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

RENDER

Deliverable D5.1.3

Understanding the Feedback-Effects of Metrics on Wikipedia

Editor:	Angelika Mühlbauer, Wikimedia
Author(s):	Angelika Mühlbauer, Wikimedia; Johannes Kroll, Wikimedia; Kai Nissen, Wikimedia
Deliverable Nature:	Report (R)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	30 September 2013
Actual Delivery Date:	30 September 2013
Suggested Readers:	Wikimedia community and researchers
Version:	1.0
Keywords:	Wikipedia, diversity, aspects of quality, increase of visibility

Disclaimer

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

In case of Public (PU):
All RENDER consortium parties have agreed to full publication of this document.

In case of Restricted to Programme (PP):
All RENDER consortium parties have agreed to make this document available on request to other framework programme participants.

In case of Restricted to Group (RE):
The information contained in this document is the proprietary confidential information of the RENDER consortium and may not be disclosed except in accordance with the consortium agreement. However, all RENDER consortium parties have agreed to make this document available to <group> / <purpose>.

In case of Consortium confidential (CO):
The information contained in this document is the proprietary confidential information of the RENDER consortium and may not be disclosed except in accordance with the consortium agreement.

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:	WP5 - Diversity Case Studies
Document Title:	D5.1.3 - Understanding the Feedback-Effects of Metrics on Wikipedia
Editor (Name, Affiliation)	Angelika Mühlbauer, Wikimedia
Work package Leader (Name, affiliation)	Javier Caminero, Telefónica

Copyright notice

Executive Summary

Wikipedia is driven by its users – the contributors but also the readers. Although its content is written by multiple editors the quality in many thematic fields is high and stable caused by the many-eyes principle. For Wikimedia Deutschland e.V. two major aspects of diversity are relevant in the context of this use case. On the one-hand side our aim is to offer and collect diverse information throughout different language versions of Wikipedia. On the other side by attracting people to edit Wikipedia by summarizing different metrics about an article or giving suggestions about what may be missing or outdated. So we hope to increase the diversity of authors.

In this deliverable we presented the result and findings of the analysis of the feedback the RENDER tools effect on Wikipedia. The results can only provide first insights. The combination of logging results and user assessment to the ability of the tools helps us to understand the influences on Wikipedia and its users.

Table of Contents

List of Figures

List of Tables

Abbreviations

ALG	Article List Generator
AM	Article Monitor
R1	Request 1
R2	Request 2
WP	Wikipedia

1 Introduction

Wikipedia is driven by its users – the contributors but also the readers. Although its content is written by multiple editors the quality in many thematic fields is high and stable caused by the many-eyes principle. For Wikimedia Deutschland e.V. two major aspects of diversity are relevant in the context of this use case. On the one-hand side our aim is to offer and collect diverse information throughout different language versions of Wikipedia. For example the results of the thematic coverage comparison in different languages can help users to contribute and so to increase the quality. On the other side by attracting people to edit Wikipedia by summarizing different metrics about an article or giving suggestions about what may be missing or outdated. So we hope to increase the diversity of authors.

As part of the Wikipedia use case study we developed two supporting tools – the Article Monitor and the Article List Generator. Both tools aim to help Wikipedia users to understand the status of articles, to flaws but also additional information related to one topic in fast and comfortable way (for details see D5.1.2).

Besides the evaluation of the tools themselves which we described in D5.1.4 we were interested in understanding which influences the results of project will have on the Wikipedia and the behaviour of the users who include the tools in their daily work and their activities in Wikipedia.

In the section 2 we will explain the methodology we followed in this process to collect the data and insights to users activities. Additionally, we are going to present the metrics we up-dated according to the software achievements of the project. The results of our analysis will be described in section 3. In the last section we will summarize the findings and our future plans.

2 Methodology and Metrics

In this section we present how we collected the data for the analysis process of the feedback-effects and describe the metrics for the analysis.

2.1 Methodology of Data Collection and Analysis

We announced the final version of the supporting tools via mailing lists, on Wikipedia and the Wikimedia blog at the end of June. As part of this announcement we invited all users to use and test them during their daily work.

Since the participation and usage took place on a voluntary basis, we have no exact knowledge about the user profiles of the testers. But we strongly suppose that the majority of the testers were Wikipedians since the announcement channels are typically used and observed by Wikipedia editors.

Additionally, we prepared two questionnaires and shared the links to the corresponding Google forms via our communication channels. Furthermore, we contacted participants of our information events like the RENDER tour (see D6.2.5 [3]) directly to ask for participating in the surveys.

In total, 25 persons participated in the questionnaires. The majority (80 %) is very experienced in Wikipedia as presented in Figure 1.

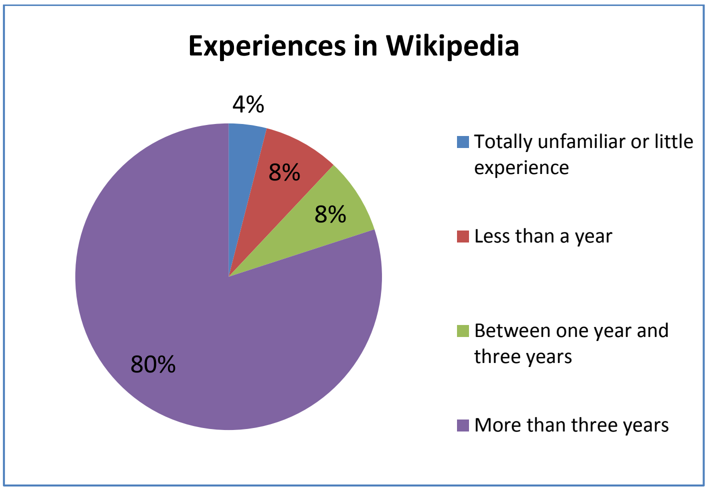


Figure 1: Distribution of the User Experience in Wikipedia

The collection took place during the evaluation and testing period of the supporting tools from June the 21st to August the 26th 2013.

In all cases we had to respect privacy aspects. We used two methods to collect information and data for the analysis and interpretation process:

Logging, Reproduction and Comparison of Request Results

To understand and observe if articles changed after the usage of the supporting tools, we logged in the first step the query strings as well as the results of the Link Extractor and the Article List Generator. In the second step we reproduced these requests with identical parameters and stored the results. Then we compared both results.

In D5.1.2 we described the Link Extractor a tool which uses internal links as one approach to analyse the thematic coverage of an article. The results are represented with help of a scale of colours - red, yellow and green. Green means a match of internal links for the three analysis language versions and the requested one. Yellow is related to the internal link is missing although the link target exists in the requested language. The colour red means the link is missing and there is no article related to this topic.

In the analysis we compare the results of the original request and the reproduced request (as visualized in Figure 2) by calculating the differences for all three result categories. An improvement would be if the editors added matches, removed missing links or created missing articles.

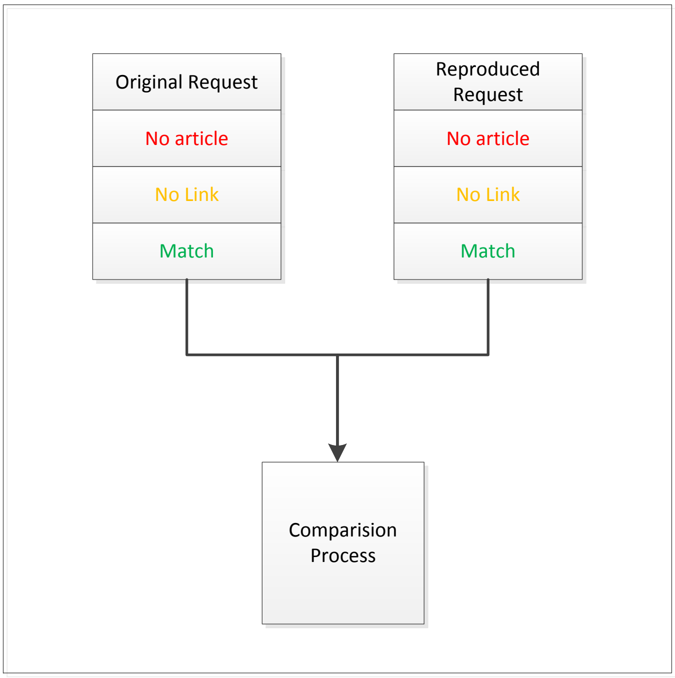


Figure 2: Comparison of the Link Extractor Results

For the ALG requests we did the analysis in a similar way. We logged the whole request parameter string and the number of the result article list as well as the number of articles within the requested search term. We calculated the number of result articles in relation to the number of articles within the search term as schematic visualized in Figure 3. That means a change in the resulting article list to a value under 100 % will be understood as an improvement, 100% as a stable situation and more than 100% as a worse result compared to the original request done by a user.

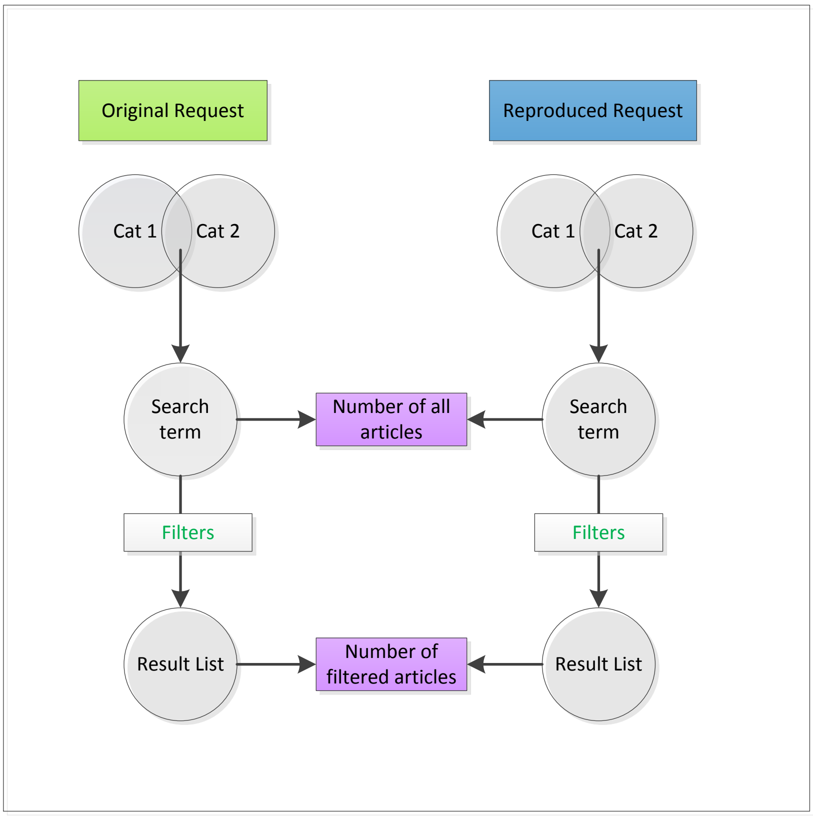


Figure 3: Comparison of ALG Requests

User Assessment to Contribute with Help of the Tools

In the second approach we used to understand the influence of the tools on Wikipedia we asked the testers in the questionnaires which task they could perform with help of the tools and analysis approaches.

In combination with the logging calculation these information might help us understand the influences, although we were not allowed to track the individual activities of single users.

2.2 Metrics for the Analysis

In D5.1.1 we described a number of metrics we planned to use to explore the influences of the RENDER developments on Wikipedia. We identified three major aspects of diversity which are motivated by the community members themself. These are core requirements for articles which could be suggested for the awarding process.

In the first deliverable we described the major aspects and identified completeness (thematic coverage), currentness, objectivity as the most important and appropriate aspects to find, to understand, and to improve the lack of knowledge diversity in Wikipedia articles, and to achieve an increase in quality. We adjusted the metrics during the project to the realized tools and analysis approaches:

Thematic Coverage: The Link Extractor analysis which is part of the Article Monitor is aimed to explore the thematic coverage of an article compared to the three largest versions of this article in other language versions. We will analyse the following metrics:

- Comparison of the number of links marked as missing before the first usage of Link Extractor and at the end of the testing period.
- Analysis of relationship between the number of requests and the change in link distribution for these articles.

Currentness: The Change Detector, the News Finder and the ALG filter Template: Out of date are aimed to call the users attention on articles which seems to be out of date or might be inspected for expansion needs in an encyclopaedic way. We will analyse the following metrics:

- The number of currentness warnings as part of the AM and the ALG
- The number of users who requests for a detailed result table to understand the warning
- The number of articles with a significant increase of editing in the first hour after the request

Neutrality: The ALG filter Template: Neutrality is aimed to call the users attention on articles which were tagged to written in a non neutral way. To measure its influence we are going to analyse the following metrics:

- The number of requests and the number of results that contained the neutrality template
- The number of articles which still contain the neutrality template after the test period

Besides these points we are going to compare the number of articles which apply to a specific ALG filter within a user request with the number at the end of the testing period.

With help of these analyses we hope to find indicators which allow us to predict the influence of our tools on Wikipedia.

3 Results and Observations

In this section we present the results and findings we collected and calculated during the testing and analysis period.

3.1 Thematic Coverage

In D5.1.4 we presented the results of the quantitative and a qualitative evaluation of the supporting tools.

The Article Monitor was installed during the testing period (21.6.-26.8.) 65 times. We observed 675 users requests of this overview tool. In 20% (136 cases) the users clicked on the offered link to get more information about the thematic coverage analysis about an article. We recognized 126 successful requests. Unsuccessful requests occurred for articles which exist in less than three additional language versions then the requested one.

For these 126 requests we reproduced the search queries and analysed if the results have been improved related to all three result types (red, yellow and green). In Figure 4 we opposed the mean values of all three types of the original (R1) and the reproduced requests (R2). The mean value for matches increased slightly as well as the number of missing links (although an article would exist) slightly decreased. The number of missing article which are visualised as red cell in the LEA result table has not changed between both request runs.

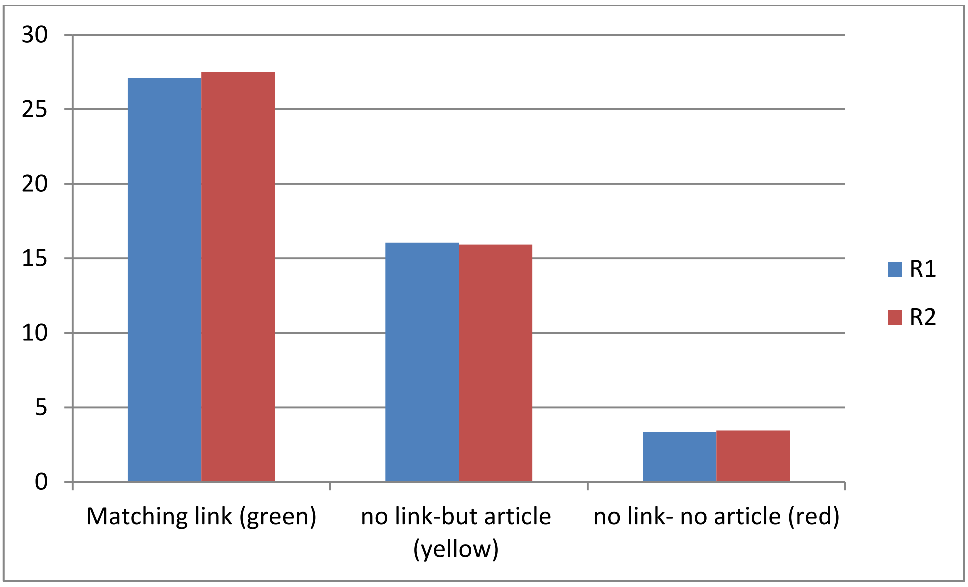


Figure 4: Results of LEA Requests - Comparison of Mean Values in R1 and R2

We analysed the data in more detail. We found that LEA was requested for 85 unique articles. Each article was requested between 1 and highest 6 times during our logging period. The results are shown **Fehler! erweisquelle konnte nicht gefunden werden.**. For 29 articles (34%) we calculated better results in the reproduced data set compared to the logged request results. The majority of articles (72%) was requested only once. For 21 of these articles we found better results in reproduced LEA analysis compared to the original user request. 12 articles contained more link matches, 11 fewer missing links and 6 fewer missing articles compared to the original request. In the majority only one of the three improvement types occurred.

Table 1: Link Extractor Analysis

Number of Requests; Number of Articles; Types of Improvement (green, yellow, red); Number of Articles with 1-3 Improvement Types

Number of Requests	Number of Articles	Type of Improvement			Positive Change in 1 type	Positive Change in 2 types	Positive Change in 3 types
		Increased number of matches	Decreased number of missing links	Decreased number of missing articles			
1	61	12	11	6	14	6	1
2	14	5	2	2	1	4	0
3	7	3	3	0	0	3	0
4	2	0	0	0	0	0	0
6	1	0	0	0	0	0	0

Additionally, we took the answers of the questionnaire into account. The users were asked to select which activities they could perform with help of Link Extractor. The summary is visualised in Figure 5.

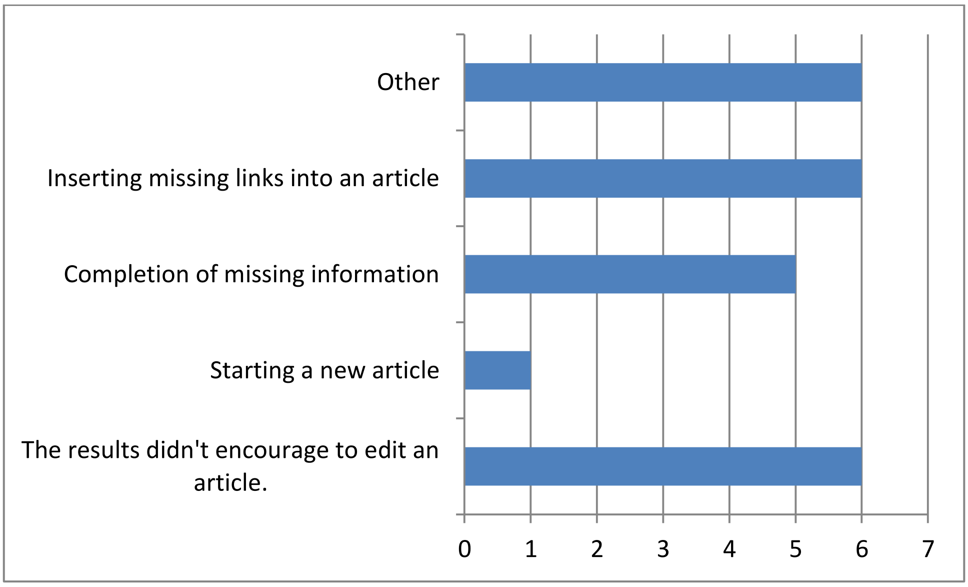


Figure 5: User assessment – Activities with Help of the Link Extractor

Six testers commented that the result encouraged them to insert missing links. In addition, 5 testers indicated that LEA supports to complete missing information and one person mentioned the possibility to start a new article. The value “The results didn't encourage to edit an article” was selected 6 times but two of these users mentioned that LEA enables the deletion of wrong links and missing information.

The answers of the tester together with the results we found in the comparison analysis leads us to a positive tendency. These results are not able to confirm our assumption that LEA causes these changes in 100% of the cases but they show us a very positive tendency.

3.2 Currentness and Neutrality

Currentness:

We logged the requests of the requests of the NewsFinder and the Change Detector out of the Article Monitor. For all logged results we checked if the articles had been edited during the first hour after the tool request to find a relation. During the testing the users requested no article the Change Detector had calculated as to be out of date. So we have no data. In similar situation occurred for the News Finder. Only in 36 article requests (5%) the News Finder could offer further information for the tested article. For 8 articles the testers asked for more information, but we could find a editing during the next hour after the request.

During the survey, 3 users assessed the News Finder and 4 testers answers the questions according to the Change Detector. Figure 6 shows the assessment for both approaches which shows that both tools are estimated as useful to complete missing information. But caused of these few results a useful interpretation is impossible.

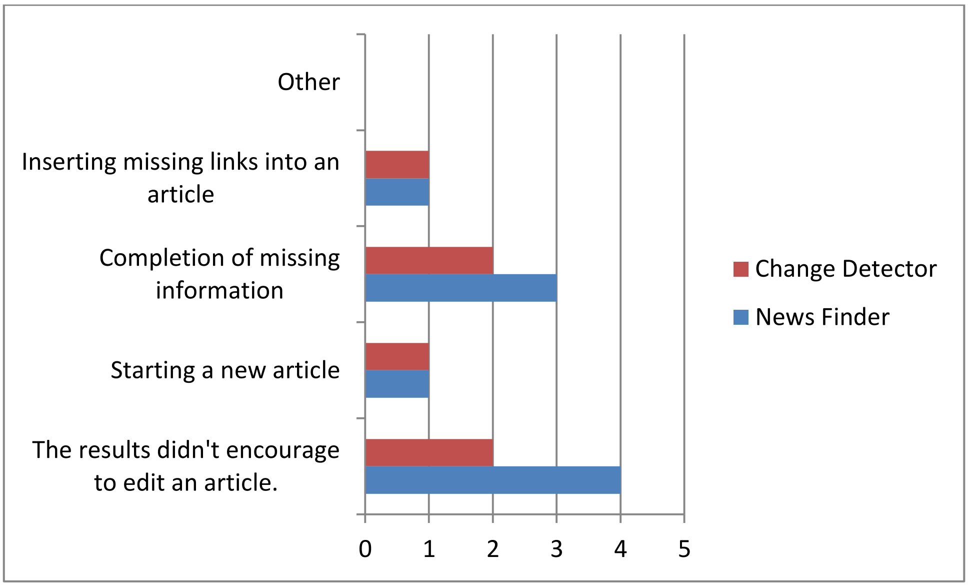


Figure 6: User assessment – Activities with Help of Change Detector and News Finder

In addition we observed the requests of the ALG for the filters Template: Out of date. In 23 cases the users searched for articles with help of this filter. Only in 6 cases a request was successful and a list of articles could be generated. For these requests we reproduced the results and found 3 cases with fewer articles containing the template but also 3 cases containing more tagged articles. Caused by this small number of articles and findings these results are not useful for an interpretation.

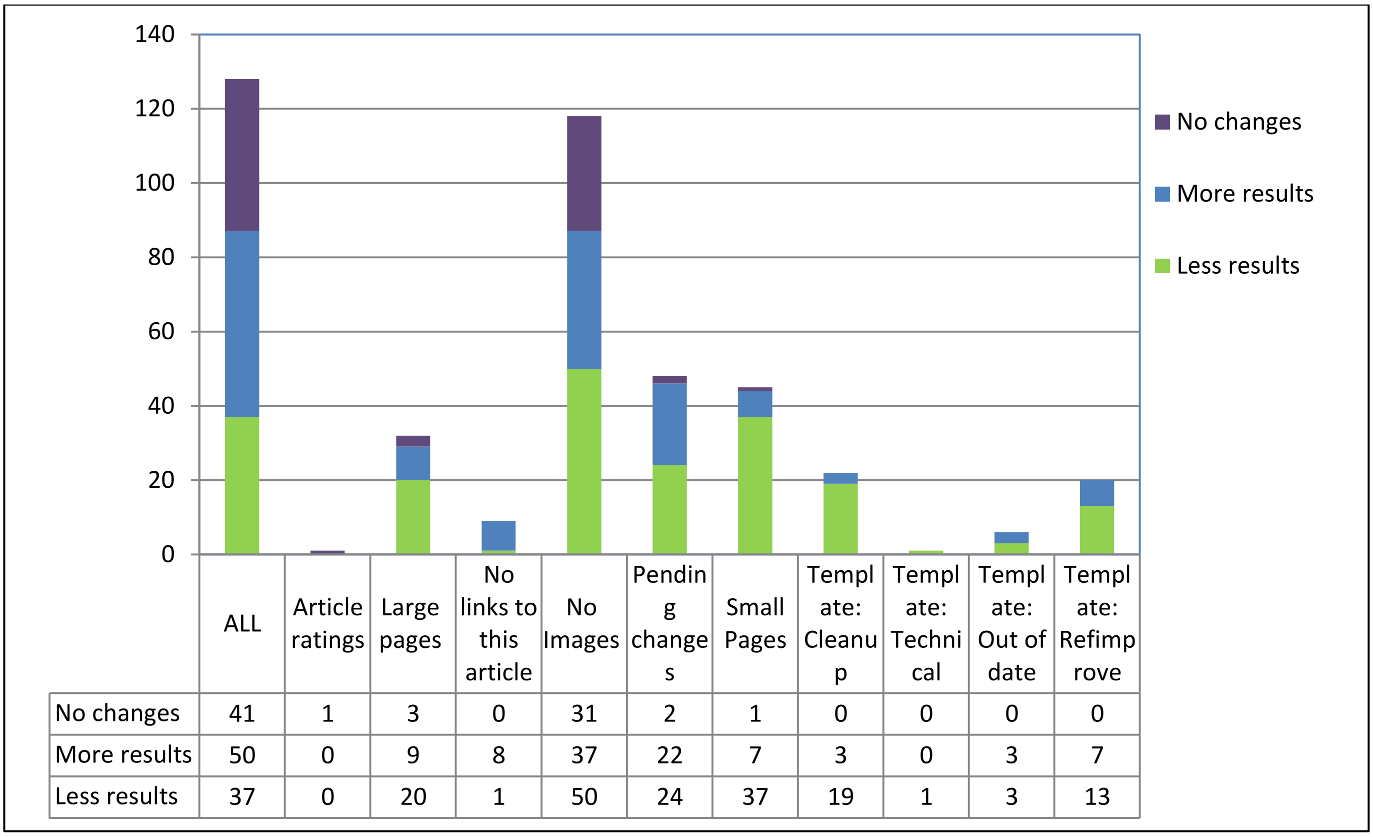
Neutrality:

We logged all requests of ALG searching which selected the Template: Neutrality filter. 23 requests could be logged but there couldn't find any article tagged with this template within the search terms.

So we have to assume about the neutrality and diversity in an indirect way. By attracting people to a certain article increases the number of eyes which check an article for flaws and increases the diversity of editors.

3.3 Further Findings

To find out if the tools were used to attract people to Wikipedia articles we logged all requests of the ALG and reproduced the findings. We calculated the increase/decrease of the number of articles in the result list in relation to the size of the article set contained in the search term. Wikipedia is growing continuously so the absolute numbers could falsify the results. We calculated the findings in percentage numbers. The results for all successful filters in the ALG are presented in Figure 7. Our calculations show that the best results could be observed for the filters *No images, Small pages, Pending changes and Template: Cleanup.*



	ALL	Article ratings	Large pages	No links to this article	No Images	Pending changes	Small Pages	Template: Cleanup	Template: Technical	Template: Out of date	Template: Refimprove
No changes	41	1	3	0	31	2	1	0	0	0	0
More results	50	0	9	8	37	22	7	3	0	3	7
Less results	37	0	20	1	50	24	37	19	1	3	13

Figure 7: Reproduction results of ALG requests - Comparison to Original Request Results

(Numbers of Results with a Better (Fewer Results), a Worse (More Results) and Stable (No Changes) Result List)

The analysis of the answers the testers gave in the questionnaires is presented in Figure 8. The majority of the users commented that the ALG helps to create missing images, but also conducting sighting and review content was mentioned by 25 % of the testers.

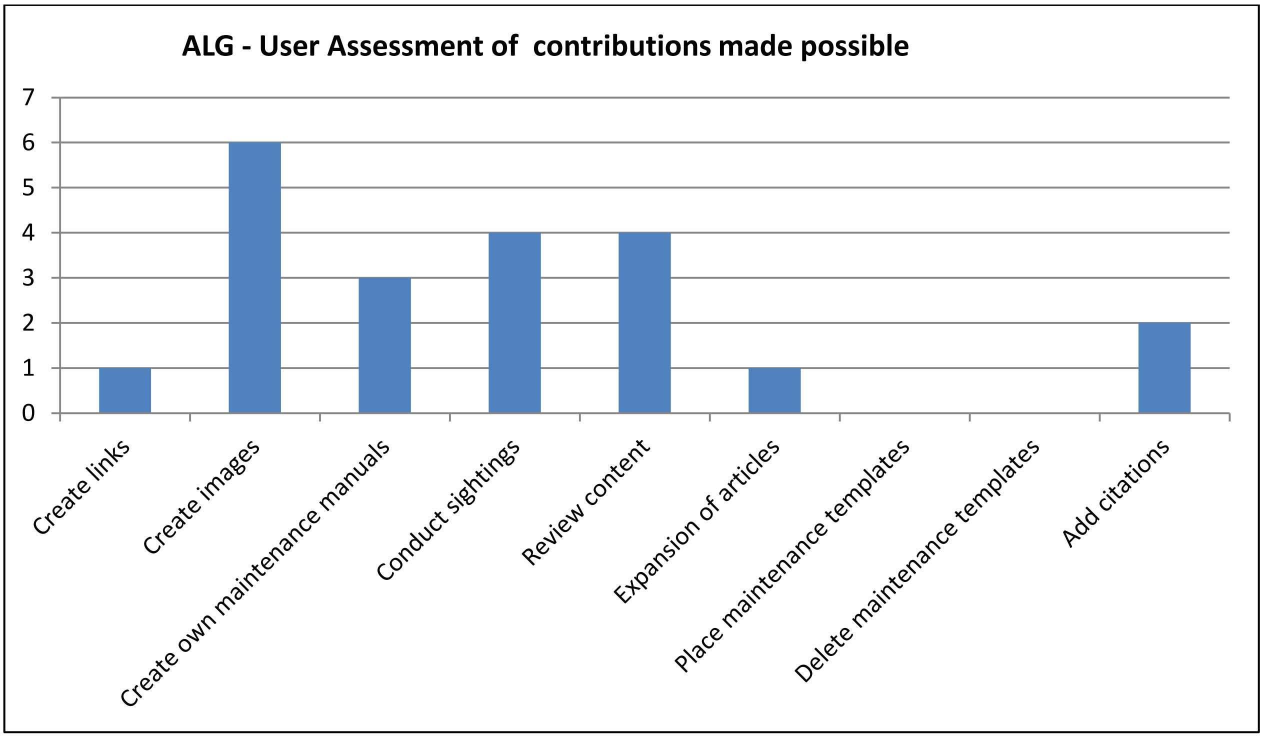


Figure 8: User Assessment – Activities with Help of Article List Generator

3.4 Discussion of Findings

Caused by these results, our observations and feedback we got during personal exchanges with users we learned that the tools we developed during the RENDER project have the power to support diversity in Wikipedia. Also, by attracting photographers to certain article we increase the visibility of articles. If she inserts a missing image to an article she might read the text and can take care of mistakes in the content. People who never had an idea where to start can use the RENDER tools. E.g. the Article List Generator is able to support established editors but also newcomers who are happy to get a starting point.

Be attracting people to use these tools and to find simple things to fix, we increase the diversity of contributors. Following the many-eyes principle this situation will lead to more quality. It is important in particular for articles which are visited not so often and are written by only a few editors.

4 Summary and Future Work

In this deliverable we presented our results of the analysis of feedback-effects on Wikipedia.

These results can only provide first but promising insights to the future improvements caused by our tools.

We described in D5.1.4 that we will continuously expand the tools and integrate them deeper into Wikipedia's infrastructure. This process will go hand in hand with the Wikipedia community. The adaption to further language versions will increase the number of users. Especially for smaller language versions the RENDER supporting tools have the potential to support the existing amount of editors but also to attract new ones but providing starting points and further information on the related topics.

The quality and the increase of Wikipedia as a worldwide project can only benefit from instruments and tools which are able to ease the entrance and the contribution for a single person.

References

[1] Angelika Mühlbauer, Kai Nissen, Johannes Kroll. D5.1.4 – Evaluation of the tools for diversity management in Wikipedia. 2013

[2] Angelika Adam, Fabian Flöck, Gerrit Holz. D5.1.1 – Definition and Evaluation of Metrics in Wikipedia. 2011

[3] Felix Leif Keppmann, Angelika Mühlbauer, Maurice Grinberg. D6.2.5 – Report on community building activities Y3. 2013