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

RENDER

Deliverable D5.1.2

Tools for diversity management in Wikipedia

Editor:	Angelika Adam, Wikimedia Deutschland (WIKI)

Author(s):	Angelika Adam, Wikimedia Deutschland (WIKI);
	Felix Keppmann, Karlsruhe Institute of Technology (KIT);
	Delia Rusu, Institute Jozef Stefan (JSI)

Deliverable Nature:	R
Dissemination Level: (Confidentiality)	PU
Contractual Delivery Date:	30.09.2012
Actual Delivery Date:	30.09.2012
Suggested Readers:	Wikipedia users, tool developers, researchers
Version:	1.0
Total number of pages:	30
Keywords:	Diversity, analysis, supporting tools, Wikipedia, community empowering

Disclaimer

This document contains material, which is the copyright of certain RENDER consortium parties, and may not be reproduced or copied without permission.

All RENDER consortium parties have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the RENDER consortium as a whole, nor a certain party of the RENDER consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Full Project Title:	RENDER – Reflecting Knowledge Diversity
Short Project Title:	RENDER
Number and Title of Work package:	WP 5: Diversity case studies
Document Title:	D5.1.2 - Tools for diversity management in Wikipedia
Editor (Name, Affiliation)	Angelika Adam (Wikimedia Deutschland)
Work package Leader (Name, affiliation)	Javier Caminero (TID)
Estimation of PM spent on the deliverable:	30

Executive Summary

This deliverable describes the tools which were developed to support Wikipedia users – readers and authors – to find, to understand and to cure non-neutral or biased articles in Wikipedia.

Diversity is a necessity for quality in Wikipedia. Articles are usually written by multiple editors, who may be biased towards a certain point of view. Either editors can transcend their personal point of view or a multitude of editors covers the significant points of view to create a balanced and all-embracing content description. In particular, articles about exotic topics or which are tagged as “controversial” could benefit from tools which help to display diversity lacks. We defined as most important diversity aspects for Wikipedia: thematic coverage, timeliness and neutrality. We also mentioned that analysing editor behaviour and interaction can give further important advices.

In this deliverable we present the analysis approaches which were developed by Wikimedia Deutschland, KIT and JSI. The results are visualised as show cases in a tool kit on the Wikimedia Toolserver. This so called RENDER Toolkit provides a central access point for interested users or researchers to test these tools and to design the development process as transparent as possible to include the community at an early stage. The results of our analysis tools combined with results of further assessments tools and Wikipedias quality assurance methods are the input for two supporting tools for Wikipedia users.

These tools are:

- The Article Statistics and Quality Monitor (ASQM): displays the diversity analysis results for every article. This tool will include the possibility to check for different metrics and provide a quick overview of the quality and the state of an article.
- The Task List Generator (TLG): enables a Wikipedia author to generate lists of articles related to a specific topic or preferred category, which need to be improved.

We present the major ideas and the plans to evaluate the usability and the benefit for Wikipedia’s readers and authors for their handling and work with Wikipedia.

List of authors

Organisation	Author
Wikimedia	Angelika Adam
KIT	Felix Leif Keppmann
JSI	Delia Rusu

Table of Contents

List of Figures and/or List of Tables

Abbreviations

ASQM – Article Statistics and Quality Monitor

LEA – LinkExtractor

TLG – Task List Generator

WP – Wikipedia

1 Introduction

This deliverable describes the tools which were developed to support Wikipedia users – readers and authors – to find, to understand and to cure non-neutral or biased articles in Wikipedia.

Articles are usually written by multiple editors, who may be biased towards a certain point of view. Either editors can transcend their personal point of view or that a multitude of editors covers the significant points of view to create a balanced and all-embracing content description. That means diversity is a necessity for quality in Wikipedia. In particular, articles about exotic topics or which are tagged as "controversial" could benefit from tools which help to display diversity lacks. We defined as most important diversity aspects for Wikipedia: thematic coverage, timeliness and neutrality in D5.1.1 [1]. In this report, we also mentioned that the analysis of editor behaviour and interaction can give further important advices.

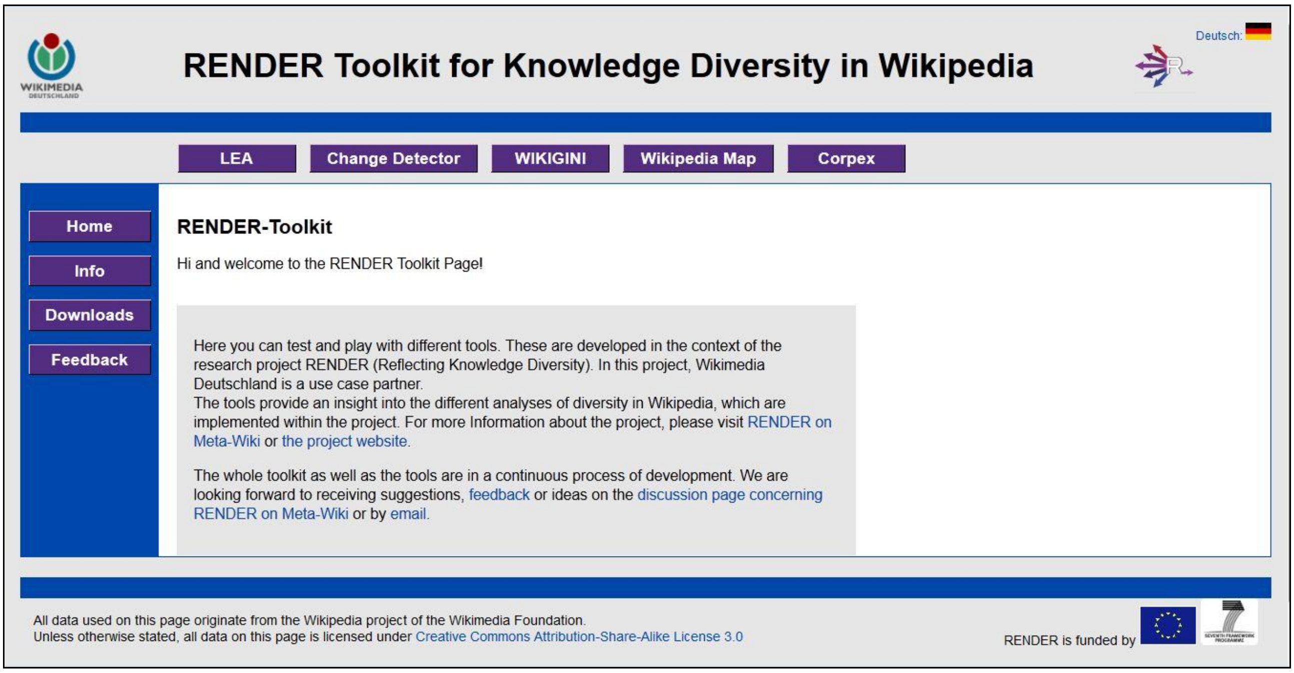
In this deliverable we present the analysis approaches which were developed by Wikimedia Deutschland, KIT and JSI. The results are visualised as show cases in a so called RENDER Toolkit. The results of our analysis tools combined with results of further assessments tools and Wikipedias quality assurance methods are the input for two supporting tools for Wikipedia users – the Article Statistics and Quality Monitor (ASQM) and the Task List Generator (TLG).

The deliverable is structured as follows. We start by describing the diversity analysis approaches and their showcase visualisation within the RENDER toolkit in Section 2; followed by additional services and tools - CatGraph and JSI's NewsFeed - we are using for the diversity management tools in Section 3. Section 4 describes the supporting tools for Wikipedia users - the Article Statistics Quality Monitor and the Task List Generator. The supporting tool evaluation plans are presented in Section 5. The final section of the deliverable presents concluding remarks and future work.

2 Diversity analysis approaches and the RENDER Toolkit

In the Wikipedia use case we are analysing different aspects of diversity – fact coverage, timeliness, neutrality, and editor behaviour and interaction. The sum of all analytical results is provided within two supporting tools for Wikipedia users. This enables readers and authors to recognize and understand articles which are non-neutral, have a bias in a certain direction, or for which there is additional information available.

We created together with our colleagues from KIT and JSI a set of individual analysis and tested various approaches. We decided to visualize these developed analysis tools as show cases on an information page - the RENDER Toolkit[1] - on the Wikimedia Toolserver[2]. Figure 1 shows a screenshot of the start page.



RENDER-Toolkit

Hi and welcome to the RENDER Toolkit Page!

Here you can test and play with different tools. These are developed in the context of the research project RENDER (Reflecting Knowledge Diversity). In this project, Wikimedia Deutschland is a use case partner.
The tools provide an insight into the different analyses of diversity in Wikipedia, which are implemented within the project. For more Information about the project, please visit RENDER on Meta-Wiki or the project website.

The whole toolkit as well as the tools are in a continuous process of development. We are looking forward to receiving suggestions, feedback or ideas on the discussion page concerning RENDER on Meta-Wiki or by email.

All data used on this page originate from the Wikipedia project of the Wikimedia Foundation.
Unless otherwise stated, all data on this page is licensed under Creative Commons Attribution-Share-Alike License 3.0

Figure 1 The RENDER Toolkit – start page

This presentation channel allows us to make the continuous development process transparent. So, we provide interested readers, Wikipedia authors, and researchers the possibility to test these tools and approaches at an early stage and to give feedback or suggestions by using an integrated feedback function. Furthermore, we use the Toolkit page as a central access point for the data sets[3] created within the project such as the neutrality template data sets or the Wikipedia Historical Article Data set (WHAD). The latter was created and provided for free download and use by our project partner Google. The Toolkit will be continuously extended to include new analytical approaches and also improved and adjusted according to user feedback or further suggestions. It is unproblematic to expand the supporting tools with new functionalities. A more detailed presentation is given in chapter 5.

In the following sections, we describe analysis approaches which have already been implemented and exist as a showcase within the RENDER Toolkit. We also give an outlook on future expansion, which are currently still under construction. The whole programme code of the toolkit and each tool is under a free licence and available on the Wikimedia Toolserver[4].

[1] http://toolserver.org/~RENDER/toolkit
[2] http://toolserver.org
[3] http://toolserver.org/~RENDER/toolkit/downloads
[4] https://svn.toolserver.org/svnroot/p_render/toolkit

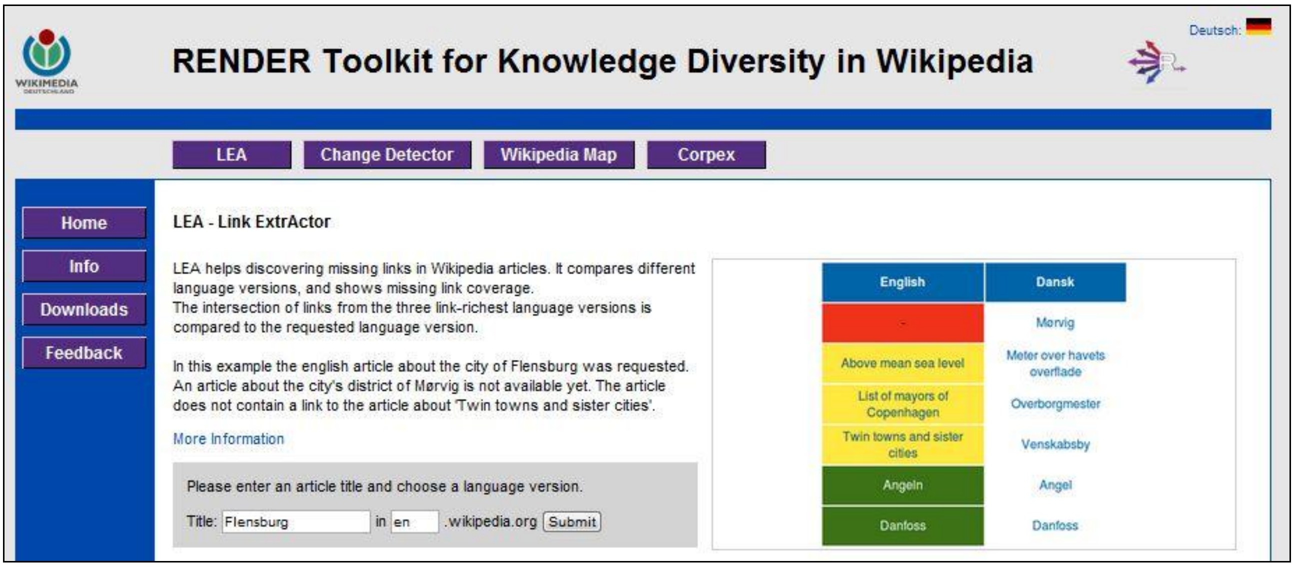
3 Analysis approaches and tools concerning thematic coverage

Fact coverage is an important aspect of knowledge diversity in Wikipedia. If fundamental information is missing in an article, it does not cover all sides of a topic and therefore it could not be neutral. Currently, we have developed one tool – the LinkExtractor – which is part of the Toolkit and described in more detail below. We are working on further approaches which are presented section 3.1.2.

3.1.1 LinkExtractor (LEA)

Wikipedia articles contain internal links[5] to other articles. These also called wikilinks are usually inserted to give further explanation about major terms or concepts. We use these internal links as one approach to analyse the thematic coverage of an article.

The LEA tool[6] explores the thematic coverage of a Wikipedia article with help of wikilinks, i.e. internal language specific Wikipedia links. A user can insert an article title and a language version to start the analysis, as shown in Figure 2.



RENDER Toolkit for Knowledge Diversity in Wikipedia

LEA - Link ExtrActor

LEA helps discovering missing links in Wikipedia articles. It compares different language versions, and shows missing link coverage.
The intersection of links from the three link-richest language versions is compared to the requested language version.

In this example the english article about the city of Flensburg was requested. An article about the city's district of Mørvig is not available yet. The article does not contain a link to the article about 'Twin towns and sister cities'.

More Information

Please enter an article title and choose a language version.

Title: Flensburg in en .wikipedia.org

English	Dansk
-	Mørvig
Above mean sea level	Meter over havets overflade
List of mayors of Copenhagen	Overborgmester
Twin towns and sister cities	Venskabsby
Angeln	Angel
Danfoss	Danfoss

Figure 2 LEA query form for the example “Flensburg” in English

Now, the LEA algorithm identifies in a first step the three largest articles[7] in other language versions which are connected by inter language links and discuss the same topic. Then the intersection is created out of all included internal links. We suppose these links (which represent unique concepts) display the most important facts about that topic. In the next step of the calculation this link set is compared to the internal links the requested article contains. A graphical overview on the algorithm is shown in Figure 3.

[5] http://en.wikipedia.org/wiki/Help:Wikilinks#Wikilinks
[6] http://toolserver.org/~RENDER/toolkit/LEA/
[7] measured by the number of the containing internal links

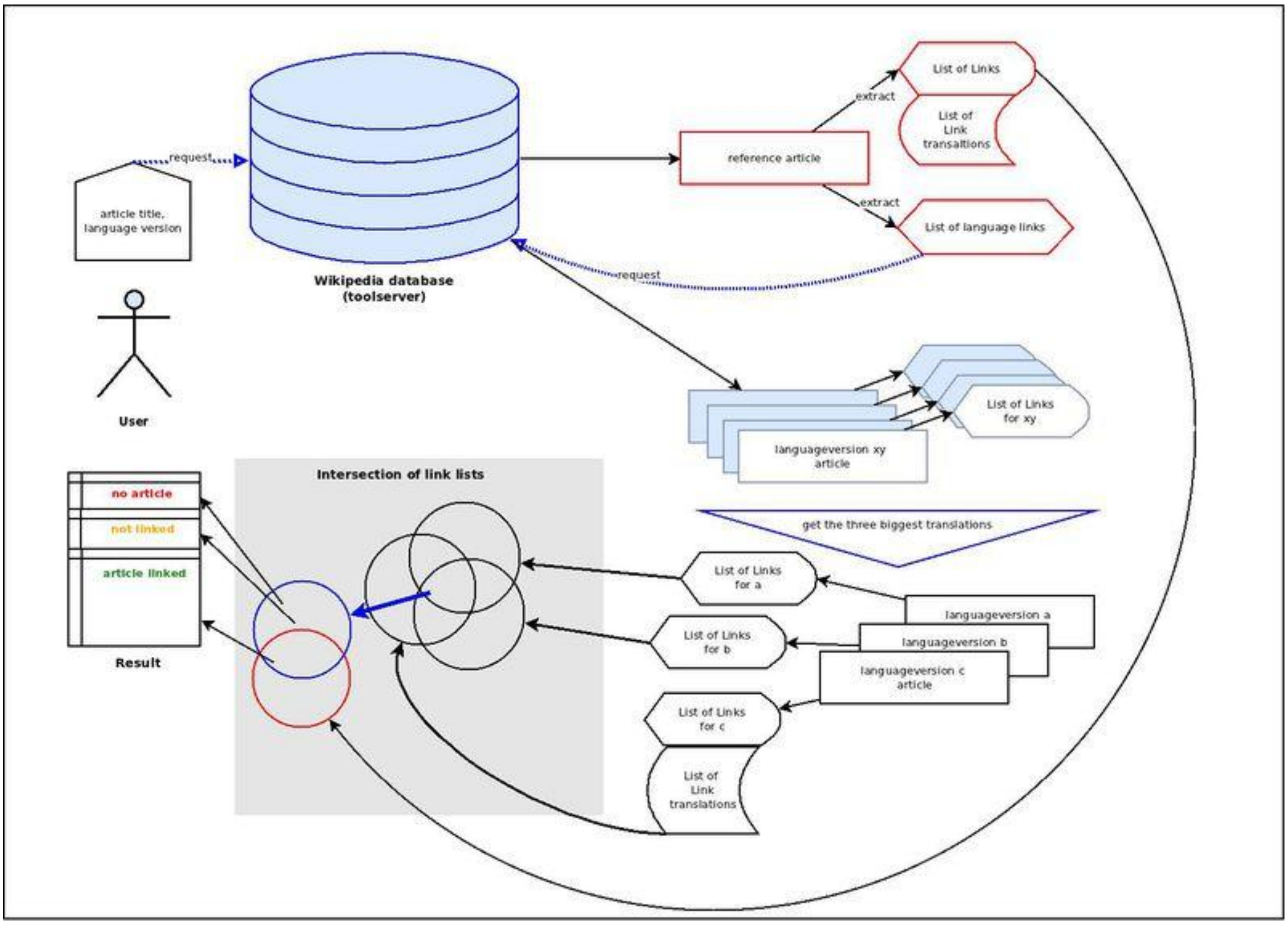
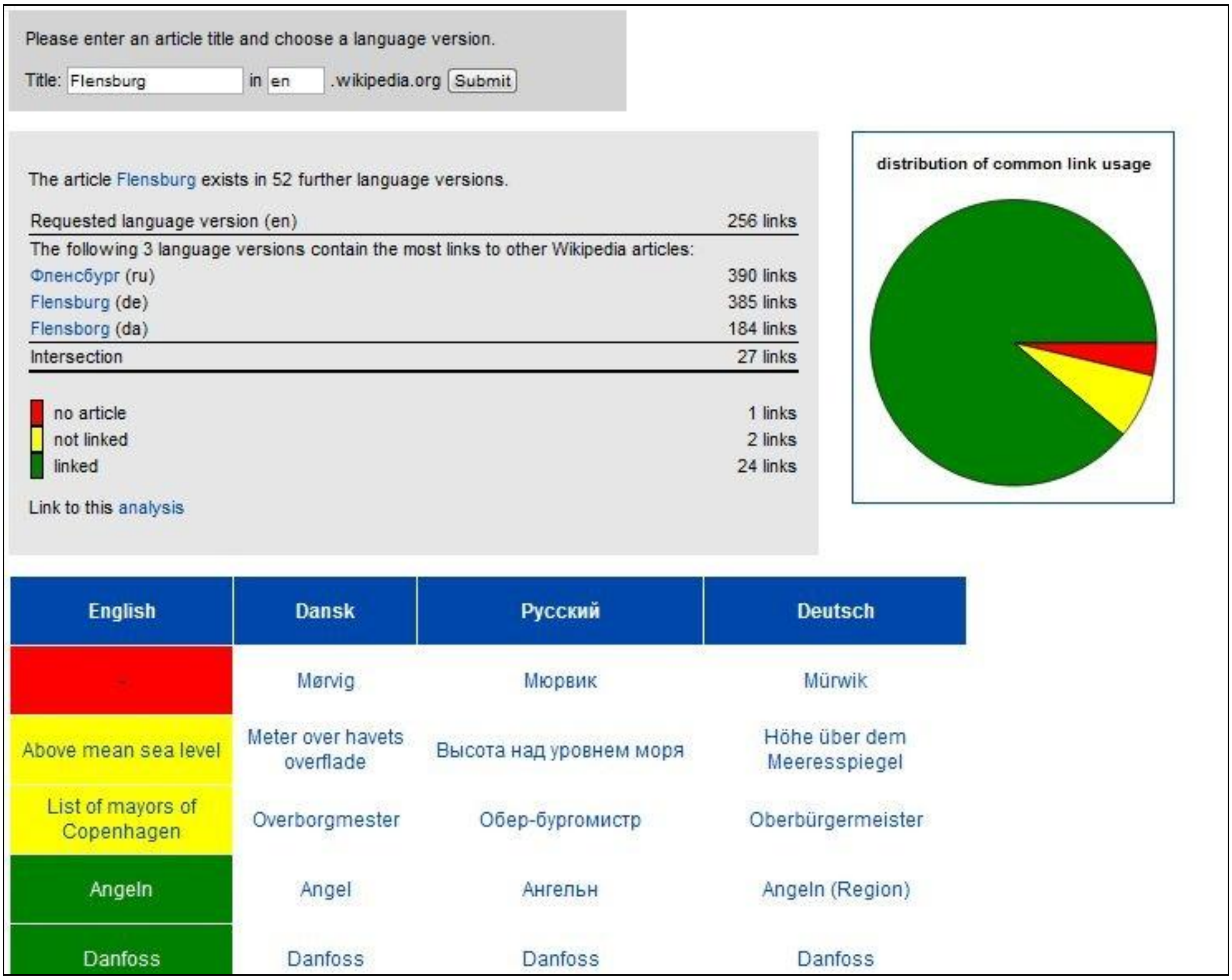


Figure 3 Functionality of the LEA Tool

The results of the analysis process for the example "Flensburg" in "en" are shown in Figure 4. Within the grey box, the user gets information about the number of internal links within this requested article, the largest corresponding articles in other language versions and the number of their internal internal links. The intersection of the internal links is calculated. Furthermore, the intersection and the results of the comparison between the intersection link set and the internal links of the requested article are provided. Additionally, the detailed results are presented in a table and a graph. Where the red coloured boxes mean that a link is missing in the requested article and there is no article in the analysed language version about that concept. The yellow boxes mean that a link is missing although there would be a corresponding article in the language edition. These findings can be caused by the fact that this information is actually missing within the article but also that the information is already part of this article but was not linked so far. We only want to provide this information and give the user the right to determine if these findings should result in a concrete action item. The green coloured boxes are the third case. These links are part of the intersection set and were found in the request article, too.

Though, LEA provides an overview of concepts that should definitely be part of an article. It offers hints about probably missing information or links in a requested article. The user gets further advice about missing articles. Some articles exist in other Wikipedia language versions and therefore seem to be important for that topic.

In future work the results of the LEA tool could be combined with other thematic coverage analysis approaches for example by applying a Named Entity Recognition. This approach allows the computation of relevant information of a topic excluding the processing of the article content.



Please enter an article title and choose a language version.

Title: Flensburg in en .wikipedia.org Submit

The article Flensburg exists in 52 further language versions.

Requested language version (en)	256 links
The following 3 language versions contain the most links to other Wikipedia articles:	
Фленсбург (ru)	390 links
Flensburg (de)	385 links
Flensborg (da)	184 links
Intersection	27 links
🟥 no article	1 links
🟨 not linked	2 links
🟩 linked	24 links

Link to this analysis

distribution of common link usage

English	Dansk	Русский	Deutsch
–	Mørvig	Мюрвик	Mürwik
Above mean sea level	Meter over havets overflade	Высота над уровнем моря	Höhe über dem Meeresspiegel
List of mayors of Copenhagen	Overborgmester	Обер-бургомистр	Oberbürgermeister
Angeln	Angel	Ангельн	Angeln (Region)
Danfoss	Danfoss	Danfoss	Danfoss

Figure 4 Screenshot LEA results for the English article “Flensburg”

3.1.2 Further approaches

We plan further analysis to recognize named entities in comparison to other language versions of one Wikipedia article with the help of JSI's Enrycher and DBPedia Spotlight[8] which will be available in a multilingual version very soon.

We are working on further approaches in collaboration with JSI to analyze the thematic coverage of a Wikipedia article compared to news articles that are related. So, we will be able to find more information about one topic which is still missing in an article. For this analysis, we will use both – the NewsFeed and the Enrycher component combined.

In addition, we are in contact with the CoSyne project[9]. In this FP7 research project the fact recognition and comparison between Wikipedia language versions are developed as part of their system. We are currently planning a meeting to discuss possible collaborations between our projects.

3.2 Analysis approaches and tools concerning timeliness

Timeliness is necessary for thematic coverage and diversity in Wikipedia. If an article is missing information about recent related events in the world, it lacks an important and essential part of the topic.

[8] http://wiki.dbpedia.org/spotlight

[9] http://cosyne.eu/

We follow two approaches to check for timeliness in Wikipedia articles. On the one hand, we are analysing the editing process in Wikipedia articles, as described in the following subsection. We use information of external sources, in particular news articles on the other hand for the examination.

3.2.1 Change Detector

The Change Detector[10] shows high editing activity in Wikipedia on a certain day and compares Wikipedias in different language versions. It enables the user to identify outdated articles in certain languages.

This is a tool to explore the timeliness of Wikipedia articles by observing the edits in different language versions. For the analysis, we compare the edits of the current day with the edits of the last 50 days for each Wikipedia article in every language version. We consider several additional factors, such as the weighting of bot edits, minor edits, and the number of unique authors. After providing a reference language the user is supported in the form of candidate articles that may need an update, triggered and ranked by amount and type of changes in other languages. An overview of the algorithm details is presented in Figure 5.

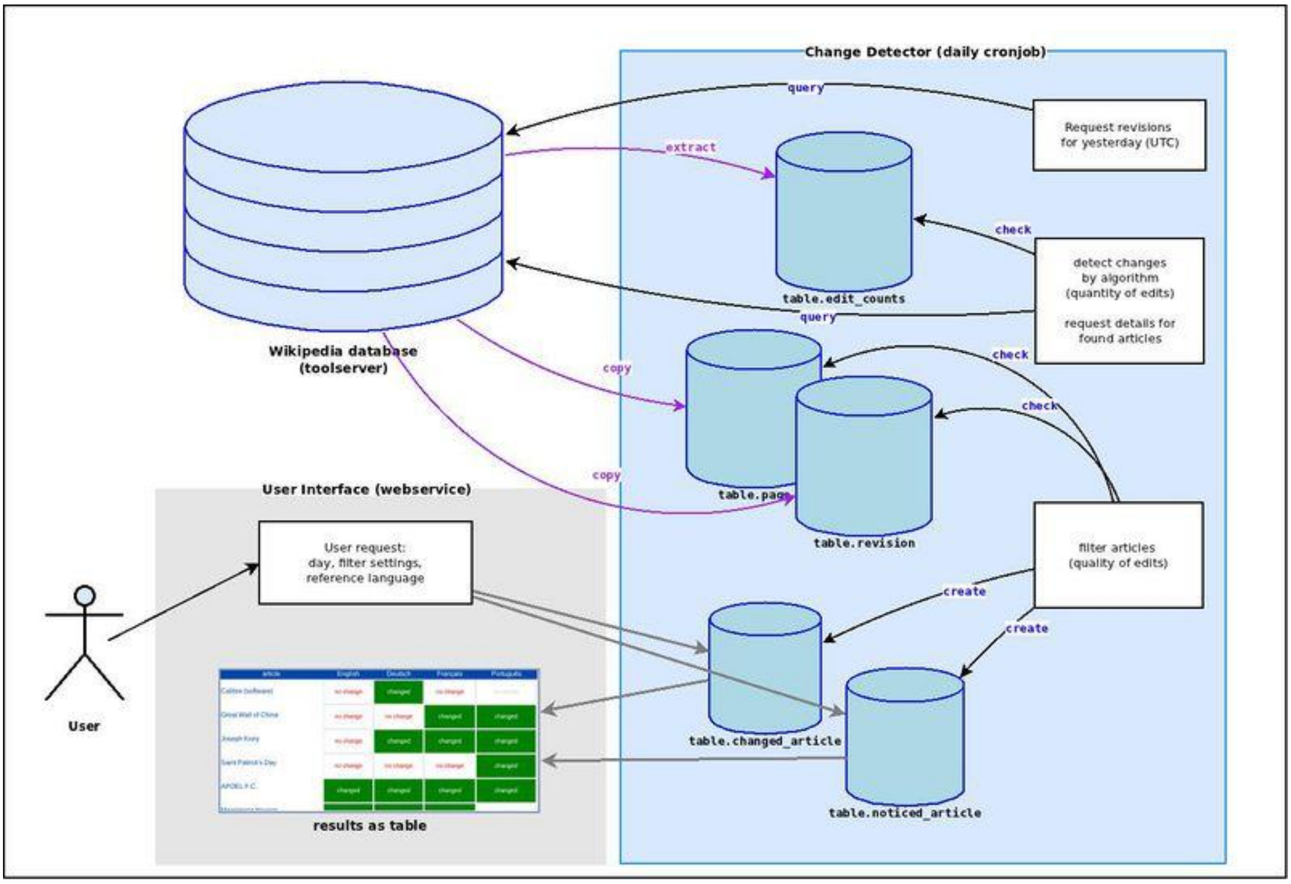
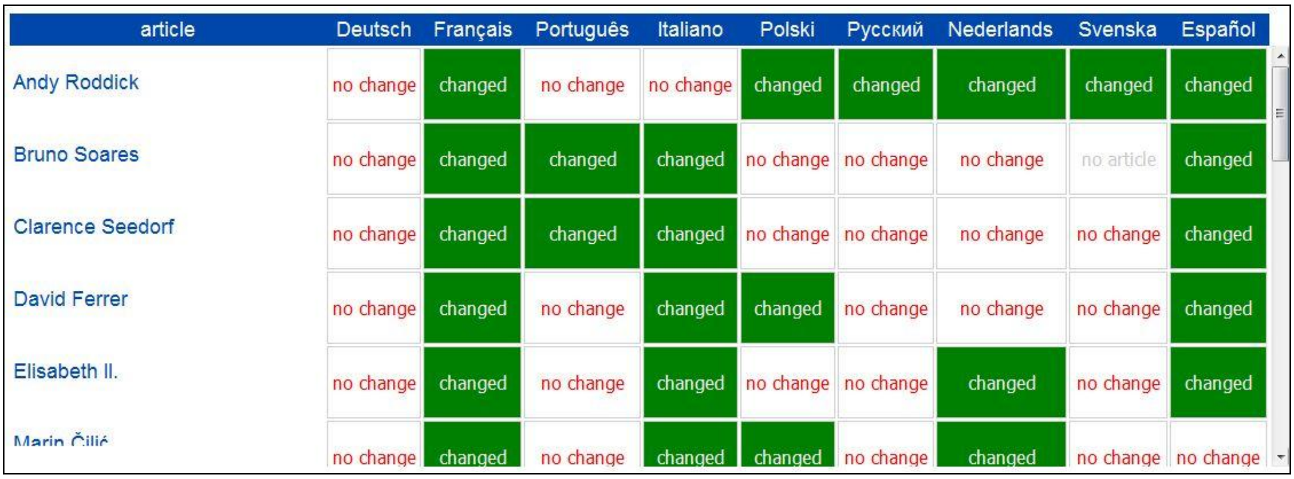

Change Detector (daily cronjob)
query
Request revisions for yesterday (UTC)
extract
check
detect changes by algorithm (quantity of edits)
request details for found articles
table.edit_counts
query
Wikipedia database (toolserver)
copy
check
check
copy
table.page
table.revision
User Interface (webservice)
User request: day, filter settings, reference language
filter articles (quality of edits)
create
create
User
table.changed_article
table.noticed_article
results as table

Figure 5 Functionality of the Change Detector Tool

The results of the daily analysis are presented in a table, as shown Figure 6. This table is ordered by the necessity of updating. Articles which were not edited in the requested language but in the most other language versions the algorithm found significant changes are sorted on the top of the table.

[10] http://toolserver.org/~RENDER/toolkit/ChangeDetector/



article	Deutsch	Français	Português	Italiano	Polski	Русский	Nederlands	Svenska	Español
Andy Roddick	no change	changed	no change	no change	changed	changed	changed	changed	changed
Bruno Soares	no change	changed	changed	changed	no change	no change	no change	no article	changed
Clarence Seedorf	no change	changed	changed	changed	no change	no change	no change	no change	changed
David Ferrer	no change	changed	no change	changed	changed	no change	no change	no change	changed
Elisabeth II.	no change	changed	no change	changed	no change	no change	changed	no change	changed
Marin Čilić	no change	changed	no change	changed	changed	no change	changed	no change	no change

Figure 6 Change Detector result table

The Change Detector cannot give details about the changed content, but it provides advices that something seems be happened without being edited in articles of a requested language version.

In future work the results of the Change Detector tool will be combined with another timeliness analysis, comparing the timestamps of news articles about a topic to the times of changes made recently in a Wikipedia article.

3.2.2 NewsFinder

An additional approach to analyse timeliness in Wikipedia is the NewsFinder. We going to provide a show case called NewsFinder as part of the RENDER toolkit. This tool is still under construction. A user can specify a Wikipedia article and a language. This service is going to request JSI's NewsFeed API (see section 0) and will present a list of news articles which are related to the requested Wikipedia article. The user can check for more information by following the links to the news articles.

Analysis approaches and tools concerning neutrality

Currently, we are following two approaches to check for non-neutral patterns within a Wikipedia article. Both approaches are still in the development process. So, we are going to present the general ideas in the following section.

On the one hand side, we are looking for words or expressions which are characteristic for a certain political party. Subsection 3.2.3 will give some more details. On the other side, JSI is investigating in machine learning algorithms which use data sets of articles which were tagged with a neutrality template, see subsection 3.2.4. These neutrality data sets can be downloaded from the RENDER Toolkit download page, as mentioned above.

3.2.3 Political Bias

To determine if a Wikipedia article contains a political bias, we are going to analyse parliamentary minutes which were provide by the The Open Knowledge Foundation. This 200 MB large data set which is pre-annotated contains additional tags like speaker or party.

We are going to identify the most frequent words[11], bi- and trigrams for each party. During the analysis process in Wikipedia article, we will display a warning if a significant number of such items of one party occur. First, this analysis is initially only for German, since we have access to this data set. But, the method to identify prototypical utterances should be easily applied to other languages. These developments are dependent on the availability and the quality of political text material in other languages.

[11] Exclude functional words like pronouns, conjunctions or prepositions

3.2.4 Opinionated Article

The opinionated Wikipedia articles dataset consist of articles which have the neutrality template[12] set by editors. There are two versions for each article – before and after setting the neutrality template. We are currently analysing opinionated articles in the English Wikipedia: 20,630 opinionated articles (out of a total of 18,719,338 articles)[13]. Most articles (17,845) have two neutrality tag changes (once the template was set, and once unset), while the article on the September 11 attacks has 38 neutrality tag changes.

Using the Diversity services (Enrycher) [3] we have identified opinionated topics based on the DMOZ[14] hierarchy of topics, named entities and part-of-speech tags. Additionally, we have registered article reference changes as well as links to related articles. All these features have been collected with the purpose of employing them in a machine learning setting. We are also investigating possibilities of transferring the acquired knowledge of which Wikipedia article is opinionated to other languages such as German.

3.3 Analysis approaches and tools concerning editor behaviour and interaction

3.3.1 WikiGini

Articles in Wikipedia are usually edited by a larger amount of users over time, including professionals in the area of expertise, usual users who modify minor sections or correct words and sentence structure, bots which automatically adding content, or even anonymous users introducing vandalism to the article. While the actual content of the article is visible and discussions are public available on the talk pages, the responsibility for the content remains hidden in the edit history of an article.

The focus of WikiGini[15] is this responsibility for text parts in Wikipedia articles. WikiGini is an application to measure and analyse the change of ownership of text in an article over time. As a measure for ownership the Gini coefficient shows the inequality in the distribution of ownership in a text. Figure 7 shows as an example the historical development of the Gini coefficient of the article "George W. Bush" for the first 1000 revisions. A Gini coefficient of 1 means total inequality, one author owns all the text in the article, while a Gini coefficient of 0 indicates a very equal distribution, all authors own the same amount of words. Easy to spot is for example vandalism, the Gini coefficient increases in a particular revision to 1 meaning one author deleted all the text or replaced it by own content.

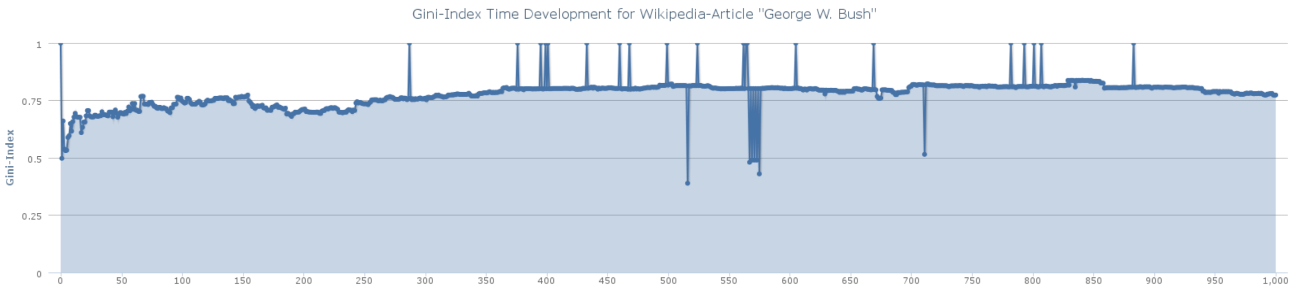


Figure 7 Gini coefficient diagram of the article "George W. Bush" for the first 1000 revisions

WikiGini introduces a new approach to calculate the ownership of text in a particular revision of an article which tends to work more accurate than prior developments. In contrast to approaches that calculate the differences between revisions based on a global hash tag for each revision the WikiGini approach uses the full revisions history on different levels of granularity for the detection of authorship changes. From paragraph level the analysis steps down to sentence and word level if necessary. Modifications are

[12] http://en.wikipedia.org/wiki/Wikipedia:Neutrality_templates

[13] The counts exclude Wikipedia infrastructure articles

[14] http://www.dmoz.org

[15] http://toolserver.org/~RENDER/toolkit/WIKIGINI

considered different depending on the modification scenario, taking into account for example the restore of former text parts by an editor which will be assigned back to the old original author independent from the amount of revisions between the two modifications.

A high level analysis of the current status of an article is already supported by the Gini coefficient. It represents a key figure of an article which can be compared with other articles. It also shows events in the history of an article like vandalism or the unusual increase or decrease of ownership in an article. Ownership data generated by WikiGini is subject of further research, e.g. in clustering approaches to identify groups of users representing a same opinion in an article, user groups who start to modify an article and thus taking ownership and possible introduce own opinions, or in contrast groups who may be tired of edit wars and stop modifying the articles, letting others opinions taking overhand. Wikipedia's statement to represent a neutral point of view is leading the further research and development of WikiGini, to find, analyse and visualize unusual editing behaviour in the Wikipedia which may contradict this neutral point of view.

3.4 Further approaches and tools concerning aspects of diversity

The following approaches cover further aspects of diversity which are in relation to cultural and language aspects in Wikipedia.

3.4.1 Corpora Explorer (Corpex)

Wikipedia is available in a large number of languages, reaching from major language versions – in terms of number of articles – like the English, German, or French Wikipedia to rather small language versions. Articles are edited, reviewed and corrected over a long time by various authors, thus leading to a relatively high quality of the text regarding the amount of linguistic mistakes, e.g. typos, or wrong sentence structure. In the aggregate a Wikipedia language version represents a large corpus of the particular language of rather good quality. While this hypothesis still remains unproven, the corpus itself can be extracted from the articles. In contrast to available corpora for widely-used languages in particular the corpora for less widely-used languages without established corpora may be a valuable source for further research.

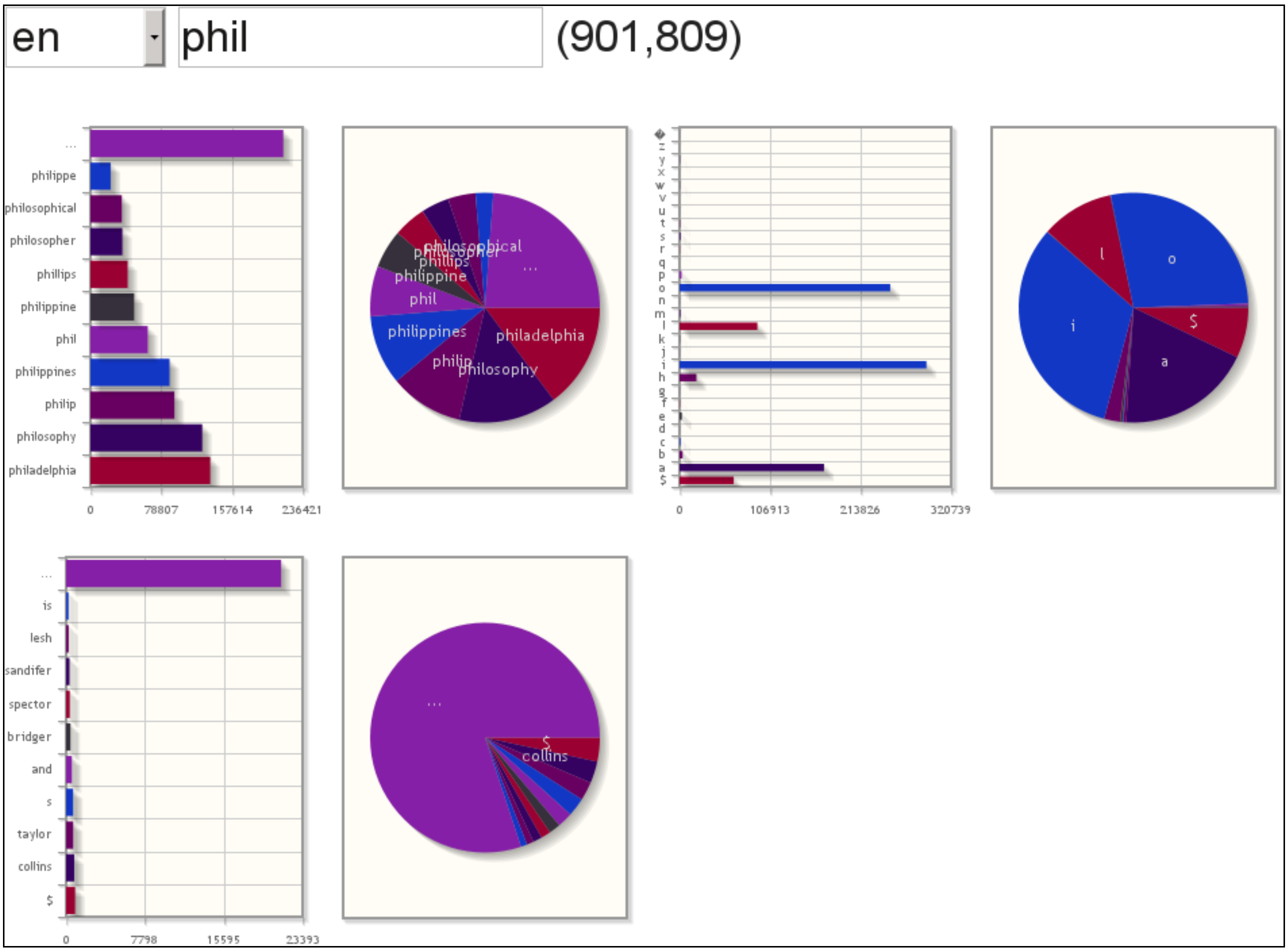


Figure 8 Corpex statistics for the input sequence "phil"

Corpex[16] generates n-gram corpora for different languages from Wikipedia. One and two grams are current available for 16 languages: Albanian, Bosnian, Bulgarian, Croatian, Czech, English, French, German, Hungarian, Italian, Romanian, Serbian, Serbo-Croatian, Simple English, Spanish, and Swedish. Corpex provides additional visualizations that allow users to browse through the available corpora. Different statistics are generated on the fly for a given input string and presented as diagrams. Figure 8 shows these statistics for the input sequence “phil”. The first row shows the ten most frequent words that start with the sequence of letters (as a bar chart and a pie chart) and the most frequent letter following the sequence of letters (again, as a bar chart and a pie chart).Both charts in the second row show the most frequent second word of all two-word-terms that start with the typed sequence as first word.

3.4.2 Wikipedia Map

A large amount of articles in Wikipedia contains location information in form of geo-coordinates. It specifies the location related to the content, in other words the locations “Wikipedia talks about”. This information is a valuable source in itself but also aggregated to a whole. The Wikipedia Map[17] is a showcase visualizing all geo-located articles of a particular language in a world map. In the world map each geo-location is highlighted by a light dot. All highlighted articles form – dependent on amount of articles in the chosen language – the contour of the world map.

Wikipedia Map is based on structured data extracted by the DBpedia[18] project. Static maps are generated from each supported language in different solutions as shown in Figure 9 for all articles in the English Wikipedia. In addition the map is also available in a dynamic version. It allows users to zoom in and out and

[16] http://toolserver.org/~RENDER/toolkit/Corpex/
[17] http://toolserver.org/~RENDER/toolkit/WikiMap
[18] http://dbpedia.org

to get additional information about particular geo-locations. The dynamic version is as well working standalone, based purely on HTML and JavaScript techniques.

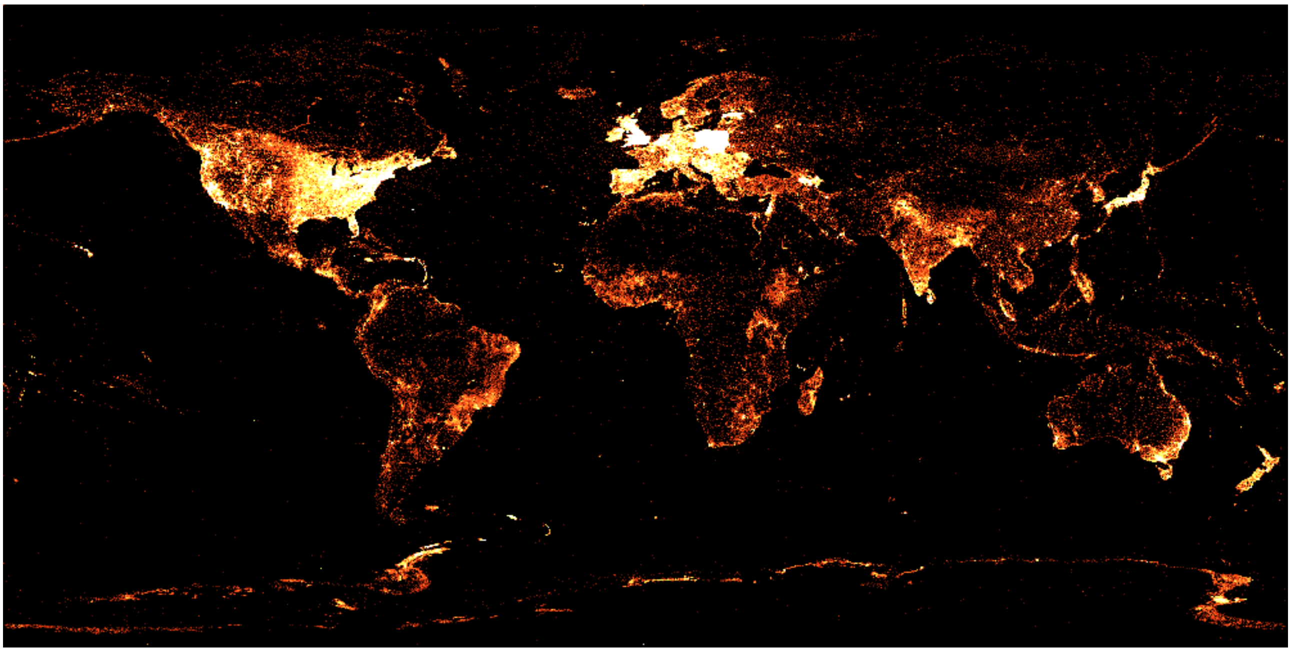


Figure 9 Wikipedia Map for all geo-located articles in the English Wikipedia

The visualization provides on the one hand information about the coverage of real world entities by Wikipedia articles from a geographical point of view. By comparing different languages regions can be identified that are covered in a higher granularity and by more information in form of articles respectively in a particular language and thus point out potential for knowledge transfer between languages. For example the Catalan Wikipedia covers in-depth Catalonia in Spain, covers several parts of Europe as well as the USA, but provides only weak coverage of other parts of the world. On the other hand data errors or abnormalities are revealed, e.g. many French articles are grouped at the zero meridians. Future improvements of the Wikipedia Map Interface will first emerge efforts by Wikimedia on better support for geographical information extraction.

4 Additional components and data

In this chapter we describe two services – CatGraph and JSI's NewsFeed. Their output is necessary for our analysis and supporting tools.

4.1 CatGraph

The CatGraph Component is a tool on the Wikimedia Toolserver which is used to search within a category graph. It is necessary to identify articles belonging to a certain category or the intersection of two categories. CatGraph allows a quick access to the Wikipedia graph structure for efficient searching of categories.

4.1.1 Graph Processor (CatGraph)

The Graph Processor project aims to develop an infrastructure for rapidly analysing and evaluating Wikipedia's category structure. Wikipedia page IDs are stored in large directed graphs in memory. The German Wikipedia graph currently holds about 6.8 million arcs (node-to-node relationships). The implementation makes things possible which cannot be done in a reasonable amount of time in SQL. For example, finding all leaf nodes (pages without successors, i.e. non-category pages) in the German Wikipedia graph takes less than 20 seconds. A query for the category 'Sport' with recursion depth 8 results in 178136 nodes and executes in less than 2 seconds. The output of this request is shown in Figure 10.

```
traverse-successors 235635 8
OK. 178136 nodes, 1.150790s:
235635
112105
236011
236583
236584
465432
566447
826737
863820
960100
1590190
2236893
[...]
```

Figure 10 CatGraph output

4.1.2 Components of the CatGraph

CatGraph consists of several components (see Figure 11) which are described in following:

GraphCore: maintains and processes large directed graphs in memory. Each instance runs as a UNIX process and models the category structure of one Wikipedia language.

GraphServ: The server process handles access to running GraphCore instances, multiplexes commands and data, and implements session handling and access control.

Client Libraries: Programming libraries[19] exist for Python and PHP to allow application programmers easy access to the CatGraph interface.

[19] https://svn.toolserver.org/svnroot/daniel/duesenstuff/trunk/gpClient/

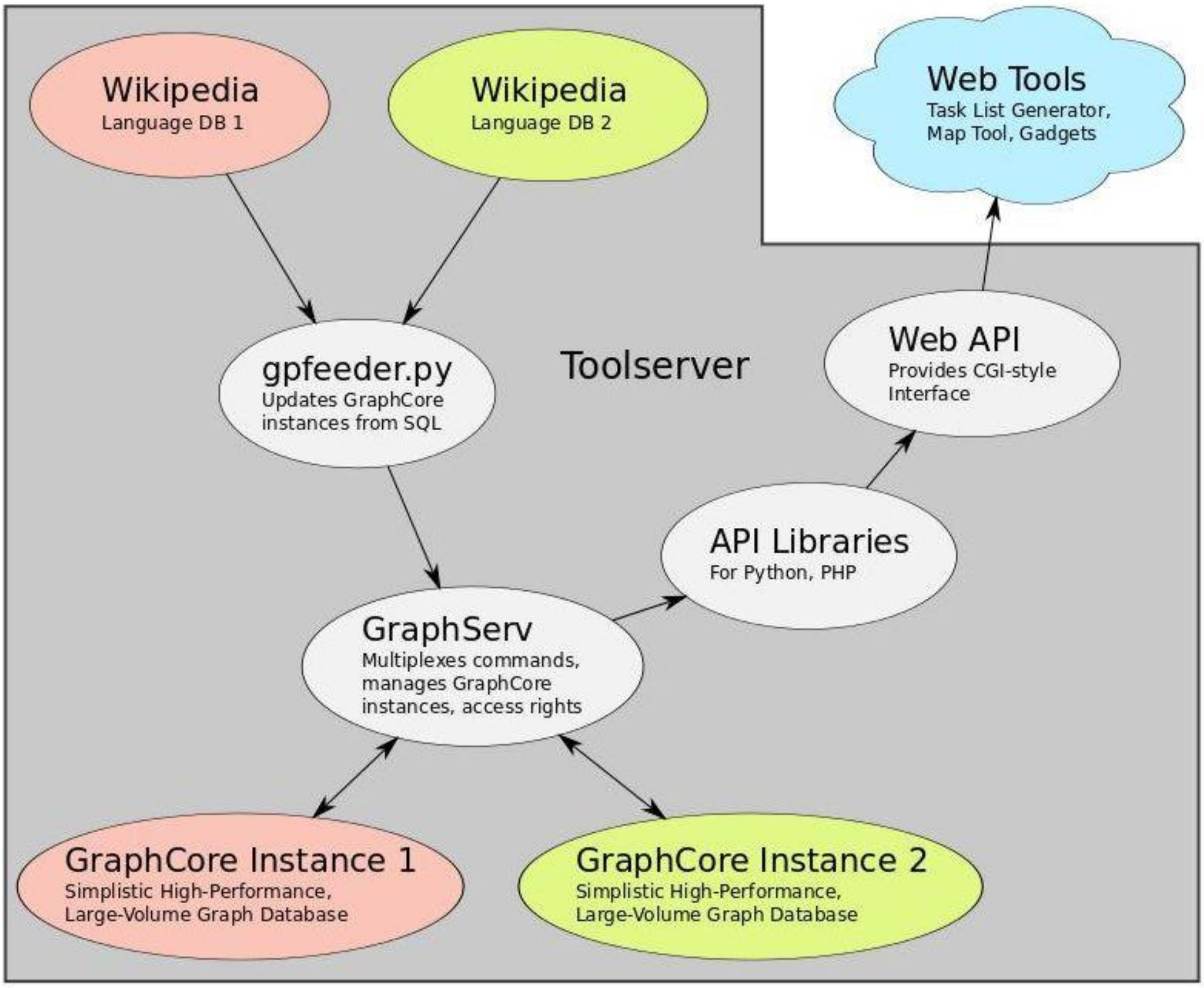
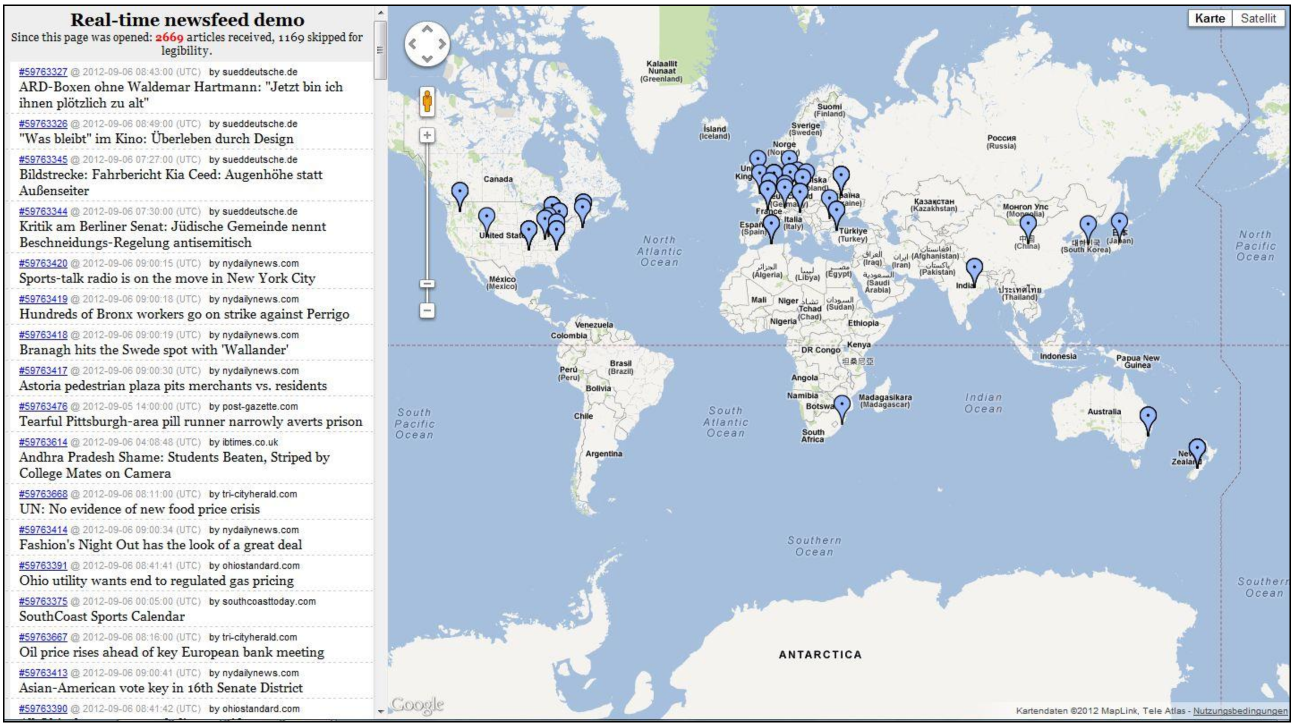

Wikipedia
Language DB 1
Wikipedia
Language DB 2
Web Tools
Task List Generator, Map Tool, Gadgets
gpfeeder.py
Updates GraphCore instances from SQL
Toolserver
Web API
Provides CGI-style Interface
API Libraries
For Python, PHP
GraphServ
Multiplexes commands, manages GraphCore instances, access rights
GraphCore Instance 1
Simplistic High-Performance, Large-Volume Graph Database
GraphCore Instance 2
Simplistic High-Performance, Large-Volume Graph Database

Figure 11 CatGraph overview

NewsFeed

The NewsFeed[20] incorporates a number of services for collecting, indexing and querying news in different languages. The NewsFeed provides a real-time aggregated stream of textual news items delivered by RSS-enabled news sources across the world. The initial version of the tool was described in more detail in D 1.3.2 - Initial corpora collection (English, French, German, Italian, and Spanish) [2]. Within the Wikipedia use case, we use this service to find news articles which are related to Wikipedia article to analyse the timeliness metric.

[20] http://newsfeed.ijs.si



Real-time newsfeed demo

Since this page was opened: 2669 articles received, 1169 skipped for legibility.

#59763327 @ 2012-09-06 08:43:00 (UTC) by sueddeutsche.de
ARD-Boxen ohne Waldemar Hartmann: "Jetzt bin ich ihnen plötzlich zu alt"

#59763326 @ 2012-09-06 08:49:00 (UTC) by sueddeutsche.de
"Was bleibt" im Kino: Überleben durch Design

#59763345 @ 2012-09-06 07:27:00 (UTC) by sueddeutsche.de
Bildstrecke: Fahrbericht Kia Ceed: Augenhöhe statt Außenseiter

#59763344 @ 2012-09-06 07:30:00 (UTC) by sueddeutsche.de
Kritik am Berliner Senat: Jüdische Gemeinde nennt Beschneidungs-Regelung antisemitisch

#59763420 @ 2012-09-06 09:00:15 (UTC) by nydailynews.com
Sports-talk radio is on the move in New York City

#59763419 @ 2012-09-06 09:00:18 (UTC) by nydailynews.com
Hundreds of Bronx workers go on strike against Perrigo

#59763418 @ 2012-09-06 09:00:19 (UTC) by nydailynews.com
Branagh hits the Swede spot with 'Wallander'

#59763417 @ 2012-09-06 09:00:30 (UTC) by nydailynews.com
Astoria pedestrian plaza pits merchants vs. residents

#59763476 @ 2012-09-05 14:00:00 (UTC) by post-gazette.com
Tearful Pittsburgh-area pill runner narrowly averts prison

#59763614 @ 2012-09-06 04:08:48 (UTC) by ibtimes.co.uk
Andhra Pradesh Shame: Students Beaten, Striped by College Mates on Camera

#59763668 @ 2012-09-06 08:11:00 (UTC) by tri-cityherald.com
UN: No evidence of new food price crisis

#59763414 @ 2012-09-06 09:00:34 (UTC) by nydailynews.com
Fashion's Night Out has the look of a great deal

#59763391 @ 2012-09-06 08:41:41 (UTC) by ohiostandard.com
Ohio utility wants end to regulated gas pricing

#59763375 @ 2012-09-06 00:05:00 (UTC) by southcoasttoday.com
SouthCoast Sports Calendar

#59763667 @ 2012-09-06 08:16:00 (UTC) by tri-cityherald.com
Oil price rises ahead of key European bank meeting

#59763413 @ 2012-09-06 09:00:41 (UTC) by nydailynews.com
Asian-American vote key in 16th Senate District

#59763390 @ 2012-09-06 08:41:42 (UTC) by ohiostandard.com

Figure 12 Screenshot of the real-time NewsFeed demo

5 Supporting Tool for Wikipedia Users

We combine the results of our analysis approaches (see section 2) with results of further assessments tools and Wikipedias quality assurance methods. This information is the input for two different tools we have developed to support the work of established Wikipedia authors on the one hand and to help readers to understand the quality and the status of a Wikipedia article.

These tools are aimed to enable Wikipedia users:

- To track biased and one-sided articles,
- To understand and improve them.

This will be achieved by identifying faults or lacking information in articles or sections of an article. Furthermore, we are going to provide further sources that are likely to contain the missing data or show more information related to the topic.

Two supporting tools:

- Article Statistics and Quality Monitor (ASQM): This application displays the diversity analysis results for every article. This tool will include the possibility to check for different metrics and provide a quick overview of the quality and the state of an article.
- Task List Generator (TLG): This application enables a Wikipedia author to generate lists of articles related to a specific topic or preferred category, which need to be improved.

Both supporting tools can be easily expended if further analysis algorithms are available or in case of Wikipedia users request for further functionalities. The whole source code is under a free licence and available for downloading on the Toolserver[21].

The following subsections give a detailed description of these tools.

5.1 Article Statistics and Quality Monitor

The Article Statistics and Quality Monitor is aimed to support in particular readers, to get a quick overview about a Wikipedia article by providing statistical data, details of diversity analysis and further assessment scores from external[22] tools. In addition, a reader will be empowered to better judge the quality of an article and can understand the collaborative editing process Wikipedia. This may motivate them to edit this Wikipedia article themselves.

5.1.1 Functionality

Currently, the ASQM covers the following statistics, RENDER metrics and further assessment tools:

Statistics:

- The title of the article is accessed from the Wikimedia API
- The status of the article: This metric is shown if an article is good or featured and the date if the nomination and is accessed from the Wikipedia API
- The date of creation and the name of the first editor: This information is retrieved from the data bases on the toolserver
- The date of the last edit and the user name or IP-address: This information is retrieved from the data bases on the toolserver

[21] https://svn.toolserver.org/svnroot/p_render

[22] means not developed within the RENDER consortium

- The number of editors: This information is retrieved the Wikipedia API
- The number of references in the article: This metric is computed by counting the number of ref-tags within the wiki text which is retrieve from the MediaWiki API
- The total number of pictures: This metric is retrieved form the Toolserver data base
- The number of accesses today and during the last 30 days: These metrics we get by requesting the Wikipedia article traffic statistics[23].

RENDER diversity metrics:

- Fact coverage: Currently, we provide a link to the LEA result page for this article in a given language.
- Timeliness: Here, the number of news articles is shown which the NewsFeed (JSI) calculated. In addition, we provide a link to the result list.
- Neutrality: For this metrics is currently no information shown. It will be expanded if the analysis results for opinionated articles (see 3.2.4) and political bias (see 3.2.3) are available.
- Editor interaction: Here, we are going to provide the Gini index as result of the WikiGini tool analysis for one article. This information is currently still missing but can be easily added if the information is available.

Further analysis tools and metrics:

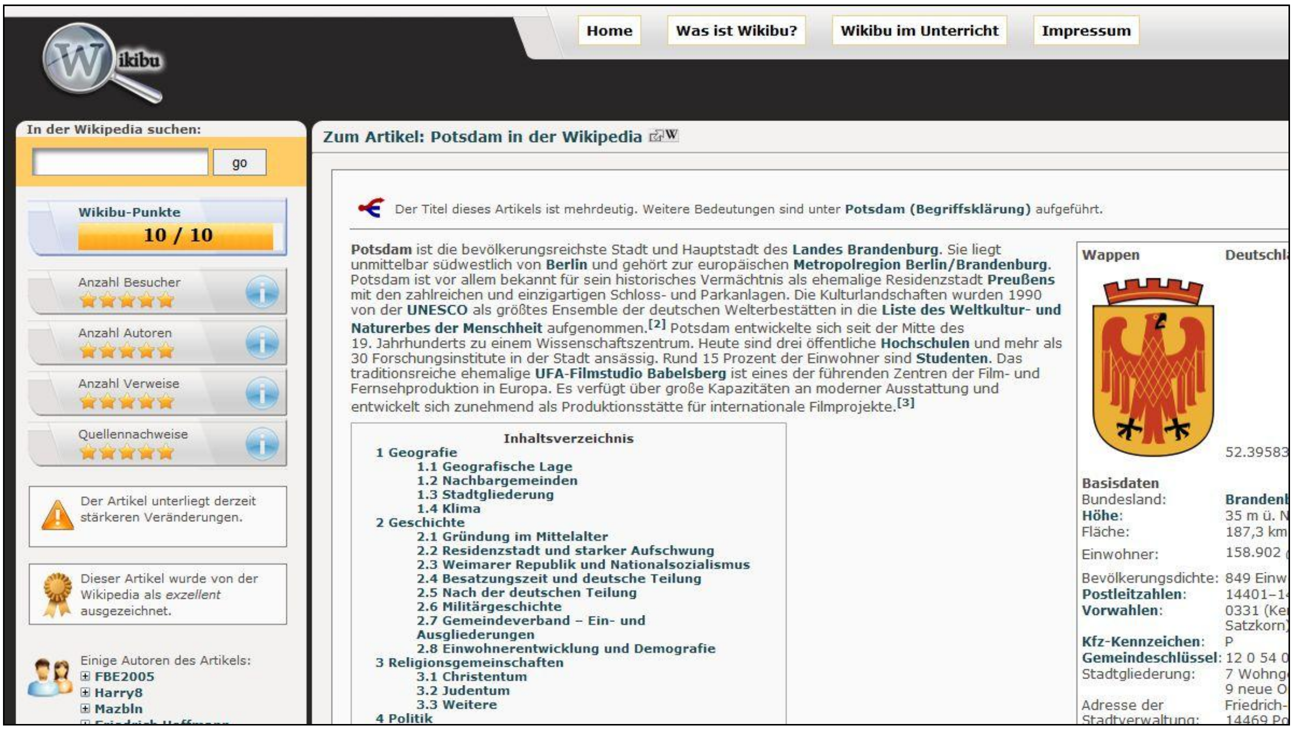
These metrics can change in order to a requested language version. Currently, we present the following tool results:

- The Article Feedback Tool Scores: Currently, these metrics are extracted for the English Wikipedia articles by accessing the MediaWiki API. There will be some changes during the next months; the interface of this tool will be totally changed. So we will adapt the new version which will be rolled out in the English Wikipedia by the end of this year and tested for different language versions at the beginning of next year[24]. Thus, we will define different metrics to represent these assessment results.
- The Wikibu.ch[25] Score: This score is calculated by using different data like number of links, editors, and sources. This tool is only available for German. So we provide the score and a link to the analysis result page, as shown in Figure 13.

[23] http://stats.grok.se/

[24] http://blog.wikimedia.org/2012/06/25/converting-readers-into-editors-new-results-from-article-feedback-v5/

[25] http://wikibu.ch



Der Artikel unterliegt derzeit stärkeren Veränderungen.

Dieser Artikel wurde von der Wikipedia als *exzellent* ausgezeichnet.

Einige Autoren des Artikels:
⊞ FBE2005
⊞ Harry8
⊞ Mazbln

Zum Artikel: Potsdam in der Wikipedia

Der Titel dieses Artikels ist mehrdeutig. Weitere Bedeutungen sind unter **Potsdam (Begriffsklärung)** aufgeführt.

Potsdam ist die bevölkerungsreichste Stadt und Hauptstadt des **Landes Brandenburg**. Sie liegt unmittelbar südwestlich von **Berlin** und gehört zur europäischen **Metropolregion Berlin/Brandenburg**. Potsdam ist vor allem bekannt für sein historisches Vermächtnis als ehemalige Residenzstadt **Preußens** mit den zahlreichen und einzigartigen Schloss- und Parkanlagen. Die Kulturlandschaften wurden 1990 von der **UNESCO** als größtes Ensemble der deutschen Welterbestätten in die **Liste des Weltkultur- und Naturerbes der Menschheit** aufgenommen.[2] Potsdam entwickelte sich seit der Mitte des 19. Jahrhunderts zu einem Wissenschaftszentrum. Heute sind drei öffentliche **Hochschulen** und mehr als 30 Forschungsinstitute in der Stadt ansässig. Rund 15 Prozent der Einwohner sind **Studenten**. Das traditionsreiche ehemalige **UFA-Filmstudio Babelsberg** ist eines der führenden Zentren der Film- und Fernsehproduktion in Europa. Es verfügt über große Kapazitäten an moderner Ausstattung und entwickelt sich zunehmend als Produktionsstätte für internationale Filmprojekte.[3]

Inhaltsverzeichnis

- 1 Geografie
  - 1.1 Geografische Lage
  - 1.2 Nachbargemeinden
  - 1.3 Stadtgliederung
  - 1.4 Klima
- 2 Geschichte
  - 2.1 Gründung im Mittelalter
  - 2.2 Residenzstadt und starker Aufschwung
  - 2.3 Weimarer Republik und Nationalsozialismus
  - 2.4 Besatzungszeit und deutsche Teilung
  - 2.5 Nach der deutschen Teilung
  - 2.6 Militärgeschichte
  - 2.7 Gemeindeverband – Ein- und Ausgliederungen
  - 2.8 Einwohnerentwicklung und Demografie
- 3 Religionsgemeinschaften
  - 3.1 Christentum
  - 3.2 Judentum
  - 3.3 Weitere
- 4 Politik

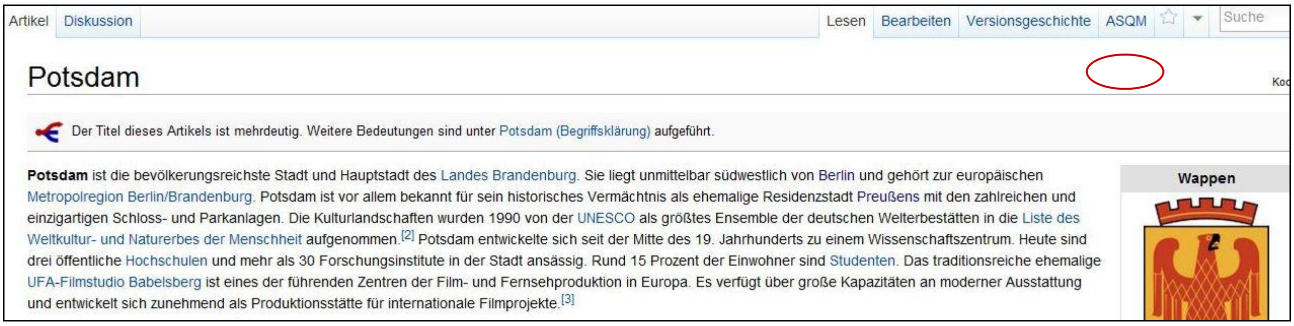
Wappen	Deutschl
	52.39583
Basisdaten	
Bundesland:	**Brandenl**
Höhe:	35 m ü. N
Fläche:	187,3 km
Einwohner:	158.902 (
Bevölkerungsdichte:	849 Einw
Postleitzahlen:	14401–1
Vorwahlen:	0331 (Ke Satzkorn)
Kfz-Kennzeichen:	P
Gemeindeschlüssel:	12 0 54 0
Stadtgliederung:	7 Wohng 9 neue O
Adresse der Stadtverwaltung:	Friedrich- 14469 Po

Figure 13 Screenshot of wikibu.ch - result for the article "Potsdam"[26]

5.1.2 Installation and usage

We established an information page on the Toolserver[27]. There one can access the code snippet and further instructions how to install this gadget within the preference settings of Wikipedia. Each user has to do this process for him. As part of the evaluation process, we decided to track the number of installations. Each time the information page is requested we generate a unique global ID within the gadget code. Thus, we can get the information about the installation but without checking out personal user information like IP-address or user name.

If the installation was successful a page within the Wikipedia article namespace will contain a new tab "ASQM" within the task bar. Figure 14 provides an example of the ASQM inclusion within the article page about the German city "Potsdam".



Potsdam

Der Titel dieses Artikels ist mehrdeutig. Weitere Bedeutungen sind unter Potsdam (Begriffsklärung) aufgeführt.

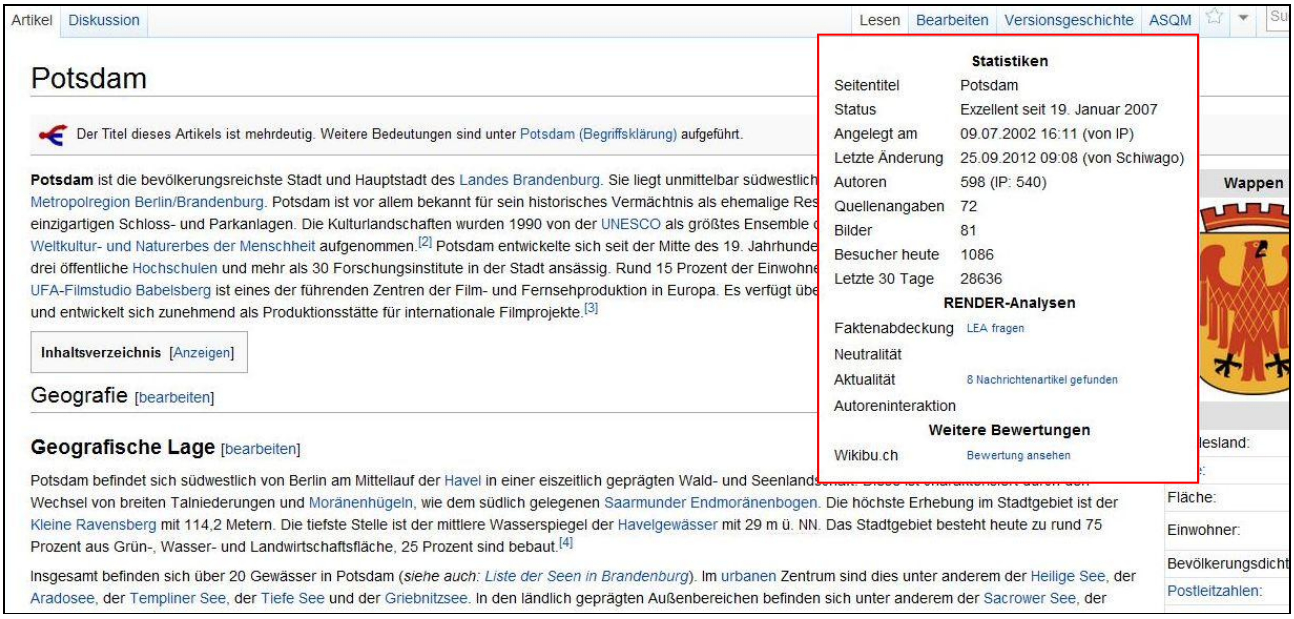
Potsdam ist die bevölkerungsreichste Stadt und Hauptstadt des Landes Brandenburg. Sie liegt unmittelbar südwestlich von Berlin und gehört zur europäischen Metropolregion Berlin/Brandenburg. Potsdam ist vor allem bekannt für sein historisches Vermächtnis als ehemalige Residenzstadt Preußens mit den zahlreichen und einzigartigen Schloss- und Parkanlagen. Die Kulturlandschaften wurden 1990 von der UNESCO als größtes Ensemble der deutschen Welterbestätten in die Liste des Weltkultur- und Naturerbes der Menschheit aufgenommen.[2] Potsdam entwickelte sich seit der Mitte des 19. Jahrhunderts zu einem Wissenschaftszentrum. Heute sind drei öffentliche Hochschulen und mehr als 30 Forschungsinstitute in der Stadt ansässig. Rund 15 Prozent der Einwohner sind Studenten. Das traditionsreiche ehemalige UFA-Filmstudio Babelsberg ist eines der führenden Zentren der Film- und Fernsehproduktion in Europa. Es verfügt über große Kapazitäten an moderner Ausstattung und entwickelt sich zunehmend als Produktionsstätte für internationale Filmprojekte.[3]

Figure 14 ASQM included as gadget in the task bar

If a user clicks on the ASQM-tab a small window with further information is opening within the article page, as displayed in the red framed box in Figure 15. There a user can get a very fast overview about the most important facts to the article and links to further analysis results and webpages. Thus, she can decide which information are most important and which additional information could be used to get a better

[26] http://wikibu.ch/search.php?search=Potsdam
[27] http://toolserver.org/~render/stools/asqm

understanding of this topic or to be inserted in the content. We do not want to assess an article, but we provide a sum of information out of Wikipedias metadata and additional external sources.



Potsdam

Der Titel dieses Artikels ist mehrdeutig. Weitere Bedeutungen sind unter Potsdam (Begriffsklärung) aufgeführt.

Potsdam ist die bevölkerungsreichste Stadt und Hauptstadt des Landes Brandenburg. Sie liegt unmittelbar südwestlich
Metropolregion Berlin/Brandenburg. Potsdam ist vor allem bekannt für sein historisches Vermächtnis als ehemalige Res
einzigartigen Schloss- und Parkanlagen. Die Kulturlandschaften wurden 1990 von der UNESCO als größtes Ensemble
Weltkultur- und Naturerbes der Menschheit aufgenommen.[2] Potsdam entwickelte sich seit der Mitte des 19. Jahrhunde
drei öffentliche Hochschulen und mehr als 30 Forschungsinstitute in der Stadt ansässig. Rund 15 Prozent der Einwohne
UFA-Filmstudio Babelsberg ist eines der führenden Zentren der Film- und Fernsehproduktion in Europa. Es verfügt übe
und entwickelt sich zunehmend als Produktionsstätte für internationale Filmprojekte.[3]

Inhaltsverzeichnis [Anzeigen]

Geografie [bearbeiten]

Geografische Lage [bearbeiten]

Potsdam befindet sich südwestlich von Berlin am Mittellauf der Havel in einer eiszeitlich geprägten Wald- und Seenlands
Wechsel von breiten Talniederungen und Moränenhügeln, wie dem südlich gelegenen Saarmunder Endmoränenbogen. Die höchste Erhebung im Stadtgebiet ist der Kleine Ravensberg mit 114,2 Metern. Die tiefste Stelle ist der mittlere Wasserspiegel der Havelgewässer mit 29 m ü. NN. Das Stadtgebiet besteht heute zu rund 75 Prozent aus Grün-, Wasser- und Landwirtschaftsfläche, 25 Prozent sind bebaut.[4]

Insgesamt befinden sich über 20 Gewässer in Potsdam (*siehe auch: Liste der Seen in Brandenburg*). Im urbanen Zentrum sind dies unter anderem der Heilige See, der Aradosee, der Templiner See, der Tiefe See und der Griebnitzsee. In den ländlich geprägten Außenbereichen befinden sich unter anderem der Sacrower See, der

Statistiken	
Seitentitel	Potsdam
Status	Exzellent seit 19. Januar 2007
Angelegt am	09.07.2002 16:11 (von IP)
Letzte Änderung	25.09.2012 09:08 (von Schiwago)
Autoren	598 (IP: 540)
Quellenangaben	72
Bilder	81
Besucher heute	1086
Letzte 30 Tage	28636
RENDER-Analysen	
Faktenabdeckung	LEA fragen
Neutralität	
Aktualität	8 Nachrichtenartikel gefunden
Autoreninteraktion	
Weitere Bewertungen	
Wikibu.ch	Bewertung ansehen

Figure 15 ASQM output for the German article “Potsdam”

5.2 Task List Generator

The Task List Generator is a tool which is aimed at supporting in particular established Wikipedia authors. With this tool editors can generate lists of articles that need to be improved, sorted by category or fields of interest editors concerning problems of the content.

5.2.1 Functionality

Currently, the TLG covers different filters which are listed below and shown in Figure 16. Besides RENDER metrics and quality analysis filters, we also take several Wikipedia maintenance templates into account. These templates have been inserted by reviewing editors and are related to our defined diversity aspects – neutrality, timeliness and thematic coverage.

Wikipedia maintenance templates:

- Cleanup Template: this filter finds articles which contain the Cleanup template[28]. This template displays that an article has general problems and has to be cleaned up.
- 'Too Technical' Template: this filter finds articles which contain the Technical template[29]. This template is used to mark an article as “too technical for the most readers to understand”.
- Out of Template: this filter finds articles which contain the Out of date template[30]. This template is used to signal that an article seems to need an update.
- Globalize Template: this filter finds articles which contain the Globalize template[31]. This template displays that an article “may not represent a worldwide view”.

[28] http://en.wikipedia.org/wiki/Template:Cleanup
[29] http://en.wikipedia.org/wiki/Template:Technical
[30] http://en.wikipedia.org/wiki/Template:Out_of_date
[31] http://en.wikipedia.org/wiki/Template:Globalize

- Missing Sources/References Template: this filter finds articles which contain the Refimprove Template[32]. This template is used to signal an article needs citations for verification.
- Neutrality Template: this filter finds Wikipedia articles which contain the Neutrality template[33]. This template is added if an article seems to dispute the NPOV[34].

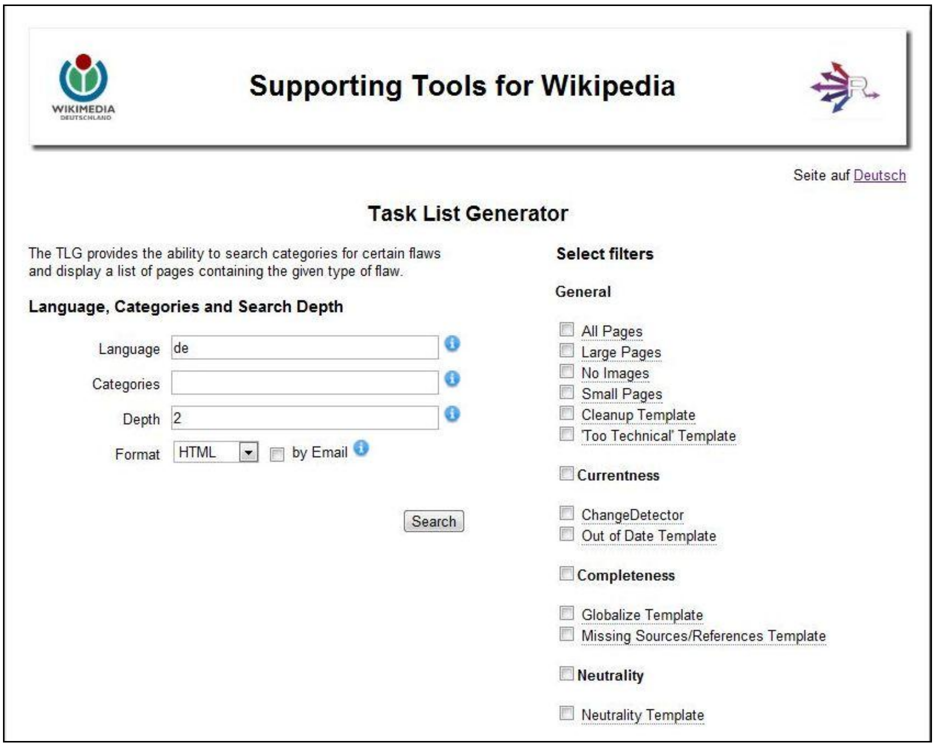
Analysis Filters:

- All Pages: if this filter is activated all pages of a requested category or a requested intersection of categories are shown in the result list.
- Large Pages: This filter finds pages which size pass a threshold calculated using the mean page size and standard deviation of all pages in the same category.
- Small Pages: this filter finds all Wikipedia articles which are ¼ shorter than the mean length in the same category.
- No Images: this filter finds Wikipedia articles which contain no image links.

RENDER analysis filters:

- Change Detector: this filter finds Wikipedia articles which are part of the result list of the daily Change Detector analysis

Further RENDER analysis approaches will be inserted during the next months.



Task List Generator

The TLG provides the ability to search categories for certain flaws and display a list of pages containing the given type of flaw.

Language, Categories and Search Depth

Language: de

Categories:

Depth: 2

Format: HTML □ by Email

Select filters

General

□ All Pages
□ Large Pages
□ No Images
□ Small Pages
□ Cleanup Template
□ 'Too Technical' Template

□ Currentness

□ ChangeDetector
□ Out of Date Template

□ Completeness

□ Globalize Template
□ Missing Sources/References Template

□ Neutrality

□ Neutrality Template

Figure 16 Screenshot of the the TLG interface in German

The Task List Generator backend currently uses CatGraph (see section 3.1) to get a quick access to the Wikipedia graph structure for efficient searching of categories. The result is an article list for the requested category or the intersection of categories. In a second step these list is checked for these flaws the user specified in the form. At this current stage TLG is available for German and English.

[32] http://en.wikipedia.org/wiki/Template:Refimprove

[33] http://en.wikipedia.org/wiki/Template:Neutrality

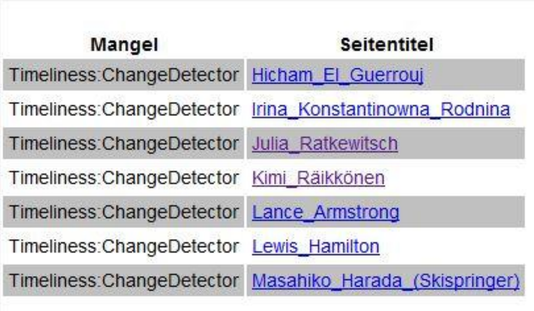
[34] http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

5.2.2 Usage

Currently, this tool is available on the Wikipedia Toolserver[35]. A user can specify within this form a language, a category or the intersection of categories; select a flaw and the search depth within the category tree. Furthermore, the users can choose between HTML or wiki text as output formats and can decide if they want to get the results on demand or via email, as shown in Figure 16. shows a screenshot of the tool interface on the Wikimedia toolserver.

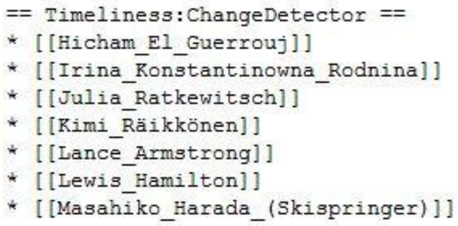
In a first step, the Task List Generator backend uses CatGraph (see section 4.1) to get a quick access to the Wikipedia graph structure for efficient searching of categories. The result is an article list for the requested category or the intersection of categories. In a second step this list is checked for the flaws the user specified in the form.

Figure 17 and Figure 18 show the output list for the category "Weltmeister" and the flaw metric timeliness – results from the Change Detector analysis.



Mangel	Seitentitel
Timeliness:ChangeDetector	Hicham_El_Guerrouj
Timeliness:ChangeDetector	Irina_Konstantinowna_Rodnina
Timeliness:ChangeDetector	Julia_Ratkewitsch
Timeliness:ChangeDetector	Kimi_Räikkönen
Timeliness:ChangeDetector	Lance_Armstrong
Timeliness:ChangeDetector	Lewis_Hamilton
Timeliness:ChangeDetector	Masahiko_Harada_(Skispringer)

Figure 17 TLG - HTML output



```
== Timeliness:ChangeDetector ==
* [[Hicham_El_Guerrouj]]
* [[Irina_Konstantinowna_Rodnina]]
* [[Julia_Ratkewitsch]]
* [[Kimi_Räikkönen]]
* [[Lance_Armstrong]]
* [[Lewis_Hamilton]]
* [[Masahiko_Harada_(Skispringer)]]
```

Figure 18 TLG - wiki text output

[35] http://toolserver.org/~render/stools/tlg

6 Evaluation plans for the supporting tools

We are going to test the usage and the functionality of the supporting tools during the evaluation process. Each supporting tool is aimed to support a specific target groups. So, the test groups are different, too.

Assuming a positive evaluation result and a reasonable number of users the supporting tools could become a regular part of Wikipedia as a MediaWiki extension.

6.1 Article Statistics and Quality Monitor (ASQM)

The ASQM tool is aimed at supporting in particular readers. We described the functionality and the benefits in section 5.1. This tool empowers the users in understanding the status and the quality of a Wikipedia article and provides advices if patterns of bias in a certain direction are detected.

We are going to evaluate the usage of ASQM in a quantitative and a qualitative way.

To assess the acceptance and usage qualitatively, we are going to measure the number of gadget installations and requests. In addition, we are going to analyse the major articles and topics for which ASQM was requested.

Besides the quantitative measuring, we are going to process a qualitative evaluation with a small group of 15 – 20 readers. These are users who have never edited in Wikipedia before. We are going to work very close together with our department of *Education and Knowledge*[36] to reach this user group and during the evaluation period. Our colleagues are organising multiple workshops to teach different user groups about writing and using Wikipedia.

We are going to support and to guide the test users during the installing process and the test phase. Each test user will be requested to perform the ASQM tool with a number of articles. During this testing several questions concerning the usability, the performance, the correctness, and the usage of these results have to be answered in a questionnaire. Additionally, we are going to collect further suggestions and feedback comments during this process.

6.2 Task List Generator (TLG)

The Task List Generator is aimed at supporting the group of established authors to cure articles belonging to a certain category or the intersection of categories as we described in section 5.2.

We are going to evaluate TLG quantitatively and qualitatively.

To assess the acceptance and usage qualitatively, we are going to measure the number of list requests. In addition, we are going to analyse the flaws and categories which are mainly requested.

During a qualitative evaluation we are going to test the TLG with two small groups of authors in Wikipedia. At the beginning, we are focused on the German community and in particular the so called Redaktionen. These are groups of authors which work together in a certain thematic field like biology, medicine or history. We request these authors to test the TLG for 3 – 4 weeks within their daily work in Wikipedia and to answer a questionnaire concerning the usability, the performance, and the correctness of single diversity and flaw analysis results. Additionally, we are going to collect further suggestions and feedback comments during this process.

More information about the evaluation plans and a description on collecting best practices and experiences in addition from developers are part of D4.2.1 section 2 [4].

[36] http://wikimedia.de/wiki/Bildung

7 Summary and future work

Within the Wikipedia use case study we aimed to increase the quality of Wikipedia. To reach this ambitious goal we are going to empower the Wikipedia users, readers and authors in finding, understanding and curing biased and non-neutral articles.

In this document we gave a description of the tools to manage different aspects of diversity in Wikipedia we realised so far. In addition, we discussed some approaches to examine diversity which we are going to finish during the next weeks.

The analysis tools are presented as a show case in the RENDER Toolkit on the Wikimedia Toolserver.

Currently the Toolkit contains 5 analysis visualisations:

- LEA: a tool which finds missing internal links in an article compared to other language versions
- Change Detector: a tool which finds out-dated article in one language version by analysing the edits compared to other language versions
- WikiGini: a tool which visualises the analysis of the change of text ownership in an article over time
- Wikipedia Map: a tool which visualises geo-tagged Wikipedia articles in one language version as a map
- Corpex: a tool for exploration corpora extracted from Wikipedia in several different languages. It provides visualizations for frequencies in a corpus, in particular the most frequent words, letters, and short two word terms.

The RENDER Toolkit is in a contiguous process of development and extension. There will be further analysis tools like the NewsFinder and tools for political bias and opinionated article identification in Wikipedia during the next month.

Furthermore, we provide the data sets extracted during the project on the download page. So, these data can be re-used by other researchers and institutes.

The results of each analysis approach are the input for the supporting tools which aimed to support editors and readers to understand and cure bias in Wikipedia article content.

- The Article Statistics and Quality Monitor: is aimed to support in particular readers, to get a quick overview about a Wikipedia article by providing statistical data, details of diversity analysis and further assessment scores from further tools.
- The Task List Generator: is aimed at supporting in particular established Wikipedia authors. Editors can generate lists of articles that need to be improved, sorted by category or fields of interest.

Both supporting tools have to be tested on usability and usage aspects during a target group specific evaluation process. There we have to check for extensions related to requirements and needs of certain user groups like a WikiProjects. There it is indispensable to respect the privacy and wishes of the Wikipedia community. With these tools we are going to provide a framework which can be extended and improved during and in particular after the duration of the RENDER project. So, we hope to conserve the project results by establishing these instruments within the Wikimedia universe.

References

[1] Angelika Adam, Fabian Flöck, Gerrit Holz. RENDER Deliverable D 5.1.1 – Definition and Evaluation of Metrics in Wikipedia. 2011

[2] Denny Vrandecic, Philipp Sorg, Rudi Studer, Delia Rusu, Mitja Trampus, Blaz Novak, Mariana Damova. RENDER Deliverable D 1.3.2 - Initial corpora collection (English, French, German, Italian, and Spanish). 2012.

[3] Delia Rusu, Mitja Trampus, Inna Novalija, Tadej Stajner, Mariana Damova. RENDER Deliverable D2.2.2 – Final Version of the Fact Mining Toolkit. 2012.

[4] Ioan Toma, Javier Caminero, Angelika Adam, Delia Rusu, Enrique Alfonseca, Fabian Flöck, Elena Simperl. RENDER Deliverable D4.1.2 – Collecting best practices for diversity-aware collaboration. 2012.