to achieve an increase of quality.

Table of Contents

List of figures

List of tables

Abbreviations

ACM	Associating for Computing Machinery
MW	MediaWiki
NPOV	Neutral Point of View
POV	Point of View
SIRD	Simple Identity Revert Detection

1 Introduction

In this deliverable we describe the metrics for the Wikipedia case study in more detail. We will discuss the characteristics of diversity, a necessary criterion for quality in Wikipedia, and the quality assurance mechanisms in Wikipedia. In addition, we define the benefits and aims of the RENDER project for Wikipedia - its community and readership - more clearly. On this foundation and our additional analysis of previous research on diversity in Wikipedia we will introduce metrics to measure different aspects of diversity.

1.1 Diversity in Wikipedia

A Wikipedia article is usually written by multiple editors. Naturally, editors are biased towards a certain point of view: their own. This is especially evident in articles tagged as “controversial”. This is often related to articles of research fields with little or no measurably or options for falsification. Any encyclopaedic coverage of those topics, including but not limited to political science and theology, will therefore have to deal with competing, conflicting and unresolvable points of view on the very nature of a subject.

Still, the real problem is those biased articles that go untagged and where the bias remains unnoticed by the reader. This is a major violation of Wikipedia’s goal to let its articles represent the so called Neutral Point Of View[1], which ”means representing fairly, proportionately, and as far as possible without bias, all significant views that have been published by reliable sources”. Therefore it is necessary that either the editors can transcend their personal point of view, or that a multitude of editors with diverse opinions about the topic covers the significant points of view. This is not always the case. An encyclopaedic project operating under the pillar of Neutral Point of View style guides will have to give appropriate representation to all relevant points of view and it can be considered a strong indicator for high quality content at Wikipedia if the diversity of opinions in research is well represented.

1.2 Organisation of this deliverable

In the following section we will give a short overview of quality assurance processes which are established by the Wikipedia community. These methods concern different aspects of quality and outline the importance of diversity to reach high-quality encyclopaedic content in Wikipedia articles. Section 3 summarizes some research related to quality measuring and editor behaviour. Section 4 introduces the metrics we identified as most important for the Wikipedia case study. The final section gives a short summary of this deliverable and a brief view in our future work.

[1] http://en.wikipedia.org/w/index.php?title=WP:Neutral_point_of_view&oldid=446756659 (accessed: 31.08.2011)

2 Background

In this chapter we will introduce the quality assurance processes in Wikipedia, our defined use case scenarios and the article feedback tool which enables Wikipedia users to assess the quality of an article.

2.1 Quality assurance processes in Wikipedia

The quality of Wikipedia is comparable to other traditional encyclopaedias despite the lack of traditional quality control mechanisms and the susceptibility to certain risks, such as vandalism, like numerous studies (e.g. [6]) confirm.

There are internal quality assurance mechanisms and rules the community has defined for contributing in Wikipedia. An editor can find support information e.g. on the community portal page[2].

Wikipedia defined the following main features to write high-quality articles[3]:

Neutral point of view: Content should be presented in a neutral point of view, thus also presenting competing views and theories.

Verifiability: Content must be verifiable by trustworthy sources.

No original research: Articles should contain only previously published knowledge. This means that Wikipedia is not an appropriate venue to present new research results.
The quality assurance process is based primarily on the users' notice. Every user has the option to comment, to tag the article with quality marks, and / or to improve an article, as described by Hammwöhner [8]. This open process is ensured by the fact that each version of an article can be reverted easily if a change leads to a deterioration of quality. Errors and vandalism are quickly corrected, as Stvilia et al. [12] demonstrated.

A qualitatively very good article on Wikipedia can get a “Featured Article” or “Good Article” status. These awards provide an incentive to create high-quality articles and can give users orientation.
An article receives such a status by a voting process after it was proposed by any user. For this judgement an article has to satisfy a set of certain criteria for Good articles[4] and Featured articles[5]. These concern different aspects of an article like accuracy, neutrality, completeness, and style. This status is not permanent and can be withdrawn, if there is a deteriorating during the development of this article.
A similar process allows for the deletion of a particularly bad article, for example, violation of central rules. The execution of deletions can be done only by administrators, users with special privileges. The quality assessment process in Wikipedia via nominating for honouring or deletion of an article is visualised in Figure 1.

[2] http://en.wikipedia.org/w/index.php?title=WP:Community_portal&oldid=442540503 (accessed: 30.08.2011)
[3] http://en.wikipedia.org/w/index.php?title=Wikipedia:SR&oldid=441963781 (accessed: 29.08.2011)
[4] http://en.wikipedia.org/w/index.php?title=WP:Good_article_criteria&oldid=437412173 (accessed: 31.08.2011)
[5] http://en.wikipedia.org/w/index.php?title=WP:Featured_article_criteria&oldid=442215638 (accessed: 31.08.2011)

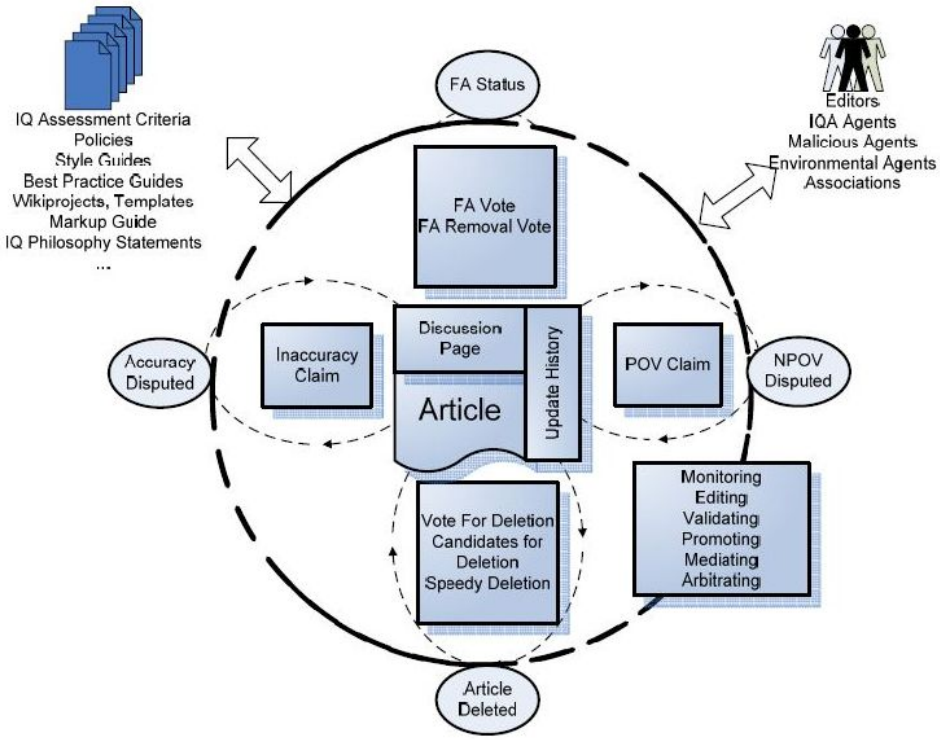


Figure 1 - Wikipedia quality assurance processes

(FA = Featured Article; IQA = information quality assurance; NPOV = neutral point of view; POV = point of view)[6]

Other instruments of quality assurance in Wikipedia are special structures like portals or WikiProjects. These content portals group thematically related articles and are often supervised by a dedicated group of users. In many cases they use additional quality mechanisms. For example, the WikiProject biology, which is named Redaktion:Biologie in the German Wikipedia[7], distinguishes validated articles in addition to the other quality levels. Furthermore, the active participants of this WikiProject are working with mostly by scripts automatically generated lists[8] to improve the quality of this topical field. These lists contain articles with certain unacceptable defects (e.g. missing references) and which thereby require a revision.

As mentioned before, every user is allowed to mark an article with a special message template[9]. The authors of [1] found 333 of these templates and identified 70 as "information quality flaws". Biased articles can be found by searching for articles which are tagged with neutrality templates[10]. This can be useful to generate a corpus of biased material for testing metrics. In addition, there are several templates, which can indicate the need for editing on a certain article for many other reasons, e.g. Template:Unreferenced[11]. This template can be added, if an article cites zero sources.

[6] Source of the figure: Stvilia et al. [12], p. 11
[7] http://de.wikipedia.org/w/index.php?title=Portal:Biologie&oldid=91953917 (accessed: 31.08.2011)
[8] http://de.wikipedia.org/w/index.php?oldid=87443863 (accessed: 30.08.2011)
[9] http://en.wikipedia.org/w/index.php?title=WP:Template_messages&oldid=439995257 (accessed: 31.08.2011)
[10] http://en.wikipedia.org/w/index.php?title=Category:Neutrality_templates&oldid=435104130 (accessed: 31.08.2011)
[11] http://en.wikipedia.org/w/index.php?title=Template:Unreferenced&oldid=452031726 (accessed: 23.09.2011)

2.2 Use Case Scenarios for Wikipedia

It is the main goal of the Wikipedia Case Study in RENDER to increase the quality of Wikipedia content and the confidence in Wikipedia by finding, understanding and curing lacks of knowledge diversity in its articles. As we showed above, there are control mechanisms to assure the quality in Wikipedia, like the criteria for high-quality and message templates. But due to Wikipedia's size many biased articles are not identified, as [1] already mentioned.

Our main motivation is supporting the Wikipedia community and to react on their requirements to improve respectively alleviate their work. On the other side, there are the users of Wikipedia. As our experiences showed, many of them did not notice the collaborative and enduring working process and the possibility of being part of that. We expect to sensibilise them by visualising assessment scores and the specifics of the editor network in the background of a certain article. So first, we identified the following possible use case scenarios (UCS), which we expect to succeed during the project period:

UCS 1: Display warnings to the reader when detecting bias

UCS 2: Notify authors that an article needs to be updated

UCS 3: Lower the barrier for readers to extend and/or correct articles

Within this case study we want to display the Wikipedia users a warning whenever we detect bias in an article or an evidence for additional information related to the topic of an article which is not included. To reach this aims we plan to compare Wikipedia articles in different language versions and with related articles in external sources, like the news. Thereby we will be able to offer sources where to find missing information. A bias could be also caused by a unilateral presentation of a certain part of a topic or by the editorial interaction and collaboration. By showing the network interaction and understand the behaviour of authors we will be able to visualise users the influence of these parameters on an article's quality.

For our second use case scenario we will analyse the editing process of a Wikipedia article in different language versions. Additional, we plan to analyse external source related to an article to detect new information which are not yet included in an article. If we find evidence that a certain Wikipedia article seems to be out of date, we will signal the need for update to the users.

With help of our research and development in this case study, we hope to find ways to lower the barriers for readers to start their first steps in Wikipedia and to contribute. If a user gets the chance to understand which facts are missing in the article and get additional information where she can find further information, she will be enabled to improve that article. Maybe this situation could lower the threshold for starting editing in Wikipedia, too.

2.3 The article feedback tool

In this subsection, we describe briefly a quality assessment tool and the metrics, which are evaluated as most important to judge the quality of an article. The feedback tool [12] is a project, which was initiated by the Wikimedia Foundation to enable Wikipedia readers to assess the quality of Wikipedia articles.

[12] http://www.mediawiki.org/w/index.php?title=Article_feedback&oldid=424702 (accessed: 31.08.2011)

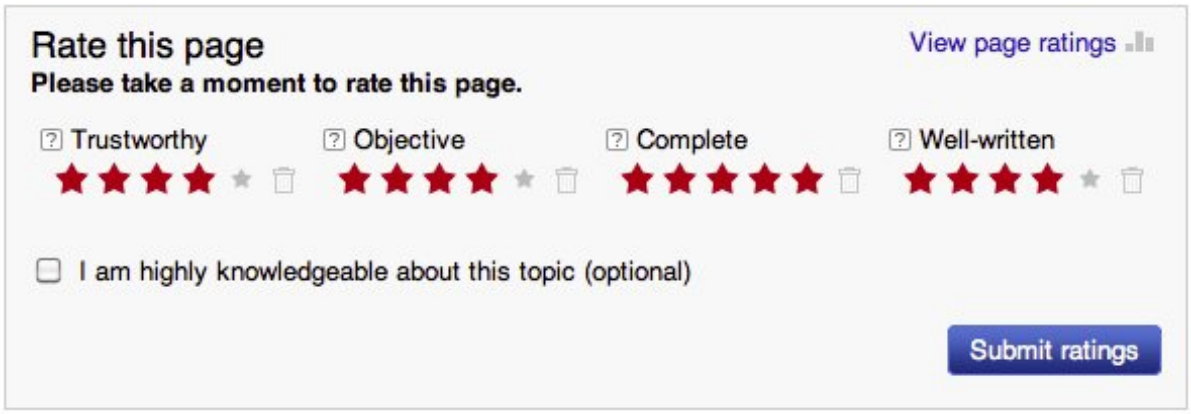


Figure 2 - Article feedback rating[13]

Figure 2 shows the article feedback form, which was included in the footer of all articles in the English Wikipedia since July 2011. With help of this tool users can assess the following metrics on a scale of 1 to 5 stars[14]:

- **Trustworthy** - "Do you feel this page has sufficient citations and that those citations come from trustworthy sources?"
- **Objective** - "Do you feel that this page shows a fair representation of all perspectives on the issue?"
- **Complete** - "Do you feel that this page covers the essential topic areas that it should?"
- **Well-written** - "Do you feel that this page is well-organized and well-written?"

The overall page rating is calculated based on the arithmetic average of all ratings. Ratings lose their relevance after 30 reverts of the article[15].

These rating data are available for downloading and analysing. We plan to use the data base - especially the results of the first three metrics - as Gold standard to evaluate our metrics.

[13] Source: http://commons.wikimedia.org/wiki/File:Aft_phase_2.jpg
[14] The description for each metric and the rating scale values (number of stars) is visible by clicking on the question mark next to the metric term respectively the star icons.
[15] http://www.mediawiki.org/w/index.php?title=Article_feedback/FAQ&oldid=427951 (accessed: 31.08.2011)

3 Related work - Research on Knowledge Diversity, Bias and Wikipedia

In this section we will give a short overview of research on Wikipedia's quality assessment of articles, editor behaviour and interaction.

3.1 Quality and diversity

Diversity is, as mentioned above, a necessity for quality in Wikipedia. Several studies have shown that Wikipedia is comparable to traditional encyclopaedias concerning quality. The most often cited, but also strong criticized study, is the Nature study [6]. This compares the Wikipedia with the *Encyclopedia Britannica* related to factual accuracy.

Hammwöhner et al. [9] compared the German Wikipedia with the printed version of the Brockhaus encyclopaedia. They extracted 50 lemmas randomly from each encyclopaedia, which were also part of the other one. The authors compared these related articles according to length and completeness of the articles, quantity and quality of the included sources and formal correctness. They found evidence for a significant improvement of Wikipedia with respect to the thematic coverage and the quality of informational reliability compared to the results of a previous study executed by Schlieker [24]. The articles of Wikipedia are more extensive and detailed compared to the Brockhaus.

A wide range of studies examined different aspects of quality in Wikipedia:

Lih [10] analysed the quality with two simple quality metrics, rigor (the number of edits for an article) and diversity (the number of unique editors). Additionally, he observed an influence of media attention on the article quality by an increasing of editing after an article was cited in the press. Brändle [4] also mentioned that the relevance of a topic and the attention of the users are the most important requirements for quality in Wikipedia-articles.

Wilkinson & Huberman [14] examined the quality according to the number of edits and editors. As an additional parameter they used the popularity of an article measured by the Google page rank. They found out that high-quality articles can be distinguished from other articles by a larger number of edits and editors and high cooperation patterns – which was measured by the revision number of the corresponding talk page of an article. A high degree of visibility or relevance of an article's topic leads to a larger number of edits, while the huge majority of articles get far less editing activity.

In a further study Hammwöhner [8] investigated Wikipedia's quality by choosing a very restricted topic – Shakespeare's work. In the first part of this research they examined which Wikipedia language versions contain articles about this topic. In 44 Wikipedias there was at least one article on Shakespeare's poetry. Not surprising was the fact, that the English Wikipedia included all. In a second step the authors analysed the completeness of these Wikipedia articles on a local article level. For this comparison they choose the English and the German articles. The decisions about the completeness of the content were made manually. Each article was assessed by collecting points for containing specific expected facts. The English Wikipedia reached slightly better results than the German language version. Hammwöhner concluded in his work that Wikipedia offers trustworthy information beyond the Featured Articles. Between the language versions of Wikipedia there are quantitative differences, which could be explained by their varying development status.

Halavais and Lackaff [7] analysed the quality related to topical scope and coverage. First, they examined the degree of Wikipedia's diversity of content by analysing a randomly selected set of English Wikipedia-articles in comparison to Bowkers Books in Print. The analysis is related to the number of articles on a specific topic, the number of edits and the average size of articles in each category. In a second study the authors analysed articles within three academic thematically fields- linguistics, poetry and physics – compared to printed scholarly encyclopaedias. Thereby, they compared the coverage of article titles or headwords in both directions. In sum they concluded, that "the degree to which Wikipedia is lacking depends heavily on one's perspective". Furthermore, they mentioned that the editing and developing process depends strongly on the interest of contributors.

Blumenstock [3] found out that the number of words to measure the length of an article is a very simple but also pretty "good predicator of whether an article will be featured on Wikipedia".

Computational information quality metrics for Wikipedia articles - Authority/Reputation, Completeness, Complexity, Informativeness, Consistency, Currency, and Volatility – were defined and analysed by Stvilia et al. [13].

The authors examined the completeness metric with the following equation:

*Completeness = 0.4 * Num. Internal Broken Links + 0.4 * Num. Internal Links + 0.2 * Article Length*

This definition does not fit with our definition of completeness, which is related to factual and topical coverage of an article. The currency – the time, when an article was updated the last time – was analysed by computing the time between the dump date and the date of the last update of the article in days. We will follow a more detailed approach to analyse this metric for Wikipedia.

As a result of their analysis based on article features and edit history metadata, the authors showed that these automatic computable metrics can successful used to discriminate high quality articles. Furthermore they concluded that these metrics are particularly suitable to assess "Wikipedia article quality along the dimensions corresponding to the set of IQ criteria (requirements) adopted by the community".

Arazy and Nov [2] analysed the influence of local and global editing activity on the article quality. As local activity they defined the activity of a set of editors in a specific article. The activity of these authors in the whole Wikipedia was named with global activity. The authors investigated the relationship of local ("inequality of editors' contribution in a particular article") and global ("inequality in overall Wikipedia activity levels for the same set of editors") inequality, coordination and article quality for the English Wikipedia. For testing their model, they used 50 Wikipedia articles, which were manually assessed on a 5-point Likert scale regarding the metrics completeness and accuracy. The authors concluded for their analysis that the communication and coordination have significant positive influence on article quality. The local inequality as well as the global one has a positive impact on the article quality.

Several further studies engaged with the quantitative analysis of Wikipedia. Ortega [11] presented in his extensive work a tool - WikiXRay[16] - to recover information from Wikipedia databases, and to use these information for a automatically quantitative analysis of this information. This tool is continuously improved and developed and has several analysing capabilities to compute and visualise Wikipedia's development. Its features are for example general statistics about e.g. active editors, active articles or talk pages; distribution of effort among editors like inequality analysis or model for editor's activity. We could use WikiXRay to visualise the development of Wikipedia with respect to article size or number of active editors during our case study.

[16] http://felipeortega.net/WikiXRay

3.2 Editor behaviour and interaction leading to bias

The paper “Towards a diversity-minded Wikipedia” which was presented at ACM WebScience 2011 Conference in Koblenz, Germany (see Appendix A) provides a survey of the most salient socio-technical mechanisms that can lead to the exclusion of the viewpoints of certain editors in Wikipedia and consequently, biased articles. They were identified analysing empirical research papers on Wikipedia that discovered recurring behavioural patterns of editors, with most of the work originally aimed at looking for quality-related patterns. These socio-technical mechanisms we identified build the theoretical base for constructing variables and metrics out of the edit history of the Wikipedia articles to find those patterns.

3.3 Conclusion

Stvilia et al. [13] states: “the Wikipedia community takes issues of quality very seriously”. For the Wikipedia case study it is important to support the community in achieving their goals. The most important parameters for diversity in Wikipedia are factual completeness, timeliness, and objectivity. These will be introduced in the following section.

4 Content-related metrics

We mentioned the necessity of completeness, timeliness, and objectivity for high-quality articles in Wikipedia in the first section. In this section we outline some more details on the content-related metrics we plan to analyse in the Wikipedia case study. A deeper look at the behaviour and the interaction of Wikipedia editors will give us important information to understand the processes which lead to biased content will follow in the next section.

4.1 Completeness

As mentioned above, a Wikipedia-article should be complete. That means it should cover all relevant facts about a specific topic. If a user understands which facts are not part of the current article, but could be found in e.g. another language version or in external sources, he will be capable to expand the content.

We will compute the fact coverage of an article. JSI developed a fact extraction service Enrycher, which was detailed in D2.2.1 [5]. This tool offers a fact extraction service for English input texts. These results could be compared with a fact extraction of the news stream extracted according to the topic of an article.

An additional possibility to reach this goal is to deal with working results of the EU-project CoSyne[17], which concerns multilingual content synchronization via a machine translation approach. The CoSyne consortium will be able to offer extractions for German, Dutch, Italian and English, later also for Bulgarian and Turkish. Additional we will use these data to compare the results of the English Enrycher fact extraction.

To understand the development of Wikipedia in different language versions, we expand the term of completeness to an overall definition. A Wikipedia (in one language version) could be seen as biased caused by a lack of important topics or articles in a certain category. So, we will look at the global coverage of Wikipedia and compare the language versions, too.

We compute the following metrics to determine the completeness:

- The article length measured by the number of words compared to articles in other language versions
- The total number of articles which contain at least 30 % additional facts compared to other language versions: these facts are detected by analysing the topical fields of an article with help of the internal links and the categories of these links in a Wikipedia article
- The total number of articles which have a lack of facts compared to at least one external source like an news article

Summarization:

We measure the completeness of Wikipedia by comparing the number of articles and the topical coverage in different language versions. The number of lacking articles in Wikipedia will decrease by generating working lists and thereby making visible these thematic lacks.

We measure the fact coverage of an article compared to its other language versions and topical facts from the news. The number of articles with a lack of information is expected to significantly decrease by showing the need of additional facts identified in other sources (news or other Wikipedias) and thereby indirect requests to improve an article.

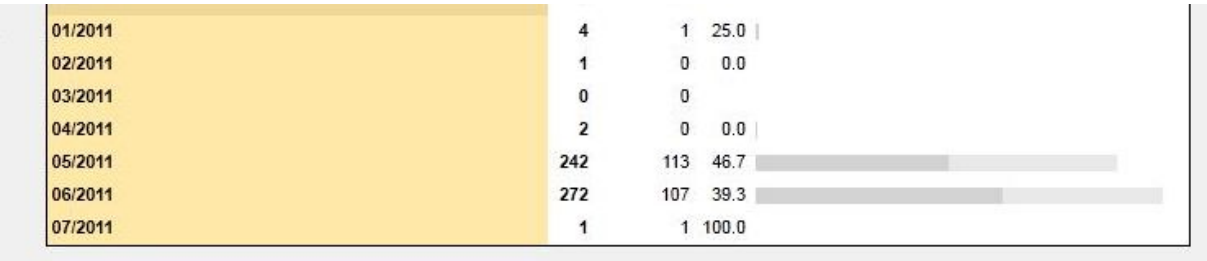
[17] Project website, www.cosyne.eu

4.2 Timeliness

The timeliness is a necessary criterion to ensure a complete coverage of information in a Wikipedia article. If a recent event is missing in the current version of an article about a topic, it is supposed to be incomplete and lacking certain facts about that topic. Following this definition, one could argue to subsume timeliness under completeness. But for our requirements and our defined use case scenarios (see section 2.2), we decided to handle this aspect as a unique metric.

We will use two ways to check the timeliness of an article. First, by checking the editing process in other language versions: If articles about the same topic in multiple other languages have been edited very frequently, this could be an indicator that this editing process is caused by an event in the world. We will compute the frequency of edits, which add content in Wikipedia articles and compare the results with the frequency in other languages about this topic. To detect abnormal editing behaviour it is important to observe the mean frequency of editing of the last days/weeks/months.

Figure 3 gives an example of the changes in the mean editing frequency of the German Wikipedia-article Enterohämorrhagische Escherichia coli between 01/2011 and 07/2011. The monthly mean editing increased extremely in May and June, caused by the many infection cases of the Hemolytic-uremic syndrome in Germany in this period.



01/2011	4	1	25.0
02/2011	1	0	0.0
03/2011	0	0	
04/2011	2	0	0.0
05/2011	242	113	46.7
06/2011	272	107	39.3
07/2011	1	1	100.0

Figure 3 - Extraction of the editing frequency for the DE-article Enterohämorrhagische Escherichia coli[18]

To determine the timeliness of Wikipedia, we compute the following metrics:

- Total number of edits per day compared to the mean daily editing of an article
- Total number of articles in Wikipedia, which are not reverted during the last day/week
- Total number of articles without content addition but high editing observation in at least five related language version articles

Our second approach will use external sources to check the timeliness of an article. We want to examine the following questions:

- Is the topic of an article a current topic in the news/the world?
- Are the new facts already part of the article about this topic?

We will observe news articles, which are related to a Wikipedia-article. For this step we will use the frequency computation of the occurrence of a certain topic in the news and compare the timestamps of Wikipedia edits and news article information. These data will be offered by JSI's news fact extraction service.

[18] Source: http://vs.aka-online.de/cgi-bin/wppagehiststat.pl

We calculate the following metric:

- The total number of articles which are out-dated in Wikipedia: The timestamp of the latest edit is at least five days older than the publishing date of a as related identified news article

Summarization:

We measure the timeliness of Wikipedia by analysing the editing processes and by comparing the timestamps of news publications with the edits in Wikipedia. To mark these articles will be particularly important for topics with a lower degree of visibility or relevance. By marking articles and generating working lists we expect to decrease significantly the number of out-dated articles in Wikipedia.

4.3 Objectivity

Following the Neutral Point Of View rule, a Wikipedia article should be written in an objective and balanced manner. Additionally, the content should be sourced by reliable references. (see section 2.1). So we analyse the objectivity of an article with two approaches:

- words or expressions, which contain subjectivity in this context
- references to an article, which seems to have a biased content

4.3.1 Opinionated words and expressions

Words or expressions, which introduce a certain kind of bias, should be avoided on Wikipedia[19]. For detecting subjectivity in Wikipedia articles, we want to determine if an article is opinionated/ non-opinionated by identifying words or word sequences which express sentiment.

We extracted Wikipedia-articles from the English Wikipedia between 01/2001 and 07/2011, which are tagged with the {{neutrality}} template (see section 2.1). For each marked article we extracted two revision versions of the article - the version at the time of the marking and the article version at the time of the deletion of the template. The latter required the marking to not be reset for the next seven days. JSI will use this data set as a trainings corpus for their machine learning algorithms.

So we will be able to compute the frequency of bias expression word or word sequences and to highlight them in an article. Additionally, we will be able to compute the frequency of as opinionated labelled articles and to observe the development over the time.

We compute the following metrics:

- The total number of articles containing subjective words or word sequences in Wikipedia
- The total number of articles, which are identified as opinionated by JSI's algorithm

4.3.2 References

References in Wikipedia change over time. Ideally, new entries presenting new information are being added; out-dated and low-quality sources are being deleted. In reality, it depends on the judgement of the editors. The content of an article changes, due to the facts and concepts presented in the references.

[19] http://en.wikipedia.org/w/index.php?title=WP:Manual_of_Style/Words_to_watch&oldid=444588981 (accessed: 31.08.2011)

We presume that availability and adoption of new information from external sources influences the content's diversity. A hypothesis might be that articles being based on mainly one source are likely to be more biased then articles being based on quite a few sources.

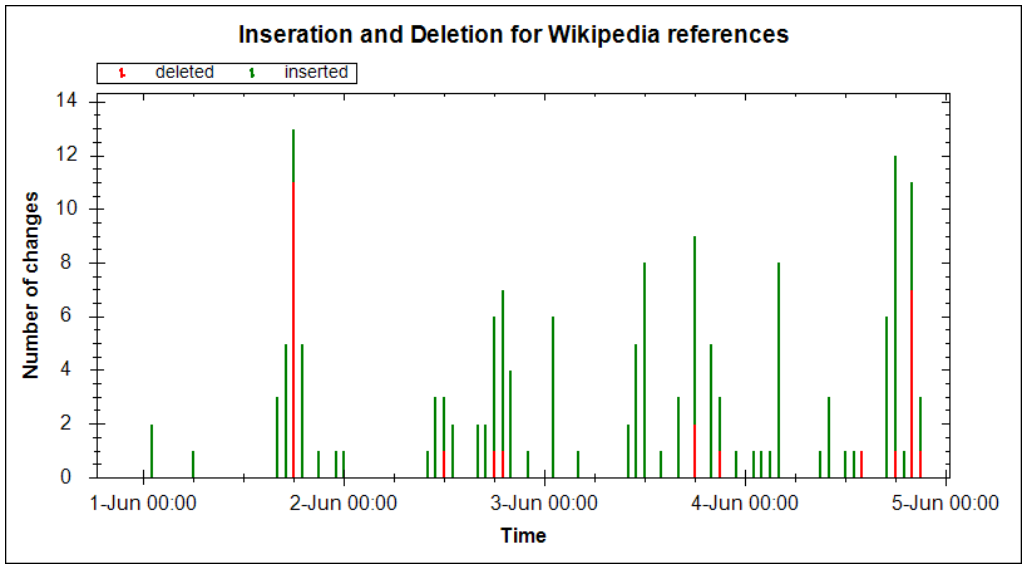


Figure 4 - Reference insertion and deletion over the time

Figure 4 shows the process of reference insertion and deletion in the English Wikipedia-article to E. coli O104:H4 outbreak[20] between 01/07/2011 and 05/07/2011.

We start our analysis of computing the number of biased references in Wikipedia articles by using JSI's Media Bias tool. We plan to offer a list of common references mentioned in Wikipedia articles as input for JSI's algorithms. For a given article JSI's Media Bias tool will be able to determine if an article is biased with regard to its reference list. With these data we will be capable to monitor the trends in a certain direction of bias in a Wikipedia-article.

We compute the following metric:

- The total number of articles which are classified as opinionated by containing as biased identified references in Wikipedia

Summarization:

We measure the objectivity of Wikipedia by detecting opinionated articles in Wikipedia. This detection will be based on identifying subjective expressions and biased references. By marking such articles and generating working lists we will decrease significantly the total number of articles which are identified as biased in Wikipedia.

[20] http://en.wikipedia.org/w/index.php?title=2011_E._coli_O104:H4_outbreak&oldid=447141231 (accessed: 30.08.2011)

5 Defining metrics based on editor behaviour

In order to quantitatively measure the existence of the socio-technical mechanisms described in section 3.2, a number of metrics are constructed and used. At the core, we are trying to find metrics for behavioural mechanisms that lead to a decreased"survival rate" of certain edits due to their carriers and other features but not their actual content. This means the same edit could have a higher or lower chance to stay in the article if it was submitted by a certain type of editor in a certain type of "social" editorial context. If specific mechanisms favour specific editor types' contributions, this is very likely leading to a biased article, as only certain editor types will have their viewpoint represented and it is highly unlikely that a fraction of dominant editors can and will transcend their own point of view to include all existing relevant points of view if they are not their own.

To find metrics that can be used to predict and identify biased articles, it is first necessary to find metrics that measure and predict the behavioural mechanisms that probably lead to, and are typical for, biased articles.

For all the socio-technical mechanisms described in section 3.2, we are therefore deriving those metrics. As a first step, we concentrated on protective ownership behaviour in an article. The metrics described below will mostly be used for exploring the other mechanisms in a similar manner and are the foundation to do so. We describe them using ownership behaviour to make a clearer example.

5.1 Variable construction - Example: Ownership behaviour

For measuring ownership behaviour, we use the following variables we constructed and that will be used in modeling via the metrics described in section 5.3:

- Reverts: The basis for modeling most of the behavioural patterns identified in section 3.2 is user reverts, i.e. undoing actions of a previous editor. This is also a key element in modeling protective ownership behaviour, since the main act of "protecting the own turf" is repealing changes made by perceived newcomers or outsiders. Modeling reverts is not as straightforward as it may first seem. Conceptually, a "revert" can have different meanings which in turn can be detected by using different techniques. We describe the current state of the art in revert detection and the improved detection method we developed in section 5.2. It was necessary to improve existing methods as reverts are essential to capture user behaviour in Wikipedia and the available methods' results in finding reverts was not sufficient.

  Based on the results derived with our revert detection techniques, we can on the one hand find correlating variables of being reverted and identify an article climate of reverts and on the other hand we can construct a network of revert relationships between the editors in the article, with the nodes being the editors and the edges being directed revert-relations, with differing strength, dependent on the number of reverts between two editors. This network can also be used to cluster editors together that never revert each other in contrast to editors that mutually revert themselves. It is also enriched by user features we talk about below.

- Editor features: what are the specific features of the editors being reverted or reverting frequently? We take into account:
  - Age:
    - Article: how many edits made in the article?
    - General: how old is the account?
  - User logged in / not logged in?
  - Activity: how much activity recently (last week, last 2 weeks), regular edits?

- (Awareness: Has article on watchlist? , not yet sure if possible)
- Revert Aggressiveness/Suffering: The number of times an editor has reverted and was reverted (suffered a revert), normalized for 1. the number of times he edited and 2. for the average wordiness of his edits (only for reverted)

- Article features ("Article climate") for every revision:
  - Edit concentration: how are the edits distributed over the editors (gini coefficient)
  - Revert concentration: active and passive reverts (analogous to edits)
  - Word possession concentration: distribution of word possession over the editors (for a certain revision and over a period of several revisions)
  - Templates used
  - How many users have the article on their watchlist?
  - Recent (1 week) number of edits, reverts, vandalism, new editors
  - Overall level of Vandalism

- Edit features: As we are interested in what social interactions shape the dynamics of edit survival in an article, the effect the information content of the edit has on its survival is not of primary interest. It has, however, to be incorporated in a comprehensive model as it has a huge effect on the probability of an edit being reverted. To get a clear picture of the non-content related features, the content-related features have to be filtered out as a kind of "data noise" and/or incorporated in calculating the correlations and causal relationships between the different variables so there are no hidden effects caused by them that get attributed to another variable.

  To approximate at least partly the different content types of edits and take into account their effects, we have identified some edit features, which are feasible to compute and have partly been shown in the literature to correlate with the probability of being reverted:
  - Length in characters/words (also negative, i.e. deletes)
  - Average word length in characters
  - (Un)commonness of used words in the Wikipedia, via the Corpex Wikipedia Corpora Explorer developed for RENDER[21]
  - Time of day of edit
  - Point in time in article life (relative to it's age in edits)
  - Length of edit comment (if not automatically generated)
  - Wiki-Syntax included (new section, reference, image, template, function, etc.)

Cleaning of the data: Before the afore mentioned variables can be used to construct valid metrics some "irregular" behaviour that would skew the results has to be filtered out of the data.

We therefore addressed:

[21] http://render-project.eu/tools/corpex/

- Vandalism edits: For example, vandals rarely log in to vandalize, or they use (not very old) sock puppet accounts. As vandalism almost invariably gets reverted eventually, this would skew some of our results. We therefore employed some proven high-quality vandalism-detection techniques (WikiTrust, Cluebot NG vandalism detection bot) to filter those edits out when computing some of our metrics.
- Bots: Bots do edit in very narrow, predefined ways and can be involved in socio-technical mechanisms (as pseudo-human agents), but are not in many cases. Therefore we filtered out bot edits for some metrics or did at least tag them as such.
- Frequent savers: Some editors hit the "save" button several times while working on an article (to be safe from data loss), whereas others just hit it once after finishing their work. This would artificially boost the edit count of certain editors. We merged edits by the same editor in a specific period (30 minutes) to adjust for this effect, as was done in previous research.

5.2 Revert detection

This section describes our extensive work on developing a comprehensive revert detection method to model intra-article, inter-editor revert behaviour as a foundation for the analysis explained in section 5.1.

5.2.1 Reverts as the basis for modeling user behavior

Editing is the most occurring, clearly measurable activity in Wikipedia. It takes place in two main domains, content and discussion/talk pages.[22] A distinction has to be made between a) article content, where editing means contribution of content and b) discussion/talk, where it is straight communication between editors. The editing and revert dynamics therefore are quite different, respectively. In the analysis method described in this section, we concentrate on the content namespaces and among those primarily target the article namespace.[23]

When an edit is performed, the following variables are directly logged: Name and ID of the article, date- and timestamp, username or IP of the editor, editor registration yes or no, a (optional) comment, article is redirect yes or no, edit is minor yes or no, text of the revision in Wiki syntax.

But there is much more, interesting information to be inferred by looking at indirectly recorded variables, as partly described in section 5.1, like length of an edit, etc. To analyze the social dynamics between the editors in an article, even more interesting are the actions of editors *relating* to each other.

There are two relevant action types to derive information about from the available editing data:

- **Co-editing**, which is editing by more than one editor that is content-wise leading the article in a similar direction (for example, two editors making edits that introduce a identic fact into an article at different times). To measure this, one would have to be looking directly at the semantic content of the edits.
- From **reverts**, on the other hand, one can draw some conclusions about the relation of the actions (edits) of two editors indirectly, *without* knowing the semantic content. A revert is broadly understood (but not clearly defined), according to the official Wikipedia guidelines, as an action of an editor

[22] More in detail, there are the following namespaces in Wikipedia: Articles, discussion, user pages, user talk, Wikipedia, Wikipedia talk.

[23] We exclude discussion/talk in this step because they follow very different edit dynamics than content namespaces that can't be modelled with the same approach. We focus on articles, furthermore, as they attract a very different subset of the editor body than user pages and the Wikipedia namespace do.

> "undoing the effects of one or more edits" and "(m)ore broadly, reverting may also refer to any action that in whole or in part reverses the actions of other editors."[24]

A revert example: If edit 1 writes an article consisting only of the word "apple", edit 2 adds "pie" after "apple" and edit 3 deletes only the word "pie", we can conclude intuitively, without understanding the semantic content, that edit 3 wanted to delete the content introduced by edit 2. If those were edits by two different editors, we can further conclude that editor 3 wanted to *revert* editor 2. By using these inferences, it is possible to build a social network graph of an article, like it has been done e.g. in Kittur et al. [20].

Wikipedia offers some tools in its interface, e.g. the "undo" button next to an edit in the article history dialog, that enables to undo the actions performed in that edit under certain circumstances.[25] But reverts can of course also be done manually, deleting words from or adding them to the revision text.[26]

5.2.1.1 Why detect reverts?

By analyzing the user interaction modeled through the detected reverts and combining them with additional data generated out of the article history, it is possible to detect patterns that are typical for the socio-technical mechanisms described in section 3.2.

5.2.2 State-of-the-art of revert detection in Wikipedia

Some related research work regarding Wikipedia editing behavior takes into account reverts ([17], [18], [19], [20], [21], [23]). Still, the work we found invariably deems only so called "identity reverts" as an adequate method to account for revert behavior. This rather simple method relies on finding two revisions that contain exactly the same text, using MD5 hashes.[27] It then defines a *reverted edit* very coarsely as all the edits that lie between two identic revisions. The second identic revision is interpreted to be the *reverting edit,* the first identic revision is the one that is reverted to. Table 1 shows an example of how reverts are detected in this manner.

Revision Number	Revision content	Words deleted/added (actions taken) in the edit	MD5 Hash (simplified)	Detected identic and reverted revisions
1	Zero	(ignored for this example)	Hash1	Like revision 5
2	Zero Apple Banana	+"Apple" +"Banana"	Hash2	Reverted by revision 5
3	Zero Apple Banana Coconut Date	+"Coconut" +"Date"	Hash3	Reverted by revision 5
4	Zero Coconut	-"Apple" -	Hash4	Reverted by

[24] http://en.wikipedia.org/wiki/Help:Reverting (accessed 13.09.11)

[25] The undo-function does only work when the section in which the edit took place has not been edited since. There is also a „rollback" function available only to administrators, which enables to go back to a certain revision and discard all revisions that were created afterwards. Administrators make up a very tiny proportion of the overall user base though.

[26] A „revision" is the version of the article created by an edit. I.e. edit 1 creates revision 1, edit 2 creates revision 2 etc. An edit is performed any time the content of the article is changed and saved.

[27] This is done by creating MD5 hashes for all the revision texts and then looking fort two identic hashes. This is a common method to compare if two text contents are identical [22].

	Date	"Banana"		revision 5
5	Zero	-"Coconut" - "Date"	Hash1	Like revision 1 → revert of revisions 2,3,4

Table 1 – Example of the result of the simple identity revert detection method

Where used in the research literature, this simple identity revert detection method (SIRD method from here) is never motivated by a definition or any theoretical discussion of what a revert actually is, as human behavior, intentionally carried out to undo someone else's actions. Rather, the motivation is that it "is computationally simple and determining exactly which editors' revisions were lost due to the revert is straightforward." [18], which seems to be a sufficient reason to use it for many researchers. By working in this particular way the method itself implies the definition of what a revert is. The coarseness of this approach is acknowledged to some extend by Priedhorsky et al. [21], who state that understanding and taking into account the *intention* of a revert is challenging, and because of this resort to using only the SIRD method. Also, in most of this research literature, the focus of the work in question is not on reverts but rather on a multitude of different metrics.

Another rationale behind using only SIRD is that it supposedly covers most of the existing reverts: Kittur et al. [20]showed that by combining a method based on edit comments (which include the expressions "revert" and "rv") and SIRD, 95% of the *reverting edits*[28] found as a result could be found using only the identity reverts. In a number of subsequent papers by other authors ([18][21]) this finding was used to conclude that the mere 5% additional reverting edits found with comments do not justify the effort of using this method on top of the SIRD method. Still, to our knowledge, there was no investigation so far if other detection methods might find even more reverts, as many users do no not attach comments to their edits ([20] [21]) and MD5 hashes cannot be used to find partial reverts [18]. Nor has there been any evaluation whatsoever of the false-positive-rate of the SIRD method from the perspective of the Wikipedia definition of a revert – meaning for example if found reverting edits are really only undoing other editors work. This is an especially crucial issue in the light of the very simplistic definition of a revert the method implies.

Although we know of at least one analysis toolkit that extends to some small degree the above described narrow definition of reverting and reverted edits,[29] we have not seen an essentially more elaborate approach at modeling revert behavior so far.

5.2.2.1 Deficiencies of existing revert detection methods

The identity revert detection method relies on the existence of two identical revisions. In reality however, there are partial reverts that undo contributions by an edit without creating a duplicate text version. These cannot be detected using only MD5 hashes [18]. In Table 1 this is exemplified by the actions of revision 4, which deletes all words introduced by revision 2 while still generating a completely new revision. Intuitively, one could say that revision 4 is reverting revision 2. The SIRD method, however, will not detect the revert relationship in this way, but, as shown, assign revision 5 as the reverting revision of revision 2. In a scenario where revision 5 would be non-identical to revision 1 the method would not even detect *any* revert of revision 2. Depending on the definition of a revert being used, this means the **recall** (finding all relevant reverts) and the **precision** (only finding true-positives) of this method can be sub-optimal. Looking further, at the conceptual model with which the SIRD method uses, there are, intuitively, some

[28] In the paper, Kittur et al. [20] refer to this as the found „reverts" while actually it is just the number of found reverting edits that either have an identic previous version or a comment including „revert" or „rv". It cannot be in all cases concluded what revision they actually reverted when there is no identical version (i.e partial revert) and no indicator in the comment (e.g. just the word „revert").

[29] The toolkit works as the described SIRD method, but makes differences between the revisions marked as „reverted": those made by the first editor after the „reverted-to" revision get marked as „possible vandalism", while the remaining reverted edits are put in a separate group [16].

inconsistencies regarding what is identified as a revert. In Table 1, Revision 2 is considered a reverted edit as well as revision 3, only because they accidentally lie between two identical revisions. Even though revision 3's actions are not undone by any following edit (the Wikipedia definition of a revert).

Moreover, those partial reverts happen quite often, as editors can do reverts manually by deleting words out of the text or adding words. Even more importantly, the prominent "undo" function described in section 5.1 also undoes only *one* revision's actions, which not automatically leads to a duplicate revision.

5.2.3 An improved revert detection method

For improving the precision and recall of the revert detection, the first step was to establish a clear concept and definition of what a revert is and then develop a model to detect all and only those edits that fit the definition.

According to the Wikipedia definition and taking into account what kind of data can be used to derive, with certainty, *intentional* acts of undoing another editors actions the following definition was derived:

> *An edit A is reverted if all of the actions of that edit are completely undone in one subsequent edit B. Edit B has then reverted edit A.*

Note that this definition includes only the content an edit adds to or removes from the Wiki syntax version of the article, not any metadata like comments. It does further not rule out that edit B performs other actions on top of undoing A's actions or the actions by a number of different edits.

Note as well that, if A's actions have been undone only partially, this is not counted as a "partial revert" and if all of A's actions have been undone in a collective effort by many partly reverts, A is not counted as "reverted" (although all its actions were undone). This is due to the fact that, then, we could not assign a single reverting edit B and thus not unambiguously determine the reverting and the reverted edit in every case. Those two restrictions do not comply with the Wikipedia concept of a revert. We had to introduce them nonetheless due to computational constraints and also because the resulting revert relations would not have been much meaningful in many instances. We are sure our approach still is superior to the state of the art method in precision, recall and the meaningfulness[30] of the found reverts.

5.2.3.1 Revert detection - Implementation

To operationalize the "actions" of the editors we use added and deleted word tokens, i.e. character chains separated by white spaces. We operate on the Wiki syntax, not on the front-end text content.

To compare the revisions, MD5 hashes and text difference comparisons (DIFFs[31]) are used. To find reverts adhering to our definition using DIFFs would suffice. Still, using MD5 hashes on top speeds up the computation significantly. We use those methods to build up a content list for each revision. First, we identify identic revisions via MD5 hashes and mark the second identic revision. Then we check via text DIFFs for the last previous edits A, A', ... , A^n , which performed the exact opposite of a subset of actions of a subsequent edit B we are analyzing. If such a negated subset is found, the content introduced or deleted by A is taken out of the content list for B. Table 2 exemplifies the procedure.

[30] Meaning: Discovering truly intended revert actions taken by an editor against another.
[31] http://en.wikipedia.org/wiki/Diff

Revision Number	Revision content (text)	Words deleted/added (actions taken) in the edit	MD5 Hash (simplified)	Content list (contains revision numbers)	Content list differences	Detected reverts
1	Zero	(ignored for this example)	Hash1	1	+1	
2	Zero Apple Banana	+“Apple” +”Banana”	Hash2	1;2	+2	Reverted by 4
3	Zero Apple Banana Coconut Date	+”Coconut” +”Date”	Hash3	1;2;3	+3	Reverted by 5
4	Zero Coconut Date	-“Apple” - “Banana”	Hash4	1;3	-2	Revert of 2
5	Zero	-“Coconut” - “Date”	Hash1	1	-3	Revert of 3
6	Zero Fig	+”Fig”	Hash5	1;6	+6	Reverted by 8
7	Zero Grape	+”Grape”	Hash6	1;6;7	+7	Reverted by 8
8	Zero Huckleberry	-“Fig” -“Grape” +”Huckleberry”	Hash7	1;8	-6; -7	Revert of 6,7

Table 2 – Example of the result of our extended revert detection method

When comparing the revisions 1 to 5 in Table 2 with the example in Table 1 it shows that, with our method and adhering to our definition, revision 5 is only reverting revision 3, while revision 4 is reverting revision 2. This means higher precision (according to ours and Wikipedia’s definition) is achieved. With our method, we additionally detect the revert by revision 8 of revisions 6 and 7, where no duplicate revisions can be found. This means our method achieves a higher recall as well.

Note that an identity-reverts-only method can also find *too many* reverts (false-positives) in the attempt to mark everything between two identic revisions as reverted. I.e. that with our method, the number of found reverts might actually decrease in some articles due to the higher precision and might by increase due to higher recall.

There are other examples of reverts where our method can extract much more meaningful revert behaviour then the SIRD method. One recurring scenario is the repair of one vandalistic revision by a couple of following revisions, all trying to recreate the revision before the vandalism occurred. If the last one in this row of repair edits recreates that original revision, all other repair-edits will be marked as reverted with the identity-revert-only detection method.

5.2.3.2 Revert detection - Results

Our evaluation so far shows the following results:

- Our method detects a mean of 10 % more reverts with the DIFF-based part of the algorithm than with identity-revert-based MD5 hashes.
- It detects up to 50% more reverts for short articles and up to 12% for very long articles with text DIFFs then with only MD5 hashes. We have still to evaluate why it finds so many more results for shorter (and probably younger articles) articles and evaluate as well if it performs better for specific kinds of articles.

We also compared our method with revert detection by looking for regular expressions like “rv”, “revert”, “undid revision” and a couple of others in the revision comments and 99,6% of reverts indicated by comments were also found. Additionally our method found more reverts than indicated in the comments (as users do not always comment on their revert) and detected the reverted edit(s) even for those edits where the comment did only indicate a revert, but not *which* revision was reverted. The results suggests that quite a number of reverts are neither covered by MD5hash- nor by comment-based detect methods.

5.2.4 Revert detection - Conclusion

As far as we can tell, our method is the first to try to detect revert behavior in Wikipedia according a specific definition of reverts, inspired by actual user behavior and Wikipedia’s own revert definition. We do this as it is crucial to build an accurate model of who is reverting whom in an article do derive an understanding of the underlying social dynamics in the article. This understanding is the key to discover recurring behavioral patterns leading to bias.

A gain of around 10% recall with a specific method is not to be dismissed, as specific ways of reverting might be used by specific types of editors and only detected with one particular method.[32] I.e. when not using one method some user groups might be underrepresented in the eventual revert model.

But, more importantly, actions which are not intentional reverts are not spuriously detected as such in our method. This gain in precision increases the quality of the interpretation of the revert relationships between users tremendously. This is the basis for our further analysis of the socio-technical mechanisms in biased articles.

5.3 Behaviour-related metrics

The following metrics are expected to be indicators for recurring proper patterns of articles that have been identified in previous research as showing strong ownership behaviour (those carrying the “maintained” template).

- We expect that “new” editors, without much previous activity in the article or in Wikipedia in general (or those without even a user account) will be reverted much more frequently than others, even when trying to constructively contribute to the article (i.e not vandalizing). “Older” editors (in terms of edits done) will be reverting other editors much more. This effect is supposed to be stronger for articles that have been identified as being prone to showing ownership behaviour, than for an average Wikipedia article.
- We check for higher-then-average concentration of reverts, edits and word possession among the users in the "maintained"-articles.

[32] E.g. administrators could be generating more identity-reverts as they have exclusive access to a “rollback” function that eases this process. It enables to specify an old revision and then undoes all subsequent edits up to the present. For X edits to be undone, average users would have to click the “undo” button X times or delete words manually to achieve this.

- From the reverting-reverted relationships, we are also constructing a "social" network graph between the editors in the articles, used to check, among other things, for high concentration clusters of edit activity and word ownership in one or several core editor group(s) (with their members not mutually reverting each other) in contrast to a marginalized, highly reverted group of editors with less edits.
- In every analysis step, we control for influencing secondary effects of the edit content and the article climate in which the edits are made (and reverted) that could mask or inflate the effects of primary interest.

As a result we aim at identifying the patterns that are typical for ownership behaviour. We are evaluating this by testing our derived patterns and metrics as predictors to classify articles that have been tagged with templates identified in the literature as usually coinciding with ownership behaviour ("maintained" articles).

5.4 Display bias warnings based on behavioural patterns

The above-described variables, metrics and patterns will eventually be condensed into some high-level indicators for bias in articles, which can be shown to the average Wikipedia user in an intuitive interface.

Although not part of this deliverable, we give some examples in Figure 5 for behavioural-pattern-based warnings that could be presented to users. It coarsely exemplifies what the eventual aim of the behavioural analysis is.

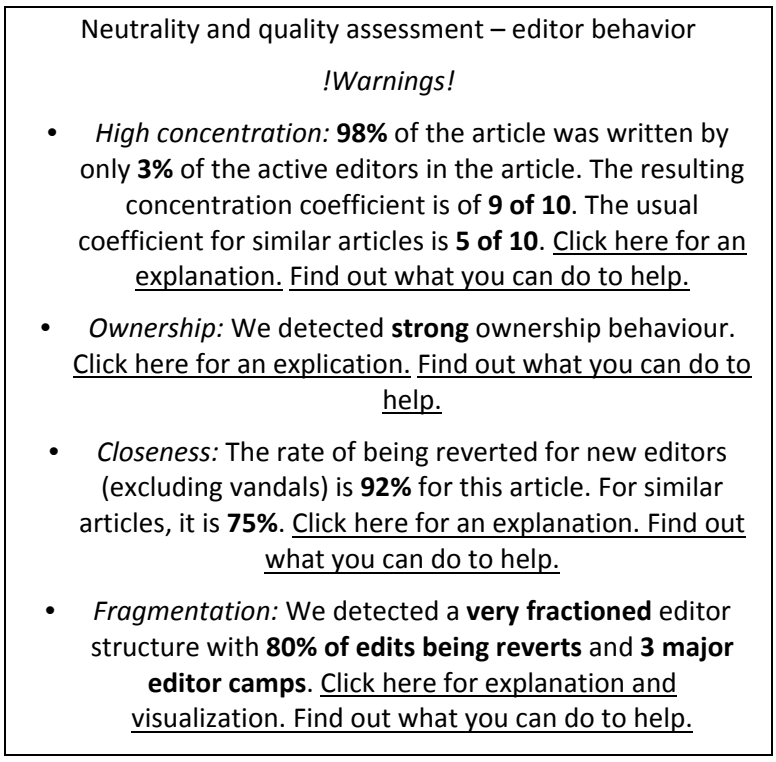


Figure 5 – Example warnings in an article quality assessment interface

The figure is just a first draft of how those bias-related warning will be presented. They would be displayed when a specified threshold in one metric or in a specific combination of metrics is reached that allows with

some certainty the conclusion that bias is present.[33] Those behavioural metrics will of course also be combined with those metrics described in section **Fehler! Textmarke nicht definiert.**. The eventual version of such a dashboard for warning will apply an elaborated visual interface with a very intuitive access to the relevant data. The link “Click here for an explanation.” then allows for a drill-down into the specific metric and its data whereas the link “Find out what you can do to help.” explains what specific actions can be taken by the editor, e.g. in case of a strong ownership group start a discussion about the issue with reference to the metric results, forward the metric results to an administrative body etc.

[33] To reach this certainty and explain why we reached it is ultimate goal of the research mentioned above in the rest of section **Fehler! Verweisquelle konnte nicht gefunden werden.**.

6 Conclusion and Future Work

In this deliverable we introduced the metrics for the case study on Wikipedia. We worked out the necessity of completeness, timeliness and objectivity as parameters for high-quality Wikipedia articles. These aspects of quality are part of the Wikipedia internal quality assurance process and the main issues, which users can assess with help of the article feedback tool. We also mentioned the need to support the Wikipedia community. Because of the size of Wikipedia, many flawed articles are not identified yet. Only articles with many views can be part of the enduring improvement process.

For this we defined in section 2.2 three use case scenarios, to underline our main motivation to increase Wikipedia's quality by supporting users (reader and editors).

These are:

UCS 1: Display warnings to the reader when detecting bias

UCS 2: Notify authors that an article needs to be updated

UCS 3: Lower the barrier for readers to extend and/or correct articles

Whenever, we recognise a hint for a bias (caused by fact incompleteness, out-of-dateness, subjective expressions, or by analysing the editors network), we will signal a warning to the users. Thereby, they will be capable to expand and improve the article. By analysing Wikipedia in matters of diversity (as essential criterion for quality), we will able to generate for instance working lists, which increase the visibility of orphaned articles with a need of revision. In a first step, we will compare our measuring results with the assessment data set of the feedback tool. These assessments are done by users of Wikipedia and could be used as a Gold Standard in our evaluating process.

Additionally, we have to evaluate the usability of presenting the measuring results for a certain article and to find out, which are the best ways of presenting and offering them to the community. At the beginning, we should address e.g. a Wiki project, which is not too complex and deal with one specific thematic field. At a later point in the project, we need to create appropriate conditions to test the metrics for a variety of languages and topics.

References:

[1] Anderka. M., Stein, B. und Lipka N., *Towards Automatic Quality Assurance in Wikipedia*, In: Proceedings of the 20th international conference companion on World wide web, 2011.

[2] Arazy, O. and Nov, O., Determinants of Wikipedia quality: the roles of global and local contribution inequality, Proceedings of the 2010 ACM conference on Computer supported cooperative work, 2010.

[3] Blumenstock, JE, Size Matters: Word Count as a Measure of Quality on Wikipedia. In WWW '08: Proceeding of the 17th international conference on World Wide Web (2008), pp. 1095-1096, 2008.

[4] Brändle, Andreas, Zu wenige Köche verderben den Brei, 2005.

[5] Fortuna, B., Rusu, D., Trampuš, M., Dali, L., Štajner, T., and Grobelnik, M. *Prototype of the Fact Mining Toolkit*. RENDER Project Deliverable D2.2.1. 2011.

[6] Giles, J., *Internet encyclopaedias go head to head*. Nature 438, S. 900-901. 2005.

[7] Halavais A. and Lackaff D. *An analysis of topical coverage of Wikipedia. Journal of Computer-Mediated Communication*, 2008, 429–440.

[8] Hammwöhner, Rainer (2007): *Qualitätsaspekte der Wikipedia*. In: Stegbauer, C., Schmidt, J., Schönberger, K. (Hrsg.): Wikis: Diskurse, Theorien und Anwendungen. Sonderausgabe von kommunikation@gesellschaft, Jg. 8. Online-Publikation: http://www.soz.uni-frankfurt.de/K.G/B3_2007_Hammwoehner.pdf .

[9] Hammwöhner, R., Fuchs, KP., Kattenbeck, M., Sax, C. *Qualität der Wikipedia - eine vergleichende Studie.* In: Achim Oßwald, Maximilian Stempfhuber, Christian Wolff (Hrsg.): Open Innovation. Neue Perspektiven im Kontext von Information und Wissen. Proc. des 10. Int. Symposiums für Informationswissenschaft. UVK, 2007, S. 77-90.

[10] Lih, Andrew, *Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource*, 5th International Symposium on Online Journalism, University of Texas, Austin, 2004.

[11] Ortega, F., PhD-Thesis: Wikipedia: A Quantitative Analysis, 2009.

[12] Stvilia, B., Twidale, M., Smith, L. C., Gasser, L., Information quality work organization in Wikipedia, JASIST, 59, 6, 983–1001. 2008.

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

[14] Wilkinson, D. and Huberman, B., *Cooperation and quality in Wikipedia*, In Proceedings of the 2007 International Symposium on Wikis. 157-164, (October, 2007), Montreal, Canada. 2007.

[15] Wöhner, T. and Peters, R., *Assessing the Quality of Wikipedia Articles with Lifecycle Based Metrics*, In: Proceedings of the 5th International Symposium on Wikis and Open Collaboration. 2009.

[16] Chichkov, D., *Pymwdat - Python MediaWiki Dump Analysis Toolkit*, http://code.google.com/p/pymwdat/

[17] Ekstrand, M. D. and Riedl, J. T. 2009. *rv you're dumb: identifying discarded work in Wiki article history*. In Proceedings of the 5th international Symposium on Wikis and Open Collaboration (Orlando, Florida, October 25 - 27, 2009). WikiSym '09. ACM, 1-10.

[18] Halfaker, A., Kittur, A., Kraut, R. and Riedl, J. *A jury of your peers: quality, experience and ownership in wikipedia.* In D. Riehle and A. Bruckman, editors, Int. Sym. Wikis, 2009.

[19] Kittur, A., Chi, E. H., Pendleton, B. A., Suh, B. and Mytkowicz, T. *Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie*. 25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007), 2007.

[20] Kittur, Suh, Pendleton, and Chi. *He says, she says: conflict and coordination in wikipedia.* In Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '07, pages 453–462, New York, NY, USA, 2007.

[21] Priedhorsky, Chen, Lam, Panciera, Terveen, and Riedl, *Creating, Destroying and Restoring Value in Wikipedia*, In GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work, pages 259–268.

[22] Rivest, R., *The MD5 Message-Digest Algorithm*, RFC 1321, MIT and RSA Data Security, Inc., April 1992.

[23] Suh, Convertino, Chi, and Pirolli. *The singularity is not near: slowing growth of Wikipedia*. In Proceedings of the 5th International Symposium on Wikis and Open Collaboration, WikiSym '09, pages 8:1–8:10, 2009

[24] Schlieker, Christian, 2005, Explorative Untersuchung von Wissen in kollektiven Hypertexten, Diplomarbeit, Fachbereich 08, Soziologie, Universität Bremen.

Annex A

A.1 Paper: Towards a diversity-minded Wikipedia

Flöck, F., Vrandecic, D., Simperl, E. 2011. ***Towards a diversity-minded Wikipedia***, in *Proceedings of the ACM 3rd International Conference on Web Science 2011*

Link: http://www.websci11.org/fileadmin/websci/Papers/112_paper.pdf